Lecture Notes in Computational Science and Engineering

11

B. Cockburn
G. E. Karniadakis
C.-W. Shu (Eds.)

Discontinuous Galerkin Methods

Theory, Computation and Applications

Springer

Lecture Notes in Computational Science and Engineering 11

Springer
Berlin
Heidelberg
New York
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Bernardo Cockburn
George E. Karniadakis
Chi-Wang Shu (Eds.)

Discontinuous Galerkin Methods

Theory, Computation and Applications

With 138 Figures

Springer

Editors

Bernardo Cockburn
School of Mathematics
University of Minnesota
206 Church Street S.E.
Minneapolis, MN 55455, USA
e-mail: cockburn@math.umn.edu

George E. Karniadakis
Chi-Wang Shu
Division of Applied Mathematics
Brown University
Providence, RI 02912, USA
e-mail: gk@cfm.brown.edu
e-mail: shu@cfm.brown.edu

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Discontinuous Galerkin methods / Bernardo Cockburn ... (ed.).
- Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ; London ;
Milan ; Paris ; Singapore ; Tokyo : Springer, 2000
(Lecture notes in computational science and engineering ; 11)
ISBN-13:978-3-642-64098-8 e-ISBN-13:978-3-642-59721-3
DOI: 10.1007/978-3-642-59721-3

Front cover: Unstructured spectral/hp element simulation of wave scattering from an F15 geometry. Contours of the nose to tail component of the magnetic field are shown. Computations by Tim Warburton, Oxford University Computing Laboratory.

Mathematics Subject Classification (1991): 65Cxx, 65Dxx, 73-XX, 76-XX

ISSN 1439-7358
ISBN-13:978-3-642-64098-8

Softcover reprint of the hardcover 1st edition 2000

Cover Design: Friedhelm Steinen-Broo, Estudio Calamar, Spain
Cover production: *design & production* GmbH, Heidelberg
Typeset by the authors using a Springer TEX macro package

Printed on acid-free paper SPIN 10735097 46/3143/LK - 5 4 3 2 1 0

Preface

A class of finite element methods, the Discontinuous Galerkin Methods (DGM), has been under rapid development recently and has found its use very quickly in such diverse applications as aeroacoustics, semi-conductor device simulation, turbomachinery, turbulent flows, materials processing, MHD and plasma simulations, and image processing. While there has been a lot of interest from mathematicians, physicists and engineers in DGM, only scattered information is available and there has been no prior effort in organizing and publishing the existing volume of knowledge on this subject.

In May 24-26, 1999 we organized in Newport (Rhode Island, USA), the first international symposium on DGM with equal emphasis on the theory, numerical implementation, and applications. Eighteen invited speakers, leaders in the field, and thirty-two contributors presented various aspects and addressed open issues on DGM. In this volume we include forty-nine papers presented in the Symposium as well as a survey paper written by the organizers. All papers were peer-reviewed. A summary of these papers is included in the survey paper, which also provides a historical perspective of the evolution of DGM and its relation to other numerical methods.

We hope this volume will become a major reference in this topic. It is intended for students and researchers who work in theory and application of numerical solution of convection dominated partial differential equations. The papers were written with the assumption that the reader has some knowledge of classical finite elements and finite volume methods.

Finally, we would like to acknowledge the financial support by the National Science Foundation, the Department of Energy, and the Army Research Office. We especially like to thank Ms. Madeline Brewster who has organized the Symposium, collected the papers, typeset this volume, and made this first symposium on DGM a success.

July 1999

The Organizers
Bernardo Cockburn
George Em Karniadakis
Chi-Wang Shu

Table of Contents

Part I

Overview

The Development of Discontinuous Galerkin Methods

Bernardo Cockburn[1], George E. Karniadakis[2], and Chi-Wang Shu[2]

[1] School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455, USA
[2] Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912, USA

Abstract. In this paper, we present an overview of the evolution of the discontinuous Galerkin methods since their introduction in 1973 by Reed and Hill, in the framework of neutron transport, until their most recent developments. We show how these methods made their way into the main stream of computational fluid dynamics and how they are quickly finding use in a wide variety of applications. We review the theoretical and algorithmic aspects of these methods as well as their applications to equations including nonlinear conservation laws, the compressible Navier-Stokes equations, and Hamilton-Jacobi-like equations.

1 Introduction

Problems of practical interest in which *convection* plays an important role arise in applications as diverse as meteorology, weather-forecasting, oceanography, gas dynamics, aeroacoustics, turbomachinery, turbulent flows, granular flows, oil recovery simulation, modeling of shallow water, transport of contaminant in porous media, viscoelastic flows, semiconductor device simulation, magneto-hydrodynamics, and electro-magnetism, among many others. This is why devising robust, accurate, and efficient methods for numerically solving these problems is of considerable importance and, as expected, has attracted the interest of many researchers and practitioners.

This endeavor, however, is far from trivial because of two main reasons. The first is that the exact solution of (nonlinear) purely convective problems develops discontinuities in finite time; the second is that these solutions might display a very rich and complicated structure near such discontinuities. Thus, when constructing numerical methods for these problems, it must be guaranteed that the discontinuities of the approximate solution are the physically relevant ones. Also, it must be ensured that the appearance of a discontinuity in the approximate solution does not induce spurious oscillations that spoil the quality of the approximation; on the other hand, while ensuring this, the method must remain sufficiently accurate near that discontinuity in order to capture the possibly rich structure of the exact solution.

These difficulties were successfully addressed during the remarkable development of the *high-resolution finite difference* and *finite volume* schemes for

nonlinear hyperbolic systems by means of suitably defined *numerical fluxes* and *slope limiters*. Since discontinuous Galerkin (DG) methods assume discontinuous approximate solutions, they can be considered as generalizations of finite volume methods. As a consequence, the DG methods incorporate the ideas of numerical fluxes and slope limiters into the *finite element* framework in a very natural way; they are able to capture the physically relevant discontinuities without producing spurious oscillations near them; see an illustration of this fact in Fig. 1. Notice that the solution itself is not monotone, however the overshoot and undershoot are not significant and the averages of the solution on the elements are monotone.

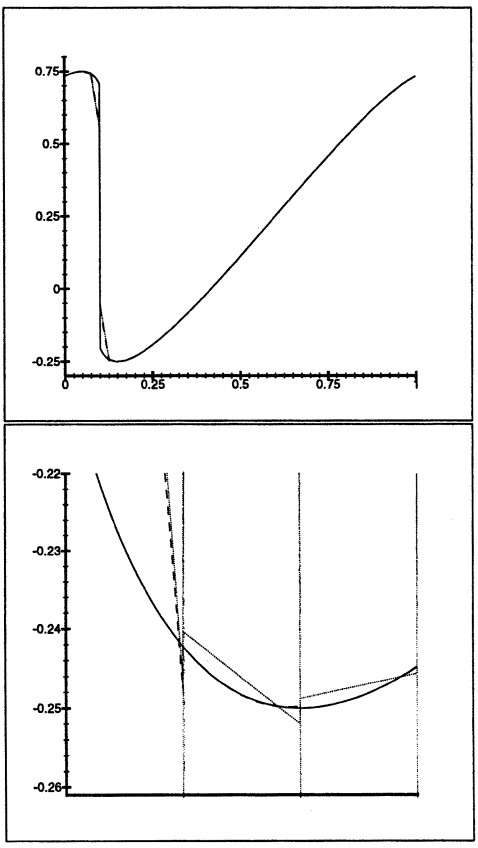


Fig. 1. Burgers equation with periodic boundary conditions and initial data $1/4 + \sin(\pi(2\,x-1))/2$. Comparison of the exact and the approximate solutions obtained with $\Delta x = 1/40$ at $T = 0.40$. Top: full domain, bottom: detail; exact solution (solid line), piecewise linear solution (dotted line), and piecewise quadratic solution (dashed line).

Owing to their finite element nature, the DG methods have the following main advantages over classical finite volume and finite difference methods:

- The actual order of accuracy of DG methods solely depends on the exact solution; DG methods of arbitrarily high formal order of accuracy can be obtained by suitably choosing the degree of the approximating polynomials.
- DG methods are highly parallelizable. Since the elements are discontinuous, the mass matrix is block diagonal and since the size of the blocks is equal to the number of degrees of freedom inside the corresponding elements, the blocks can be inverted by hand (or by using a symbolic manipulator) once and for all.
- DG methods are very well suited to handling complicated geometries and require an extremely simple treatment of the boundary conditions in order to achieve uniformly high-order accuracy.
- DG methods can easily handle adaptivity strategies since refinement or unrefinement of the grid can be achieved without taking into account the continuity restrictions typical of conforming finite element methods. Moreover, the degree of the approximating polynomial can be easily changed from one element to the other. Adaptivity is of particular importance in hyperbolic problems given the complexity of the structure of the discontinuities.

Although the original DG method has been known since 1973, it was only recently that DG methods have evolved in a manner that made them suitable for use in computational fluid dynamics and the aforementioned applications. In this paper, we introduce the DG methods and give an overview of their evolution since their introduction in 1973 by Reed and Hill [145], in the framework of transport of neutrons, until their most recent applications, as well as their theoretical and computational developments.

This paper is organized as follows. In section 2, we present the original DG method and describe its theoretical and computational developments in the framework of linear hyperbolic systems and ordinary differential equations. We also review other early applications, its use to discretize in time parabolic problems, and its introduction to the numerical approximation of viscoelastic flows.

In section 3, we present the evolution of the DG method for nonlinear hyperbolic problems. We show how the first attempts to extend the original DG method lead to implicit schemes and how the efforts to use explicit schemes lead to the construction of the so-called Runge-Kutta DG (RKDG) methods. We show how the RKDG methods incorporated the ideas of numerical flux and slope limiter into the finite element framework to produce formally high-order accurate, nonlinearly stable schemes. Finally, numerical applications to the Euler equations of gas dynamics are displayed.

In section 4, we review how the DG methods were extended to convection-diffusion systems. After presenting some early attempts involving the use of

standard mixed methods, we describe the method used by Bassi and Rebay whose generalization lead to the so-called local DG (LDG) methods. Then, we display applications to the compressible Navier-Stokes equations and to MHD. Finally, we mention the Baumann-Oden DG method for the discretization of second-order equations and several new developments.

In section 5, we describe the extension of the RKDG method to Hamilton-Jacobi equations and the extension of the LDG method to second-order nonlinear degenerate parabolic equations. We present an application to movement by mean curvature.

In section 6, we briefly discuss parallelization and adaptivity for the DG methods. We also discuss several implementational issues. The first is the use of an orthogonal, tensor-product basis for unstructured grids in 2D and 3D. We also discuss quadrature-free implementations of DG methods and point out the object-oriented codes currently in use.

We end this review in section 7, which is devoted to the discussion of open problems and future developments.

2 Linear hyperbolic systems

2.1 The original DG method for the neutron transport problem

The original DG finite element method was introduced in 1973 by Reed and Hill [145] for solving the neutron transport equation

$$\sigma u + \nabla \cdot (\mathbf{a} u) = f, \qquad \text{in } \Omega,$$

where σ is a real number and $\mathbf{a}$ a constant vector. The relevance of the method was recognized by LeSaint and Raviart who in 1974 [117] published its first mathematical analysis.

To display the method, we multiply the equation by a test function v and integrate over an arbitrary subset of Ω, say K. After a formal integration by parts, we get

$$\sigma (u,v)_K - (u, \mathbf{a} \cdot \nabla v)_K + \langle \mathbf{a} \cdot \mathbf{n}_K u, v \rangle_{\partial K} = (f,v)_K,$$

where $\mathbf{n}_K$ denotes the outward unit normal of ∂K, and

$$(u,v)_K = \int_K u v \, dx, \qquad \langle w, v \rangle_{\partial K} = \int_{\partial K} w v \, ds.$$

Next, we construct a triangulation $\mathcal{T}_h = \{ K \}$ of Ω, and take our approximate solution u_h to be a polynomial of degree at most k on each element K of the triangulation. The approximate solution u_h is then determined as the unique solution of the following weak formulation:

$$\begin{aligned} &\forall K \in \mathcal{T}_h : \\ &\sigma (u_h, v)_K - (u_h, \mathbf{a} \cdot \nabla v)_K + \langle \hat{h}, v \rangle_{\partial K} = (f,v)_K, \qquad \forall v \in P^k(K), \end{aligned}$$

where $P^k(K)$ denotes the space of polynomials of degree at most k on the element K and $\hat{h}$ is the numerical flux given by

$$\hat{h}(\mathbf{x}) = \mathbf{a} \cdot \mathbf{n}_K(x) \lim_{s \downarrow 0} u_h(\mathbf{x} - s\,\mathbf{a}).$$

Note that the value $\lim_{s\downarrow 0} u_h(\mathbf{x} - s\,\mathbf{a})$ is nothing but the value of u_h upstream the characteristic direction $\mathbf{a}$. As a consequence, the degrees of freedom of the approximate solution u_h in the element K can be computed in terms of the values of u_h upstream the characteristics hitting ∂K. In other words, the approximate solution u_h can be computed element by element when the elements are suitably ordered according to the characteristic direction $\mathbf{a}$.

2.2 The DG method for ODEs

The first analysis of the DG method as applied to ODEs, was performed in 1974 by LeSaint and Raviart [117] who showed that the method is strongly A-stable of order $2\,k+1$ at mesh points, and that the Gauss-Radau discretization of the DG method is also of order $2\,k+1$ when piecewise polynomials of degree k are used.

It is interesting to note that only one year before the introduction of the DG method by Reed and Hill, Hulme [107,108] had studied a method for ODEs which used the *same* weak formulation as the DG method but employed a *continuous* approximate solution u_h; this method is, however, only of order $2\,k$ at mesh points. A study of global error control for ODEs for this method was carried out in 1994 by Estep and French [84]. Another very interesting work on DG methods for ODEs was done in 1981 by Delfour, Hager and Trochu [70]; they introduce a class of DG methods which are proven to give an order of accuracy up to $2\,k+2$ at the mesh points. Recently, Schötzau and Schwab have obtained a new estimate on the size of the time step needed to solve the implicit system of equations determined by the DG method by means of a simple fixed point iteration technique; see the reference in the lecture notes by Schwab [152].

In 1988, Johnson [112] gave an analysis of error control for the DG method for stiff ODEs and later in 1995, Estep [83] extended this analysis to general non-autonomous ODEs. Finally, in 1996, Böttcher and Rannacher [37] introduced a new adaptive error control technique for ODEs by using the DG method.

2.3 Analysis of the original DG method

A priori error estimates. In 1974, LeSaint and Raviart [117] made the first analysis of the DG method and proved a rate of convergence of $(\Delta x)^k$ in the $L^2(\Omega)$-norm for general triangulations and of $(\Delta x)^{k+1}$ for tensor products of polynomials of degree k in one variable defined on Cartesian grids. In 1986, Johnson and Pitkäranta [113] proved a rate of convergence of $(\Delta x)^{k+1/2}$ for

general triangulations and in 1991, Peterson [140] numerically confirmed this rate to be optimal. In 1988, Richter [146] obtained the optimal rate of convergence of $(\Delta x)^{k+1}$ for some structured two-dimensional non-Cartesian grids. The issue of the loss of order of convergence was addressed again in 1991 by Lin and Zhou [120] who proved that the *standard Galerkin* method using bilinear approximations defined on almost uniform Cartesian is of order 2; the order of this method for arbitrary meshes is only one. In 1994, Zhou and Lin [180] extended this result to piecewise-linear approximations in almost uniform triangulations. Then, in 1996 Lin, Yan, and Zhou [119] showed first order convergence for the DG method using piecewise-constant approximations. Their result holds for almost uniform grids of rectangles and for almost uniform grids of triangles; their technique is based on a key approximation result. In this volume, Lin [118] reviews this technique and applies it to several finite element approximations for hyperbolic problems. Also in this volume, Falk [86] reviews several techniques of analysis for finite element methods for hyperbolic problems including the DG method and the continuous Galerkin method.

All the above mentioned papers assume that the exact solution is smooth. In 1993, Lin and Zhou [121] proved convergence to the weak solution assuming only that the exact solution belongs to $H^{1/2}(\Omega)$. More recently, Houston, Schwab and Süli [102] proved spectral convergence of the DG method assuming that the exact solution is piecewise analytic. In this volume, E. Süli, Ch. Schwab, and P. Houston [162] review these results and extends them to *hp*-DGFEM for PDEs with non-negative characteristic form. Finally, Cockburn, Luskin, Shu and Süli [54] showed that if the exact solution is in L^2 but is locally smoother, error estimates can be obtained between the exact solution and a suitably post-processed approximate solution.

Concerning the issue of super-convergence, in 1994, Biswas, Devine and Flaherty [36] discovered that the approximate solution of the DG method super-converges at the Gauss-Radau points. A rigorous proof of this fact was recently found by Adjerid, Flaherty, and Krivodonova [3]; the groundwork for this analysis was carried out in 1998 by Adjerid, Aiffa and Flaherty [1]. Another indication of super-convergence was obtained by Lowrie [127] who reported numerical evidence of the existence of a component of the error of the DG method that was $(2k+1)$-th. order accurate. This experimental indication was put on firm mathematical basis by Cockburn, Luskin, Shu and Süli [55] who showed that, assuming that the exact solution is sufficiently smooth, a simple post-processing of the approximate solution obtained with polynomials of degree k does produce an approximation of order $2k+1$; in this volume, they present a short version of this result. Also in this volume, Lin [118] proposes a new error estimation technique for finite element approximations of hyperbolic problems.

A posteriori error analysis. In 1990, Stroubolis and Oden [158] studied a posteriori error estimates for the DG method. Later, Bey and Oden [33]

obtained the first *hp*- a posteriori error estimates for the DG method; parallelization strategies based on these estimates were developed in 1995 by Bey, Patra, and Oden [35] and in 1996 by Bey, Oden and Patra [34].

A posteriori error analysis of finite element methods for hyperbolic problems, including a slight modification of the original DG method, have been studied in 1996 by Süli [159] and in 1997 by Süli and Houston [161]; see also the 1999 lectures notes on this subject by Süli [160].

Wave propagation analysis. An analysis of wave propagation for the DG method is given in this volume by Rasetarinera, Hussaini, and Hu [143]. Also in this volume is a paper by Sherwin [154] which is devoted to the study of numerical phase properties of continuous and discontinuous Galerkin methods using a high-order basis (see section 6.2).

2.4 Early applications of the DG method

Besides the application of the DG to the simulations of neutron transport and to ODEs, applications of this method to the analysis of wave propagation in elastic media was done from 1975 to 1976 by Oden and Wellford [173,134,174,175], and to optimal control in 1978 by Delfour and Trochu [71].

2.5 Time discretization of parabolic equations

Also in 1978, Jamet [110] used the DG method to discretize *in time* parabolic equations and showed that the method was of order k. Since then, several authors have studied this method. Thus, in 1985, K. Eriksson, C. Johnson and V. Thomée [82] proved that the method was of order $2k+1$ at the nodes and later Erikson and Johnson studied the issue of error control in a series of papers [77–81] starting in 1987 and ending in 1995. In 1997, Makridakis and Babuška [132] studied the effect on adaptive mechanisms on the stability of the method. In this volume, Machiels [130] investigates an adaptive procedure for this method based on a new a posteriori error control. Also in this volume, Estep and Freund [85] use it to solve nonlinear reaction-diffusion systems; they show how to use an inexact Newton method preconditioned with Krylov-subspace iteration. Finally, Schötzau and Schwab have studied how to actually solve the system of equations defined by the DG methods; they show that it is possible to *decouple* the system into several scalar equations of the same type; see the lecture notes by Schwab [152].

2.6 DG methods for viscoelastic flows

In 1989, the DG method of Reed and Hill was applied for the first time for the numerical computation of viscoelastic flows by Fortin and Fortin [93]. The idea was to apply the DG method to the constitutive law relating the

so-called extra-stress tensor in terms of the velocity. In this volume, Fortin, Béliveau, Heuzey and Lioret [92] review the development of this idea and Baaijiens, Bogaerds and Verbeeten [13] study the successes and failures of the use of these methods in viscoelastic fluid analysis. A recent application of the DG to these problems was pursued in 1998 by Sun, Smith, Armstrong, and Brown [163].

Mathematical analysis of these methods have been carried out in 1992 by Baranger and Sandri [21], in 1995 by Baranger and Wardi [22], in 1997 by Baranger and Machmoum [20], and in 1998 by Bahhar, Baranger and Sandri [14]. See also the 1996 paper by Baranger and Machmoum [19].

2.7 New developments: DG methods for Maxwell's equations

The equations of (viscous) magneto-hydrodynamics, that include the Maxwell's equations, have been discretized with DG methods by Warburton and Karniadakis [171]. Other applications to the Maxwell's equations are presented in three papers in this volume. Warburton [169] presents the use of the DG method with unstructured polymorphic *hp*-finite elements; Kopriva, Woodruff and Hussaini [116] consider a spectral discontinuous method; and Cai [39] deals with the problem of defining the basis functions for electromagnetic scattering of curved surfaces.

3 Nonlinear hyperbolic systems

3.1 The space DG-discretization

The success of the DG method for linear hyperbolic problems, made the extension to the nonlinear hyperbolic systems

$$\mathbf{u}_t + \sum_{i=1}^{d} (\mathbf{f}_i(\mathbf{u}))_{x_i} = 0,$$

the natural step in the development of the method. An extension of the original DG method can be obtained as follows. To simplify the presentation, let us assume that u is a scalar-valued function; in the case of a vector-valued $\mathbf{u}$, we proceed similarly component by component. Thus, we multiply the above equation by a test function and formally integrate by parts to get

$$(u_t, v)_K - \sum_{i=1}^{d} (\mathbf{f}_i(u), \partial_{x_i} v)_K + \sum_{i=1}^{d} \langle \mathbf{f}_i(u)(\mathbf{n}_K)_i, v \rangle_{\partial K} = 0.$$

The approximate solution u_h is now defined as the solution of the following weak formulation:

$$\forall K \in \mathcal{T}_h :$$

$$((u_h)_t, v)_K - \sum_{i=1}^{d} (\mathbf{f}_i(u_h), \partial_{x_i} v)_K + \langle \hat{h}, v \rangle_{\partial K} = 0, \qquad \forall v \in P^k(K),$$

where $\hat{h}$ is an approximation to the trace of $\sum_{i=1}^{d} \mathbf{f}_i(u)(\mathbf{n}_K)_i$ on the boundary of the element K, in other workds, it is nothing but an *approximate Riemann solver*; see, for example, Toro [166] and the references therein. This shows that the treatment of the boundary conditions is natural and extremely simple. Chavent and Salzano [44] used the above DG-space discretization in 1982 for the first time in the framework of nonlinear conservation laws.

Now, it only remains to discretize the above equations in time. However, it is not simple to find a time discretization that would result in a stable, efficient, and formally high-order accurate method. At this point in the development of the DG methods for hyperbolic conservation laws, this was the main difficulty.

3.2 Implicit time-discretizations

Global time-discretizations. The presence of the nonlinearities $\mathbf{f}_i$ prevents the element-by-element computation of the solution that was possible in the linear case considered by Reed and Hill [145]. This is so because it is no longer possible to determine the characteristics explicitly. Thus, one is forced to use implicit time discretizations and hence, to solve at each time step a new nonlinear system of equations. This renders the method computationally very inefficient for hyperbolic problems. In 1989, Bar-Yoseph [17] and in 1990, Bar-Yoseph and Elata [18] explored this approach.

Local time-discretizations. A way around this difficulty was found independently in 1994 by Richter [148], in 1996 by Lowrie [127] and Lowrie, Roe and, van Leer [129], [126] and the work in [157], [179], and [40] by the group of Haber and his collaborators. It consists of using space-time elements constructed in such a way that a local element-by-element computation is still possible. In this volume, Lowrie and Morel [128] use this approach to deal with hyperbolic systems with stiff relaxation; Carranza, Fang, and Haber [40] use a space-time DM method to the simulation of oxidation-driven fractures in super-alloys; and Yin, Acharya, Sobh, Haber, and Tortorelli [179] apply this technique to perform elastodynamic analysis (see the application to precipitate nucleation and growth in aluminum alloy quench processes by Sobh, Huang, Yin, Haber, and Tortorelli [157]). Also in this volume, Richter [149] considers several ways to carry out this approach.

It is interesting to note that in 1990, Hulbert and Hughes [106] proposed a space-time finite element method for elastodynamics that used a DG method in time and a continuous-in-space approximation. This work, reminiscent of the approach took in 1978 by Jamet [110] for parabolic equations, can now be seen as a step toward the development of fully space-time DG methods for hyperbolic problems.

Analysis of the DG method. To rigorously analyze the DG method in the nonlinear case is very difficult; in fact, up to date there are only three results

in this direction. The first was established in 1994 by Jiang and Shu [111] for one-dimensional nonlinear conservation laws with strictly convex or concave nonlinearities. It states that if the approximate solution converges, it converges to the entropy solution; this holds for any degree of the approximating polynomials.

The other two results hold for a version of the space-time DG method for scalar nonlinear conservation laws that contains an *additional* term called the *shock-capturing* term. In 1995, Jaffré, Johnson, and Szepessy [109] proved the convergence of the approximate solution to the entropy solution. In 1996, Cockburn and Gremaud [50] obtained the only a posteriori error estimate for this method in the nonlinear case; they also proved, not only convergence, but also an error estimate that yields the order of convergence of $1/4$ in $L^\infty(0,T;L^1)$ for possibly discontinuous solutions. These results hold in any number of space dimensions and for any value of the polynomial degree.

3.3 Explicit schemes: The Runge-Kutta Discontinuous Galerkin methods

The Euler method. To avoid the difficulty of implicit time discretizations, in 1982, Chavent and Salzano [44] constructed an explicit version of the DG method in the case of a one-dimensional scalar conservation law. They discretized in space by using the DG method with piecewise linear elements and then discretized in time by using the simple forward Euler method. Unfortunately, a classical von Neumann analysis shows that the resulting method is unconditionally unstable when the ratio $\frac{\Delta t}{\Delta x}$ is held constant; it is stable if $\frac{\Delta t}{\Delta x}$ is of order $\sqrt{\Delta x}$. This condition is reasonable if the method is used in conjunction with explicit methods for convection-diffusion schemes, as done for secondary oil recovery by Chavent and Jaffré [43], but it is a very restrictive condition for hyperbolic problems.

Incorporation of the slope limiter. To improve the stability of the scheme, in 1989, Chavent and Cockburn [42] modified the scheme by introducing a suitably defined slope limiter, following the ideas introduced in 1974 by van Leer [168]. They thus obtained a scheme that was proven to be total variation diminishing in the means (TVDM) and total variation bounded (TVB) provided that the CFL number, $f' \frac{\Delta t}{\Delta x}$, is less than or equal to $1/2$; convergence of a subsequence is thus guaranteed. Although the numerical results indicate convergence to the correct entropy solutions, the scheme is only first order accurate in time. Moreover, the slope limiter has to balance the spurious oscillations in smooth regions caused by linear instability, hence adversely affecting the quality of the approximation in these regions.

The first RKDG method. These difficulties were overcome by Cockburn and Shu in [58], where the first Runge Kutta Discontinuous Galerkin (RKDG)

method was introduced. This method was constructed by (i) retaining the piecewise linear DG method for the space discretization, (ii) using a special explicit TVD second order Runge-Kutta type discretization introduced by Shu and Osher in 1988 [155] and in 1989 [156], and (iii) *modifying* the slope limiter to maintain the formal accuracy of the scheme at extrema. The resulting explicit scheme was then proven to be linearly stable for CFL numbers less than 1/3, formally uniformly second order accurate in space and time, and total variation bounded in the means (TVBM). The numerical results show second order convergence in smooth regions including at extrema, sharp shock transitions (usually in one or two elements) without oscillations, and convergence to entropy solutions even for non convex fluxes.

High-order accurate RKDG methods. In 1989, Cockburn and Shu [56] generalized this approach and constructed (formally) high-order accurate RKDG methods for the scalar hyperbolic conservation law. To device RKDG methods of order $k+1$, they used (i) the DG-space discretization method with polynomials of degree k for the space discretization, (ii) a TVD Runge-Kutta $(k+1)$-th order accurate explicit time discretization, and (iii) a generalized slope limiter. The generalized slope limiter was carefully devised to enforce the TVBM property without destroying the accuracy of the scheme. The numerical results, for $k = 1, 2$, indicate $(k+1)$-th order order in smooth regions away from discontinuities as well as sharp shock transitions with no oscillations; convergence to the entropy solutions was observed in all the tests.

In 1994, Biswas, Devine, and Flaherty [36] introduced a new generalized slope limiter. Although no stability results have been proven for this generalized slope limiter, it has the advantage of dealing with local critical points without the aid of any auxiliary parameter. Another distinctive feature is that it can be readily used for *hp*-adaptivity purposes.

One-dimensional systems. These RKDG schemes were extended to one-dimensional systems in 1989 by Cockburn, Lin and Shu [53].

Multi-dimensional scalar equations. The extension of the RKDG method to the scalar multi-dimensional case was done in 1990 by Cockburn, Hou, and Shu [51]. The main contributions of this extension are (i) some accuracy considerations, and (ii) the extension of the generalized slope limiter.

It was found that in order to ensure formal accuracy of order $k+1$ when using polynomials of degree k, quadrature rules, exact for polynomials of degree $2k$, should be used for the integrals inside the elements and quadrature rules, exact for polynomials of degree $2k+1$, should be used for the integrals on the faces of the elements.

The construction of the generalized slope limiter was not simple. This is so, not only because of the more complicated form of the elements but

also because of inherent accuracy barriers imposed by the stability properties. Indeed, since the main purpose of the slope limiter is to enforce the nonlinear stability of the scheme, it is essential to realize that in the multi-dimensional case the constraints imposed by the stability of a scheme on its accuracy are even greater than in the one-dimensional case. Although in the one-dimensional case it is possible to devise high-order accurate schemes with the TVD property, this is not true in several space dimensions, since in 1985, Goodman and LeVeque [96] proved that any TVD scheme is at most first order accurate. Thus, any generalized slope limiter that enforces the TVD property, or the TVDM property for that matter, would unavoidably reduce the accuracy of the scheme to first order accuracy. This is why Cockburn, Hou and Shu [51] devised a generalized slope limiter that enforced a *local* maximum principle only; maximum principles are not incompatible with high-order accuracy. No other class of schemes of second or higher order of accuracy has a proven maximum principle for general nonlinearities $\mathbf{f}$ and unstructured triangulations.

In 1997, Wierse [177] introduced and studied several interesting new *slope limiters* for formally high-order accurate schemes defined in unstructured triangulations.

Multi-dimensional systems. The extension of the RKDG methods to general multi-dimensional systems was initiated in 1991 by Cockburn and Shu in [57] and was completed in 1998 in [60] where applications to the Euler equations of gas dynamics were displayed. One of the contributions of [60] is the construction of a new, practical generalized slope limiter which works very well in triangles and rectangles and with piecewise linear and quadratic elements.

In 1996, Devine and Flaherty [73] introduced a parallel adaptive *hp*-refinement techniques for conservation laws using the RKDG methods.

Numerical experiments for the Euler equations of gas dynamics were performed in 1991 by Bey and Oden [32], in 1997 by Bassi and Rebay [26], in 1998 by Baumann and Oden [31] and by Warburton, Lomtev, Kirby, and Karniadakis [172].

3.4 Other explicit time-stepping schemes

Time-stepping schemes different from the TVD Runge-Kutta time-stepping used by the RKDG can give very good computational results. However, it remains to be proven that those methods share with the RKDG methods the same nonlinear stability properties.

3.5 The DG methods of Allmaras and Halt

Totally independently of the just described development of DG methods, Allmaras and then Halt explored schemes which would now be considered

DG methods. In 1989, Allmaras [5] introduced a DG method for the Euler equations of gas dynamics; an earlier version of his algorithm appeared in the 1987 paper by Allmaras and Giles [6]. He used the Roe parametric variables, piecewise linear test functions, a two-point Gauss quadrature rule on the edges, and a three-stage second-order Runge-Kutta time stepping method; he also took into account the curvature of the boundaries of the domain. In 1992, Halt [99] extended Allmaras' work to higher degree polynomials and to general unstructured grids in two- and three-space dimensions. His numerical test cases include the Ringleb flow, 2-D airfoils and the 3-D Onera M6 wing. See also the 1991 and 1992 papers by Halt and Agarwall [100] and [101], respectively. No slope limiters were considered by the above-mentioned authors.

3.6 Numerical experiments: Gas Dynamics

In what follows, we present some numerical results from some of the papers mentioned above and some new results that display the performance of the method when applied to the Euler equations of gas dynamics.

Approximation of the boundaries. In 1997, Bassi and Rebay [26] showed with a remarkable experiment, the importance of using a good approximation of the boundaries of the space domain. Here, we reproduce some of their results to illustrate that point.

The test problem is the classical two-dimensional isentropic flow around a circle. In Fig. 2, part of the grid is displayed and the corresponding solution using P^1 elements is shown. Note that in this grid, the circle is approximated by a polygon; since each of the kinks of the polygon introduces non-physical entropy production, the approximate solution presents a non-physical wake which does not disappear by further refining the grid! By simply taking into account the exact shape of the boundary, a remarkably improved approximation is obtained, as can be seen in Fig. 3. Note also the improvement of the approximation as the degree of the polynomials is increased from 1 to 3!

Spectral convergence. We consider the isentropic flow in the geometry shown in Fig. 4; the numerical results we show are from the work of Warburton, Lomtev, Kirby, and Karniadakis [172]. Low-order methods erroneously produce entropy from inlet to outlet for this problem. In Fig. 4 (bottom), we show that the entropy errors converge *exponentially fast* to zero as the degree of the polynomials increases. A comparison is shown on the plot of the bottom between a fully unstructured and a hybrid discretization; more elements are used in the unstructured grid.

Approximation of contact discontinuities. Now, we consider the classical double-Mach reflection problem; we show results from the work of Cockburn and Shu [60]. In Fig. 5, we show details of the approximation of the

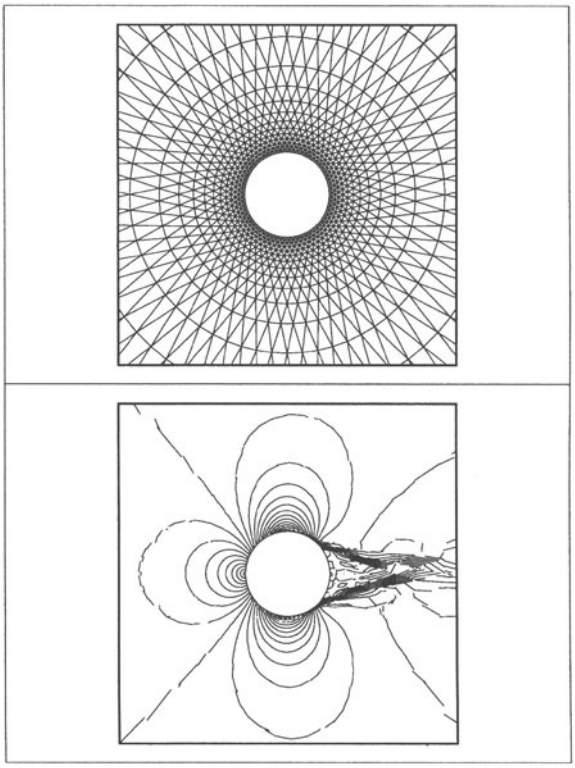


Fig. 2. Grid "64 × 16" with a piecewise linear approximation of the circle (top) and the corresponding solution (Mach isolines) using P^1 elements (bottom).

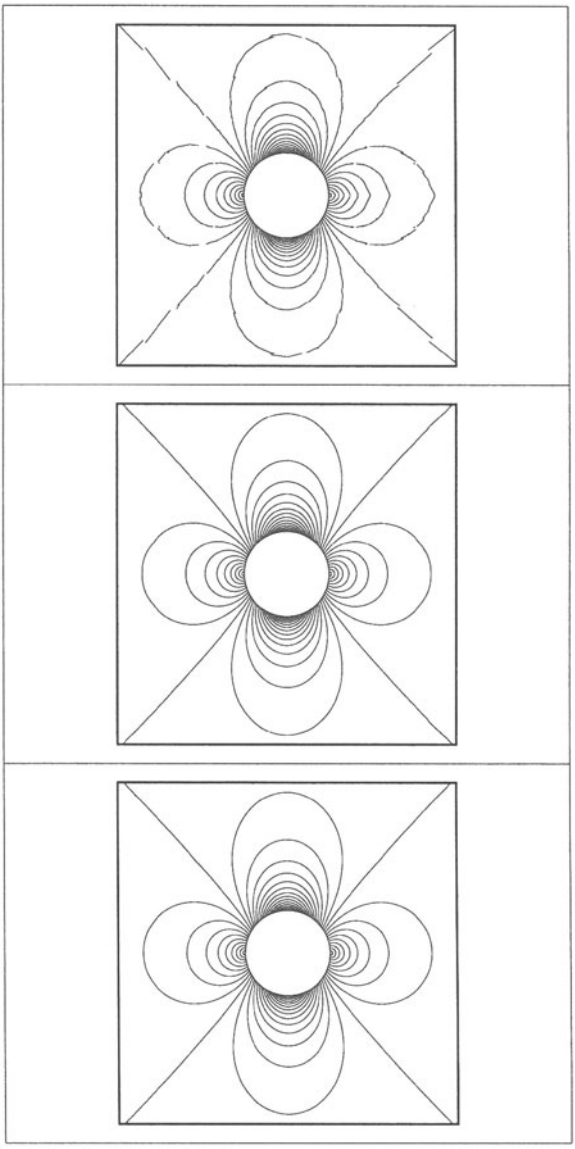


Fig. 3. Grid "64 × 16" with exact rendering of the circle and the corresponding P^1 (top), P^2(middle), and P^3 (bottom) approximations (Mach isolines).

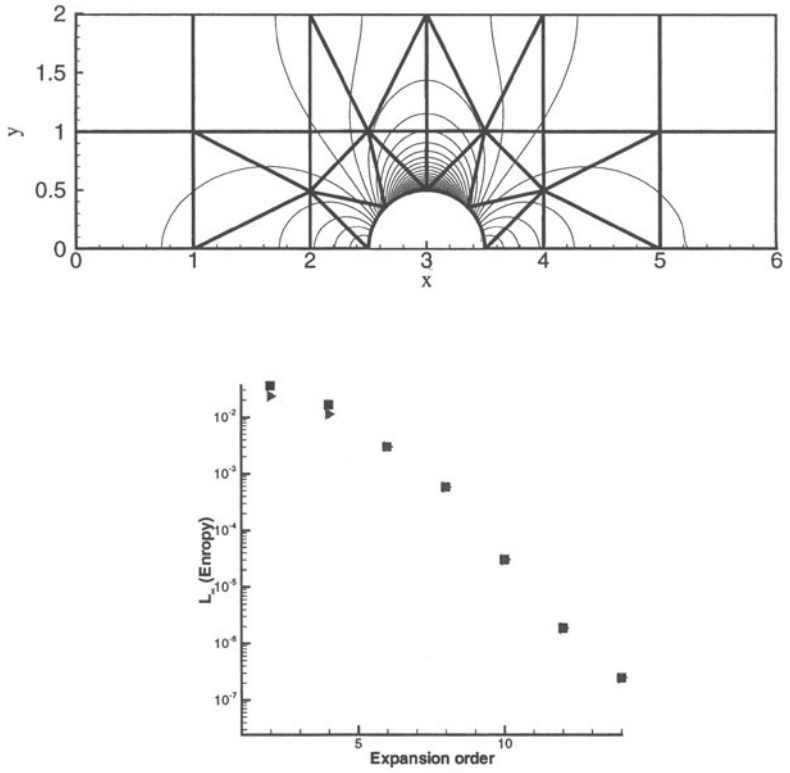


Fig. 4. Density contours (top) obtained on a hybrid grid for an inviscid $M = 0.3$ flow (left). History of convergence (bottom): Exponential convergence of the error is obtained for an unstructured (triangles) and a hybrid (squares) grid.

density. Note that the strong shocks are very well resolved with both P^1 and P^2 elements. Also, note that there is a remarkable improvement in the approximation of the density near the contacts when going from P^1 elements to P^2 elements.

Conclusions. The first experiment illustrates the fact that very good approximations of the boundaries are crucial. The second experiment shows that the DG method can achieve spectral accuracy, and so, that polynomials of high degree should be used when dealing with a smooth solution. Finally, the last experiment shows that this is also desirable even when the solution is not smooth.

3.7 New developments and applications

In this volume, Barth [23] presents a new simplified version of the DG methods for conservation laws that incorporates a symmetrization technique; Despres [72] presents a DG method for solving the Euler equation in an axisymmetric geometry; Gremaud [98] presents an application of the DG method to granular flow; and, van der Ven and van der Vegt [167] present a study of the accuracy, resolution and computational complexity of a DG method.

4 Convection-diffusion systems

4.1 A DG method for convection-diffusion problems

In 1992, Richter [147] proposed a direct extension of the original DG method to linear convection-diffusion equations. Richter proved that if the convection is dominant, that is, if the viscosity coefficients were of the order of the meshsize, the optimal order of convergence is $k + 1/2$ when polynomials of degree k are used.

4.2 A coupled Euler/Navier-Stokes solver

In 1989, Allmaras [5], see also Allmaras and Giles [7], proposed to couple his DG method for the Euler equations of gas dynamics and a compressible Navier-Stokes solver. The two solvers were applied to different, overlapping regions of the computational domain.

4.3 The Upwind-mixed methods for advection-diffusion equations

In 1991, Dawson [63] introduced the so-called upwind-mixed methods (UMM) for advection-diffusion problems. The main idea of these methods is to combine a mixed finite element approximation for the second-order terms with an

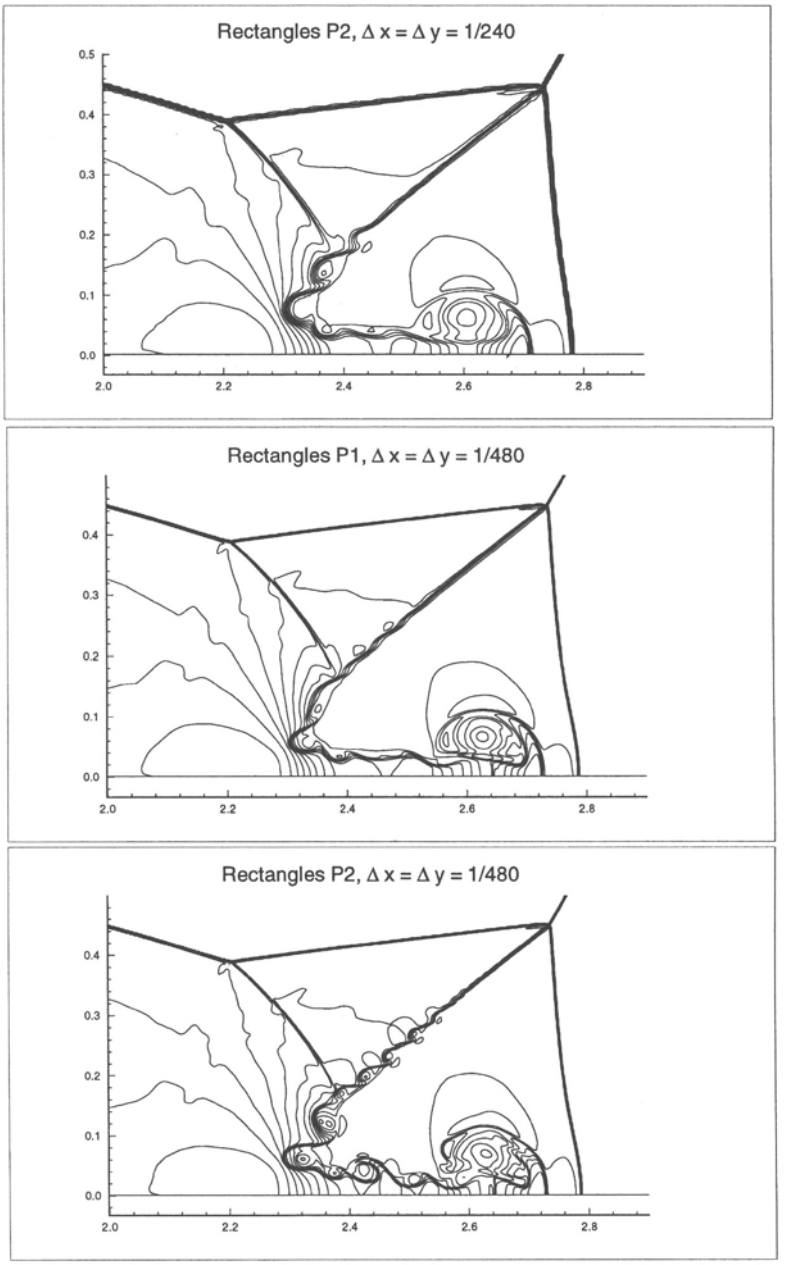


Fig. 5. Double Mach reflection problem. Blown-up region around the double Mach stems. Density ρ. Third order P^2 with $\Delta x = \Delta y = \frac{1}{240}$ (top); second order P^1 with $\Delta x = \Delta y = \frac{1}{480}$ (middle); and third order P^2 with $\Delta x = \Delta y = \frac{1}{480}$ (bottom).

upwinding for the advective terms. Since the UUM always use discontinuous approximations for the solution, this is a very natural combination. In 1993, Dawson [64] extended his analysis to multi-dimensions, and in 1998, [66], analyzed the application of the method to nonlinear contaminant transport equations. All this work was done for the lowest-order Raviart-Thomas space. Recently, Dawson and Aizinger [67] considered the UMM that uses the DG method as its so-called upwinding scheme and obtained error estimates for arbitrary degree polynomial spaces. For applications of UMM to transport problems arising in porous media, see the references in [67].

4.4 A DG method for semiconductor device simulation

Strongly related with the above UMM method are the extensions of the RKDG method to nonlinear, convection-diffusion systems of the form

$$\partial_t \mathbf{u} + \nabla \cdot \mathbf{F}(\mathbf{u}, D\,\mathbf{u}) = 0, \text{ in } (0,T) \times \Omega,$$

proposed in 1995 by Chen, Cockburn, Jerome, and Shu [46] for the hydrodynamic model for semiconductor device simulation and by Chen, Cockburn, Gardner, and Jerome [45] for the quantum hydrodynamic model for semiconductor device simulation. In these extensions, approximations of second and third-order derivatives of the discontinuous approximate solution were obtained by using simple *projections* into suitable finite elements spaces. This projection requires the inversion of global mass matrices, which in [46] and [45] are 'lumped' in order to maintain the high parallelizability of the method. Since in [46] and [45] polynomials of degree one are used, the 'mass lumping' is justified; however, if polynomials of higher degree were used, the 'mass lumping' needed to enforce the full parallelizability of the method could cause a degradation of the formal order of accuracy.

In this volume, Cockburn, Jerome, and Shu [52] review some of the above results and addresses the issue of the utility of modeling and simulation in determinining properties of semiconductors.

4.5 DG-mixed methods for Compressible Navier Stokes

In 1998, Lomtev, Quillen and Karniadakis [125] used the DG-space discretization method to deal with the convective part of the compressible Navier-Stokes equations and used a mixed method to approximate the diffusive part of the equations.

4.6 The method of Bassi and Rebay and the LDG method

In 1997, Bassi and Rebay [25] proposed an extension of the DG-space discretization method for the compressible Navier-Stokes equations. In this approach, the original idea of the DG-space discretization method is applied to

both u and Du which are now considered as *independent* unknowns. Like the RKDG methods, the resulting methods are highly parallelizable methods of high-order accuracy which are very efficient for time-dependent, convection-dominated flows. In 1998, Cockburn and Shu [59] introduced the local discontinuous Galerkin (LDG) methods, which are a generalization of Bassi and Rebay's approach, and proved stability and error estimates for the method.

The basic idea to construct the LDG methods is to *suitably rewrite* the original system as a larger, degenerate, first-order system and then discretize it in space by the DG method. By a careful choice of this rewriting and of the numerical fluxes, nonlinear stability can be achieved even without slope limiters, just as for the RKDG method in the purely hyperbolic case; see Jiang and Shu [111]. The resulting method is element-wise conservative, a property which is particularly difficult to preserve with high-order finite elements.

The large amount of degrees of freedom and the restrictive conditions of the size of the time step for explicit time-discretizations, render the LDG methods inefficient for diffusion-dominated problems; in this situation, the use of methods with continuous-in-space approximate solutions is recommended. However, as for the successful RKDG methods for purely hyperbolic problems, the extremely local domain of dependency of the LDG methods allows a very efficient parallelization that by far compensates for the extra amount of degrees of freedom in the case of convection-dominated flows.

4.7 The LDG method for purely diffusive problems

The parabolic case. Next, we illustrate the definition of the LDG method as applied to the heat equation with variable diffusion coefficient $\nu(\mathbf{x})$:

$$u_t - \nabla \cdot (\nu \nabla u) = f, \qquad \text{in } (0,T) \times \Omega,$$

We then rewrite the above equation as the following first-order degenerate system:

$$\begin{aligned} u_t + \nabla \cdot \mathbf{q} &= f, && \text{in } (0,T) \times \Omega, \\ \mathbf{q} + \nu \nabla u &= 0, && \text{in } (0,T) \times \Omega. \end{aligned}$$

and after multiplying by test functions w and $\mathbf{v}$ and formally integrating by parts, we obtain

$$\begin{aligned} &(u_t, w)_K - (\mathbf{q}, \nabla w)_K + \langle \mathbf{q} \cdot \mathbf{n}, w \rangle_{\partial K} = (f, w)_K, \\ &(\frac{1}{\nu}\mathbf{q}, \mathbf{v})_K - (u, \nabla \cdot \mathbf{v})_K + \langle u, \mathbf{v} \cdot \mathbf{n}_K \rangle_{\partial K} = 0. \end{aligned}$$

We are now ready to define the LDG-space discretization method:

$$\begin{aligned} &\forall K \in \mathcal{T}_h : \\ &((u_h)_t, w)_K - (\mathbf{q}, \nabla w)_K + \langle \hat{\mathbf{h}}, w \rangle_{\partial K} = (f, w)_K, \qquad \forall w \in P^k(K), \\ &(\frac{1}{\nu}\mathbf{q}_h, \mathbf{v})_K - (u_h, \nabla \cdot \mathbf{v})_K + \langle \hat{u}, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial K} = 0, \qquad \forall \mathbf{v} \in (P^k(K))^d, \end{aligned}$$

where $\hat{\mathbf{q}}$ and $\hat{u}$ are numerical fluxes that must be carefully defined.

Independently, in 1994 Giannakouros [95] and in 1997 Bassi and Rebay [25], took their numerical fluxes $\hat{\mathbf{q}}$ and $\hat{u}$ to be the arithmetic average of the two values of $\mathbf{q}_h$ and u_h at the boundary of the elements. Bassi and Rebay [25] reported an order of convergence of order $k+1$ for even values of the polynomial degree k and of order k for odd values and in 1998, Cockburn and Shu [59] proved this result. In 1999, Lomtev and Karniadakis [123], showed numerical evidence that the method is exponentially accurate even for highly distorted grids.

In 1998, Cockburn and Shu [59] showed that for a fairly general class of numerical fluxes, the LDG methods are of order k when polynomials of degree k are used. However, their numerical experiments indicate that the order of convergence varies with the definition of the numerical fluxes and that a simple choice gives the optimal rate of $k+1$; in this volume, Castillo [41] gives a proof of this fact.

The elliptic case. It must be pointed out that when applied to *elliptic* problems, the LDG method can be ill-posed if the numerical fluxes are not carefully chosen; this happens, for example, for the fluxes chosen by Giannakouros and by Bassi and Rebay for their original DG scheme. This difficulty was overcome in 1997 by Bassi, Rebay, Mariotti, Pedinotti, and Savini [28] by means of a suitable modification of their original DG scheme; the resulting scheme was then further developed in 1998 by Bassi and Rebay [27]. At the same time, Brezzi, Manzini, Marini, Pietra and Russo [38] analyzed this problem and found several modifications resulting in well posed numerical methods for which they proved optimal error estimates; the scheme developed in [28] and [27] is one of these methods.

4.8 Numerical experiments: Compressible Navier-Stokes

From now on, as is customary in the finite element community, we use p instead of k to denote the degree of polynomials. The numerical results we show next are from Lomtev and Karniadakis [123].

Transonic flow past an airfoil. First we consider a refinement study for a transonic flow past an airfoil NACA0012 at an angle of attack $\alpha = 10°$, freestream Mach number $Ma = 0.8$, and Reynolds number based on the freestream velocity and the airfoil chord equal to $Re = 73$. The wall temperature is equal to the freestream total temperature. The same problem is considered in [25] and is one of the benchmark problems suggested in the GAMM (1986) workshop [94]. The mesh is shown in Fig. 6; it extends 4 chords downstream and consists of 592 elements, which is about one-fourth of the number used in [25]. Three different discretizations with p-refinement were used corresponding to order 2, 4 and 6. The maximum order used in [25]

was 3. In Fig. 7, we plot Mach contours for the first two discretizations ($p = 2$ and 4) that show the improvement in the solution as the polynomial order is increased. A more quantitative comparison is shown in Table 1 where we present the drag and lift coefficients for the three meshes; very good agreement with the results of [25] is obtained. The same is true for the distribution of the pressure and friction coefficients around the airfoil as shown in Fig. 8.

Table 1. Drag and lift coefficients corresponding to different p-refinements.

Item	$p = 2$	$p = 4$	$p = 6$
C_d	0.68287	0.67858	0.6758
C_l	0.47625	0.53022	0.53173

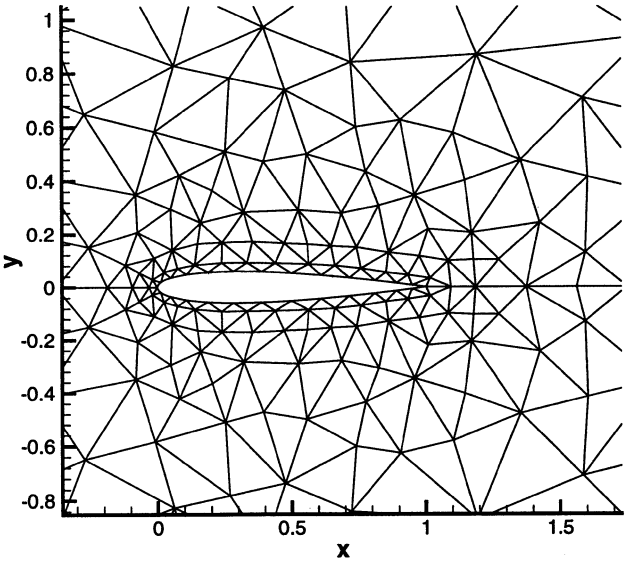


Fig. 6. Discretization around a NACA0012 airfoil; 592 elements are used.

Supersonic flow past an airfoil. We now consider a supersonic flow past a NACA 4420 airfoil at Mach number 2 and Reynolds number (based on

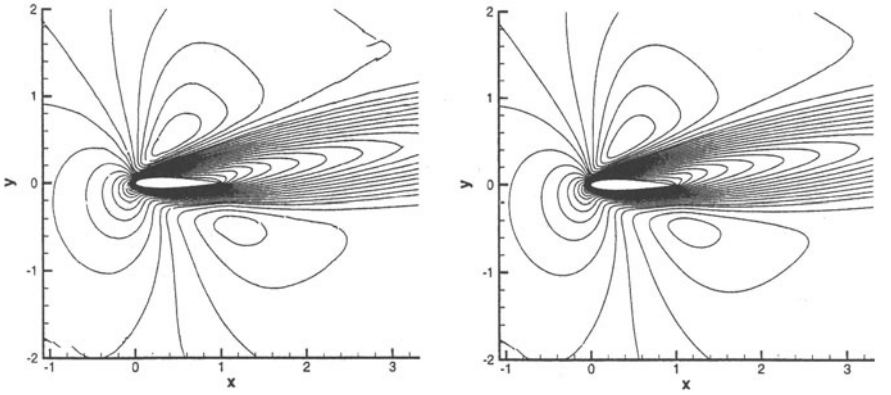


Fig. 7. Mach contour lines for discretization with $p = 2$ (left) and $p = 4$ (right).

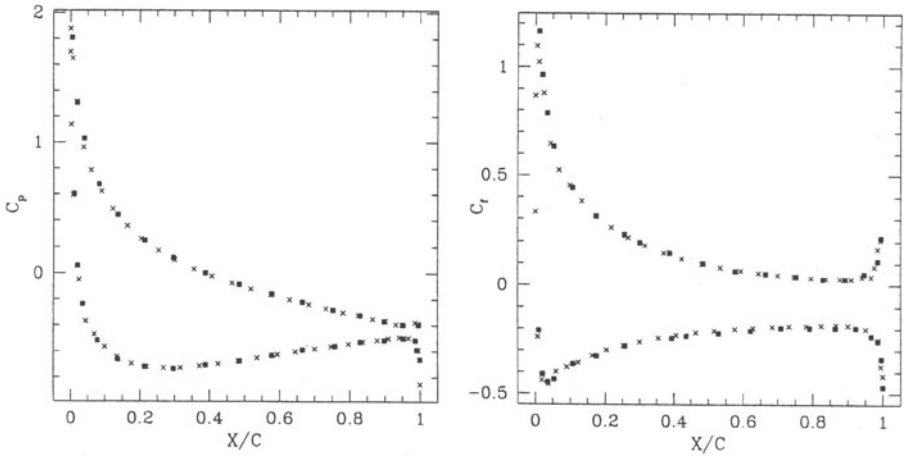


Fig. 8. Pressure (left) and drag (right) coefficients. Solid squares are data from [25] and crosses are from the current simulation for $p = 6$.

the chord length) 2400; the angle of attack is 20°. The domain extends from 1.25 chords upstream to 3.75 chords downstream and is discretized with 1492 triangles. Discretization and density contours and streamlines are shown in Fig. 9; the results are identical to earlier results obtained with results using a mixed formulation in [125]. Variable polynomial order is used from zero (constant elements) around the shock to $p = 5$ in the wake. No flux limiters or filtering were used in this simulation.

4.9 Numerical experiments: viscous MHD

The numerical results we show next are from the work of Warburton and Karniadakis [171].

Simulation of the Orszag-Tang Vortex. We have performed a series of detailed simulations in order to investigate the small-scale structure exhibited in MHD turbulence. In particular, we consider a problem first studied by Orszag & Tang (1979) [136] in the compressible case and later extended by Dahlburg & Picone (1989) [62] to the compressible case. The initial conditions are non-random, periodic fields with the velocity field being solenoidal. The total initial pressure consists of the superposition of appropriate incompressible pressure distribution upon a flat pressure field corresponding to an initial average Mach number below unity. It was found in [136] and [62] that the coupling of the two-dimensional flow with the magnetic field causes the formation of singularities, i.e. excited small-scale structure, which although not as strong as the singularities in three-dimensional turbulence, they are certainly much stronger than two-dimensional hydrodynamic turbulence. Moreover, it was found in [62] that compressibility causes formation of additional small-scale structure such as massive jets and bifurcation of eddies. Our interest here is to investigate if we can capture these fine features both on structured and unstructured meshes, as shown in Fig 10.

The initial conditions we used were:

$$\rho = 1, \quad u = -sin(\frac{2\pi y}{L}), v = sin(\frac{2\pi x}{L}), B_x = -sin(\frac{2\pi y}{L}), B_y = sin(\frac{4\pi x}{L}),$$
$$p = C + \frac{1}{4}cos(\frac{8\pi x}{L}) + \frac{4}{5}cos(\frac{4\pi x}{L})cos(\frac{2\pi y}{L}) - cos(\frac{2\pi x}{L})cos(\frac{2\pi y}{L}) + \frac{1}{4}cos(\frac{4\pi y}{L}),$$

where C fixes the initial average Mach number and p is the instantaneous pressure for the equivalent incompressible flow.

We first simulate this MHD flow on a hybrid grid consisting of quadrilaterals and triangles as shown in Fig.10. We perform the simulations using the formulation of Powell [141] for the magnetic field as well as the streamfunction formulation with the objective of investigating divergence errors in the magnetic field. The rest of the parameters of this simulation are given in the

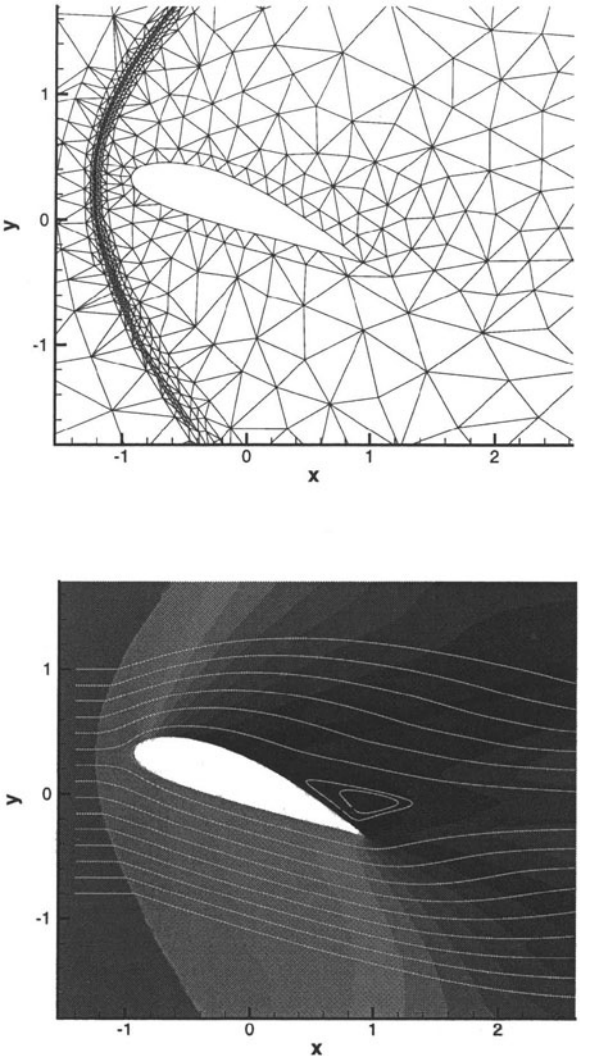


Fig. 9. Discretization around a NACA 4420 airfoil (top) and density contours and streamlines (bottom) at Mach number 2.

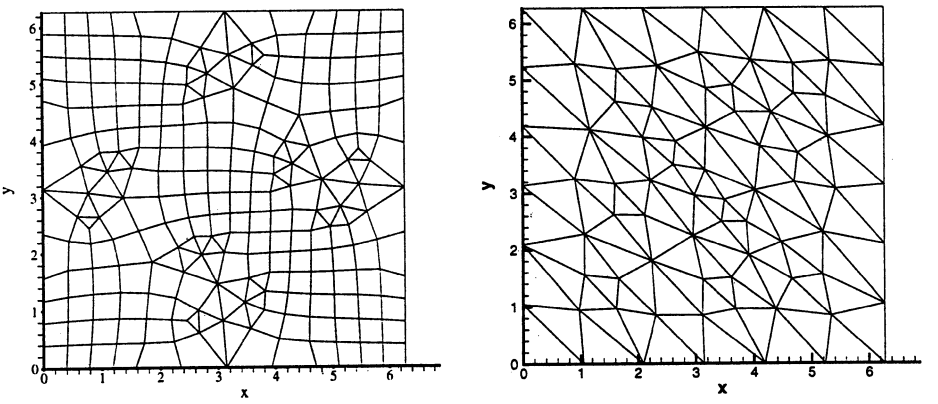


Fig. 10. Hybrid mesh on the left and unstructured mesh on the right used for the Orszag-Tang vortex simulations.

paper by Warburton and Karniadakis [171]. In Fig. 11, we plot streamlines of the incompressible flow as well as the compressible flow at Mach number 0.4 and non-dimensional time $t = 2.0$. These results agree very well with the simulations of [62] at the same set of parameters. We note here that the compressible flow exhibits structures of finer features compared to the incompressible flow but the differences in the magnetic field are less obvious.

4.10 Baumann-Oden DG method

In 1998, Baumann and Oden [29] introduced a new DG method for the discretization of second-order problems; see also the paper by Oden, Babuška and Baumann [133]. Since the method is not a mixed method, it results in fewer degrees of freedom per element, a property that may make it competitive in Navier-Stokes approximations. For diffusion problems, this method is stable when polynomials of degree greater or equal to 2; adaptive *hp*-versions for Navier-Stokes equations have been implemented which exhibit exponential convergence rates. In this volume, Oden and Baumann [135] consider their method for convection-diffusion and the Navier-Stokes equations; see also the work done in 1998 by Baumann and Oden in [30]. In this volume, Arnold, Brezzi, Cockburn and Marini [8] propose a unified framework which contains almost all DG methods for elliptic equations including the Baumann-Oden method.

In this volume, Béatrice Rivière and Mary Wheeler [150] presents an error analysis of three interior penalty methods, some of which are related

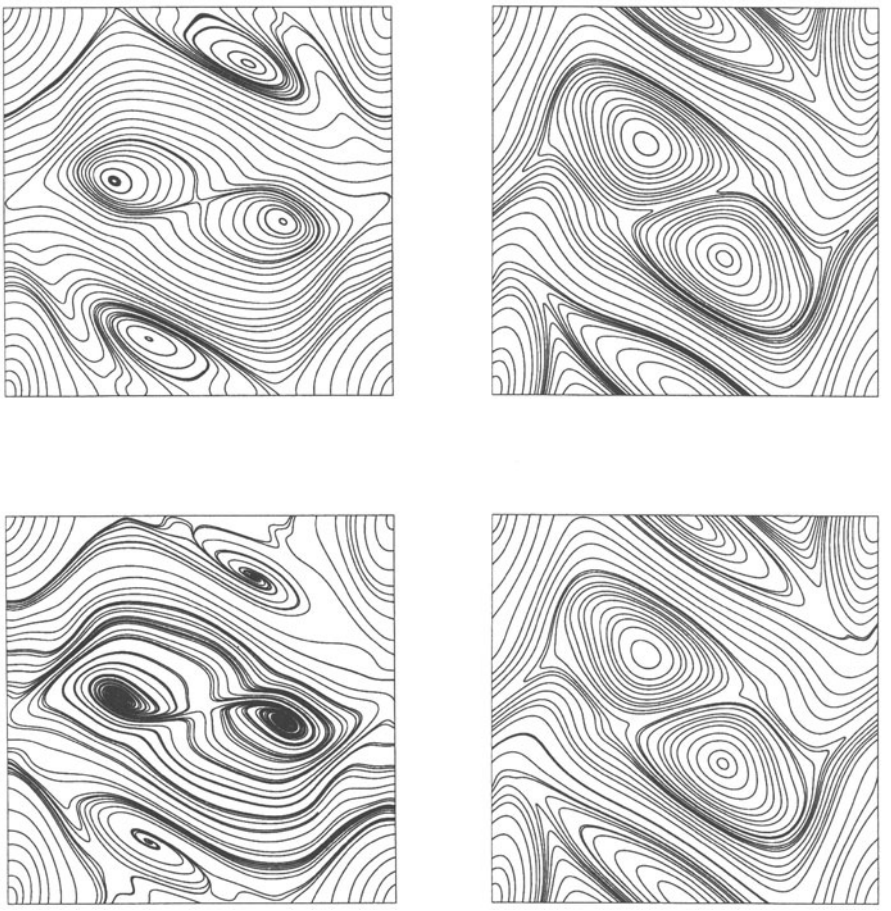


Fig. 11. Compressible Orszag-Tang vortex (t=2, instantaneous fields, $Mach = 0.4$). Top: Incompressible flow; Left: Flow streamlines; Right: Magnetic Streamlines. Bottom: Compressible flow; Left: Flow streamlines; Right: Magnetic Streamlines.

to the Baumann and Oden method. Also in this volume, an extension of the error analysis of the Baumann-Oden DG method to partial differential equations with non-negative characteristic form, is presented by Süli, Schwab, and Houston [162].

4.11 New developments and applications

In this volume, there are several new contributions to the development of DG methods for convection-diffusion problems:

- Bassi [24] reviews his recent work on the high-order, implicit, DG solution of the Reynolds Averaged Navier-Stokes equations coupled with the k-omega turbulence model closure; Rebay [144] shows how to use a GMRES solver in conjunction with the DG method for the compressible Navier-Stokes equations; and Liu and Shu [122] consider the use of the DG method for 2D incompressible flows.
- Prasad, Milovich, Shestakov, Kershaw, and Shaw [142] present a 3D unstructured ALE hydrodynamic DG method; and Lomtev, Kirby, and Karniadakis [124] introduce a discontinuous Galerkin ALE method for compressible flows in moving domains.
- Dawson, Aizinger and Cockburn [68] apply the LDG method to contaminant transport; Schwanenberg and Kongeter [153] use the method for shallow water equations; and Carranza, Fang, and Haber [40] introduce an adaptive DG method for coupled viscoplastic crack growth and chemical transport.
- Atkins [10] outlines the construction of a robust, high-order simulation tool for aerospace applications.

5 Hamilton-Jacobi and second-order nonlinear equations

5.1 The method

Recently, Hu and Shu [104] extended both the RKDG and the LDG method to the Hamilton-Jacobi equation

$$u_t + H(Du) = f,$$

and to the general nonlinear second-order differential equation

$$u_t + F(u, Du, D^2u) = f,$$

for which the mapping $r \mapsto F(r, \cdot, \cdot)$ is increasing and $F(\cdot, \cdot, X) \geq F(\cdot, \cdot, Y)$ provided that $X \leq Y$.

The main idea is to exploit the equivalence of the Hamilton-Jacobi equation with the conservation law systems satisfied by the gradient of the solution and utilize the advantage of finite elements in maintaining the solution

(not individual components of its gradient) as piecewise polynomials. A least square procedure is used to apply the discontinuous Galerkin framework to the conservation law system satisfied by the gradients, and the solution is recovered from its gradients by using again the Hamilton-Jacobi equation. Numerical results indicate that it is very important to keep the solution itself a polynomial and use the least square procedure.

5.2 Numerical experiment: Motion by mean curvature

To illustrate the performance of the DG method, we consider the initial value problem for the following nonlinear second-order differential equation:

$$\varphi_t - (1 - \varepsilon K)\sqrt{1 + \varphi_x^2 + \varphi_y^2} = 0, \qquad 0 < x < 1, 0 < y < 1,$$
$$\varphi(x, y, 0) = 1 - \tfrac{1}{4}(\cos 2\pi x - 1)(\cos 2\pi y - 1),$$

where K is the mean curvature defined by

$$K = -\frac{\varphi_{xx}(1 + \varphi_y^2) - 2\varphi_{xy}\varphi_x\varphi_y + \varphi_{yy}(1 + \varphi_x^2)}{(1 + \varphi_x^2 + \varphi_y^2)^{\frac{3}{2}}},$$

and ε is a small constant. Periodic boundary conditions are imposed. This problem was studied in 1998 by Osher and Sethian [137] by using the finite difference ENO schemes.

We can see that the resolution is excellent even without using any limiters. The singularity of the solution is captured sharply without noticeable oscillations.

5.3 New developments

In this volume, Hu, Lepsky and Shu [103] present a study of the least square procedure for DG methods for Hamilton-Jacobi equations. Also, Augoula and Abgrall [12] develop a new algorithm for Hamilton-Jacobi equations.

6 Parallelization, adaptivity and implementational issues

6.1 Parallelization and adaptivity

Parallelization strategies for the steady-state transport equation were developed in 1995 by Bey, Patra, and Oden [35] and in 1996 by Bey, Oden and Patra [34]. They were based on the 1996 *hp*- a posteriori error estimate obtained by Bey and Oden [33].

In 1994, Biswas, Devine, and Flaherty [36] carried out the first study of parallelization and adaptivity for RKDG methods for nonlinear conservation laws; a remarkable feature of this study is the use of their generalized slope

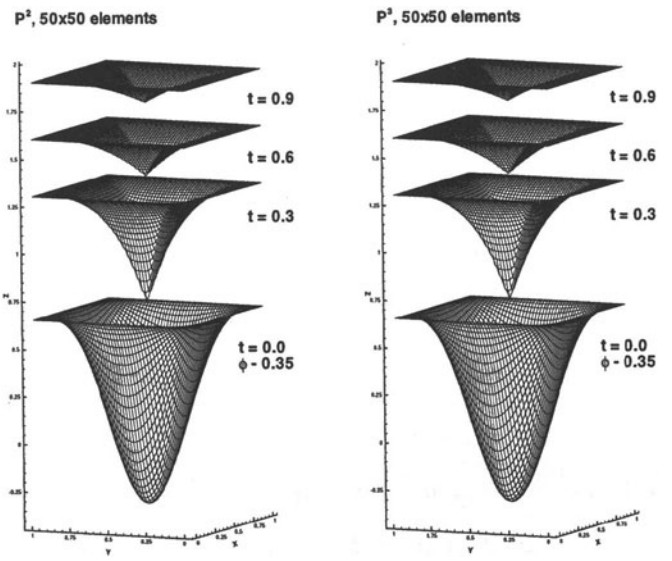


Fig. 12. Propagating surfaces, rectangular mesh, $\varepsilon = 0$.

limiter to perform adaptive limiting. In 1996, Devine and Flaherty [73] devised a parallel adaptive *hp*-refinement techniques for hyperbolic systems in 2 space dimensions; a local time-stepping was used. The crucial issue of load balancing, which addresses the tension between parallelization and adaptivity, was considered in 1993 by Devine, Flaherty, Wheat, and Maccabe [75] for two-space dimension problems, and for three-space dimension problems in 1994 by deCougny, Devine, Flaherty, Loy, and Ozturan [69], and by Ozturan, deCougny, Shephard, and Flaherty [138], and in 1995 by Devine, Flaherty, Loy, and Wheat [74]. Parallel strategies, like *predictive* load-balancing, with local time-stepping techniques in the three-dimensional case have been devised and tested in 1997 by Flaherty, Loy, Shephard, Szymanski, Teresco, and Ziantz [91], in 1998 by Flaherty, Loy, Özturan, Shephard, Szymanski, Teresco and Ziantz [89], and in 1999 by Flaherty, Loy, Shephard, Simone, Szymanski, Teresco and Ziantz [90]. Recently, Teresco, Beall, Flaherty, and Shephard [165] extended this work and developed the TRELLIS framework and the RPM parallel data management. In this volume, Flaherty, Loy, Shephard and Teresco [88] report on the newest development of this technique.

In this volume, Aharoni and Barak [4] present an iterative, asynchronous parallel algorithm for PDEs using DG discretizations.

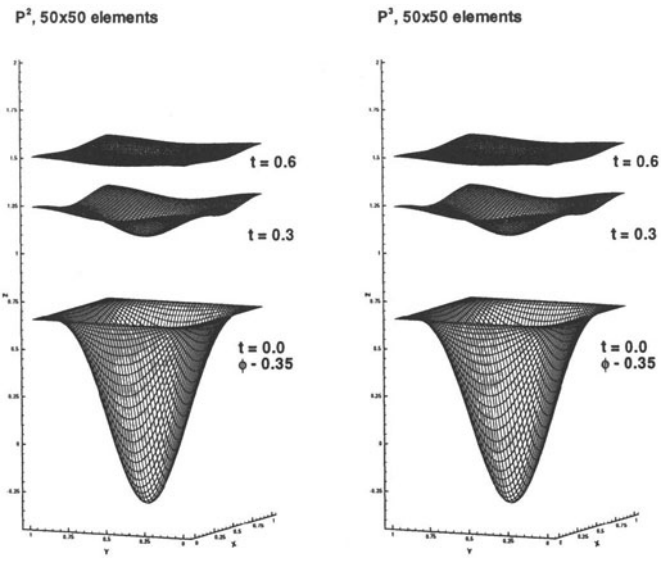


Fig. 13. Propagating surfaces, rectangular mesh, $\varepsilon = 0.1$.

6.2 Spectral/hp Element Methods - The Basis

DG methods are particularly efficient when they are combined with high-order discretization. To this end, Karniadakis and Sherwin [114] and Warburton [170] have developed a hierarchical tensor-type basis extending the original ideas of Dubiner [76]. This basis is appropriate for hybrid discretizations, it is a combination of structured and unstructured domains consisting of polymorphic subdomains; for example, tetrahedra, hexahedra, triangular prisms, and pyramids. For each of these subdomains they have developed a polynomial expansion based upon a new local co-ordinate system [114]. These expansions are polynomials in terms of the local co-ordinates as well as the Cartesian co-ordinates (ξ_1, ξ_2, ξ_3). This is a significant property as primary operations such as integration and differentiation can be performed with respect to the local co-ordinates but the expansion may still be considered as a polynomial expansion in terms of the Cartesian system.

An important property is that these expansions are orthogonal in the Legendre inner product. To wit, we define three principle functions $\phi_i^a(z), \phi_{ij}^b(z)$ and $\phi_{ijk}^c(z)$ in terms of the Jacobi polynomial $P_p^{\alpha,\beta}(z)$ as:

$$\phi_i^a(z) = P_i^{0,0}(z), \qquad \phi_{ij}^b(z) = \left(\tfrac{1-z}{2}\right)^i P_j^{2i+1,0}(z),$$

$$\phi^c_{ijk}(z) = \left(\tfrac{1-z}{2}\right)^{i+j} P_k^{2i+2j+2,0}(z).$$

Using these functions we can construct the orthogonal polynomial expansions:

Hexahedral expansion:
$$\phi_{pqr}(\xi_1,\xi_2,\xi_3) = \phi^a_p(\xi_1)\phi^a_q(\xi_2)\phi^a_r(\xi_3)$$

Prismatic expansion:
$$\phi_{pqr}(\xi_1,\xi_2,\xi_3) = \phi^a_p(\xi_1)\phi^a_q(\eta_2)\phi^b_{qr}(\xi_3)$$

Pyramidic expansion:
$$\phi_{pqr}(\xi_1,\xi_2,\xi_3) = \phi^a_p(\overline{\eta_1})\phi^a_q(\eta_2)\phi^c_{pqr}(\eta_3)$$

Tetrahedral expansion:
$$\phi_{pqr}(\xi_1,\xi_2,\xi_3) = \phi^a_p(\eta_1)\phi^b_{pq}(\eta_2)\phi^c_{pqr}(\eta_3)$$

where

$$\eta_1 = \frac{2(1+\xi_1)}{(-\xi_2-\xi_3)} - 1, \quad \overline{\eta_1} = \frac{2(1+\xi_1)}{(1-\xi_3)} - 1, \quad \eta_2 = \frac{2(1+\xi_2)}{(1-\xi_3)} - 1, \quad \eta_3 = \xi_3,$$

are the local co-ordinates illustrated in figure 14.

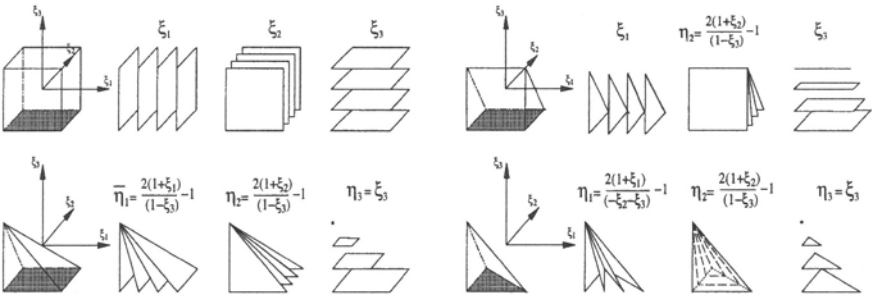


Fig. 14. The local coordinates (ξ_1,ξ_2,ξ_3).

The hexahedral expansion is simply a standard tensor product of Legendre polynomials (since $P_p^{0,0}(z) = L_p(z)$). In the other expansions the introduction of the degenerate local co-ordinate systems is linked to the use of the more unusual functions $\phi^b_{ij}(z)$ and $\phi^c_{ijk}(z)$. Both these functions contain factors of the form $\left(\frac{1-z}{2}\right)^p$ which is necessary to keep the expansion a polynomial of the Cartesian co-ordinates (ξ_1,ξ_2,ξ_3). For example, the co-ordinate η_2 in the prismatic expansion necessitates the use of the function $\phi^b_{qr}(\xi_3)$ which introduces a factor of $\left(\frac{1-\xi_3}{2}\right)^q$. The product of this factor with $\phi^a_q(\eta_2)$ is a polynomial function in ξ_2 and ξ_3. Since the remaining part of the prismatic expansion, $\phi^a_p(\xi_1)$, is already in terms of a Cartesian co-ordinate, the whole expansion is a polynomial in terms of the Cartesian system.

The polynomial space, in Cartesian co-ordinates, for each expansion is:

$$\mathcal{P} = \text{Span}\{\xi_1^p\ \xi_2^q\ \xi_3^r\} \tag{1}$$

where pqr for each domain is

$$\begin{array}{lll} \text{Hexahedron} & 0 \le p \le P_1, 0 \le q \le P_2, & 0 \le r \le P_3, \\ \text{Prism} & 0 \le p \le P_1, 0 \le q \le P_2, & 0 \le q + r \le P_3, \\ \text{Pyramidic} & 0 \le p \le P_1, 0 \le q \le P_2, & 0 \le p + q + r \le P_3, \\ \text{Tetrahedron} & 0 \le p \le P_1, 0 \le p + q \le P_2, & 0 \le p + q + r \le P_3. \end{array} \tag{2}$$

The range of the p, q and r indices indicate how the expansions should be formed to generate a complete polynomial space. We note that if $P_1 = P_2 = P_3$ then the tetrahedral and pyramidic expansions span the same space and are in a subspace of the prismatic expansion which is in turn a subspace of the hexahedral expansion.

An important property of the hybrid spectral basis is that it is orthogonal in the new coordinate system. This simplifies greatly the discontinuous Galerkin formulation, since all mass matrices are diagonal and their inversion is trivial.

6.3 Quadrature-free implementations

In 1998, Atkins and Shu [11] introduced the first quadrature-free implementation of the RKDG and LDG methods. The idea is to use an easily manipulated local basis, such as the local basis used in the Taylor expansions at the center of the cell, and expand the nonlinear terms in a (suitably truncated) polynomial in this local basis based on the solution itself. The integration of products of polynomials in this local basis can be precomputed and stored, in fact a similarity transformation allows one to only store extensive data for one reference object in each class of elements (triangles, quadralaterals, elements with curved boundaries, etc.). Significant speed up can be obtained for linear problems and simple nonlinear problems such as Euler equations with only multiplicative nonlinearity and one division (by density). The coding structure is also simplified in this formulation.

6.4 Object-oriented implementations

Several object-oriented implementations of DG methods have already been developed. The code NEKTAR (freeware), developed at Brown University, is written in C++ and MPI for parallel implementation, is being currently used in more than twenty universities, national laboratories, and industries. As pre-processor, it uses the code FELISA [139] to generate 2D and 3D grids and METIS [115] for parallel domain decomposition. Specifically, to obtain the partition of the 3D mesh for simulation of compressible Navier-Stokes and viscous MHD, NEKTAR uses a multi-level graph theoretical approach, similar to the one used in METIS, which, however, takes into account the p-modes on each element by using appropriate weights. The code allows for a variable polynomial order per element and for all different shapes of elements

including tetrahedra, hexahedra, triangular prisms, and pyramids using the tensor-product Jacobi basis described earlier. This discretization flexibility is useful for complex geometry simulations, especially for viscous compressible and MHD flows and leads to large parallel efficiencies due to the high volume-over-surface ratio associated with the p-expansions.

The TRELLIS framework and the RPM data management system are also object-oriented implementations; see the 1998 paper by Teresco, Beall, Flaherty and Shephard [165].

In this volume, Atkins [10] reports on efforts towards the contruction of a high-order simulation tool for aerospace applications based on DG methods. Also, Prasad, Milovich, Shestakov, Kershaw and Shaw [142] describe a 3D ALE version of a DG method for solving hydrodynamic problems relevant to inertial confinement fusion.

7 Open problems and concluding remarks

7.1 Open problems and future developments

One of the main challenges for the development of finite element methods is the construction and analysis of efficient techniques for problems in computational fluid dynamics. In what follows, we discuss open questions and future developments for one of those methods, the DG methods.

Local conservativity. It is well known that practioners in the area of numerics for nonlinear conservation laws, overwhelmingly prefer locally conservative numerical schemes. This is not the case for elliptic or parabolic equations, however, for which the widely used classical finite element methods are not locally conservative. These points of view clearly clash when the issue of how to approximate convection-diffusion problems arises. A deep analysis of these two properties constitutes a very interesting open problem. An effort in this direction can be found in this volume in the paper by Hughes, Engel, Mazzei, and Larson [105].

Slope limiters. An important component of the RKDG method for transient nonlinear hyperbolic systems is the generalized slope limiter. Although this slope limiter does not seem to be needed in computations involving diffusive flows, it is necessary for the current DG methods for purely hyperbolic problems. The slope limiter used in the RKDG methods involve a parameter (which in one-dimensional scalar conservation laws is nothing but an upper bound of the second-order derivative of the solution at critical points) by means of which the limiting does not destroy accuracy at critical points. An efficient way of estimating this parameter in terms of the computed approximate solution remains to be obtained. Another challenging problem is how to devise a slope limiter that is free from such a parameter. Finally, since the

slope limiter is computationally expensive, it would be very useful to devise a DG method that does not have a slope limiter and remains nonlinearly stable and high-order accurate.

Time-steppping. High-order accurate time-discretizations which are capable of treating the convective terms explicitly and the diffusive term implicitly, if necessary, have not been developed yet and are in high demand nowadays.

Also, in order to be able to do adaptivity while maintaining the high parallelizability of the DG methods, new high-order accurate time-stepping methods would have to be created which could use different time steps at different locations. The use of space-time DG methods could be a possible way of overcoming this difficulty, but they tend to be rather difficult to code– and are not very efficient. Another possibility is to extend to high-order accurate schemes the approach used in 1995 by Dawson [65] to devise a first-order accurate, *conservative* variable time-stepping schemes. Non-conservative time-stepping methods can also lead to efficient time-discretizations, but one has to be very careful to exert a tight control on the loss of mass, especially near the discontinuities. A very interesting example is the local time stepping technique introduced in 1997 by Flaherty, Loy, Shephard, Szymanski, Teresco, and Zianz [91].

Quadrature crimes and over-integration. In 1998, Atkins and Shu [11] introduced a quadrature-free implementation of the RKDG method. They used truncated expansions of the nonlinear integrands that could then be evaluated exactly. A challenging problem is to determine the way the above mentioned expansions have to be truncated to ensure both stability and accuracy of the resulting DG method.

In this volume, Lomtev, Kirby, and Karniadakis [124] show that, in order to produce high-quality approximations, over-integration of one or even two extra degrees of accuracy is necessary when steep gradients on the approximate solution appear near the boundary. Although the LDG method have been proven to be stable, even for nonlinear convection, see [59], such result assumes *exact* integration. A systematic study of the dependence of the stability of the LDG method for nonlinear convection with respect to the quadrature rules is an interesting open problem. For elliptic equations, work in this direction was pursued in 1990 by Maday and Ronquist [131] on the *hp*-Galerkin/spectral method.

Approximation of singularities. It is very well known that singularities often appear in nonlinear and even linear flows. In the past, to deal with those, *ad hoc* strategies have been employed. An example is the idea introduced in 1984 by Woodward and Colela [178] to deal with the corner singularity of the forward facing step test problem. In 1998, Cockburn and Shu [60] showed that to deal with that singularity, it is enough to simply refine the mesh

around the corner, as it is customary in standard finite element methods. A significant effort towards a systematic handling of singularities, in particular in fluid dynamics, is being carried out by Schwab [151], [152].

Steady-state computations Efficient solvers for steady-state computations need to be developed; these solvers are also needed when implicit discretizations in time are used. In this volume, Rebay [144] shows how to use the GMRES for DG methods for compressible Navier-Stokes equations.

Error estimation. Several a priori and a posteriori error estimates are currently available for DG methods. For the linear case, see, for example, the review of a priori error estimates in the lecture notes of Cockburn [47], [48]; for a posteriori error estimates, see the lecture notes of Süli [159], [160]. See also the recent papers by Falk and Richter [87], by Houston, Schwab, and Süli [102] and the paper in this volume by Süli, Schwab, and Houston [162]. For the nonlinear case, see the 1996 paper by Cockburn and Gremaud [50] and the 1999 lecture notes by Cockburn [49]. This is a rapidly developing area that deserves special attention since its development will lead to computations with a preassigned, guaranteed accuracy. Refinements of the above mentioned results and adaptivity strategies based on them which fully take advantage of the use of discontinuous approximations will be developed in the coming years.

Super-convergence. In 1994, Biswas, Devine, and Flaherty [36] gathered numerical evidence that, when rectangular elements are used, the approximate solution of the DG method super-converges at the Gauss-Radau points and exploited this for adaptivity purposes. This fact was recently proven [3]; see also the papers by Adjerid, Aiffa and Flaherty [2] and [1]. The search for super-convergence points in simplexes remains an interesting open problem. Also, the way to exploit the super-convergence of the postprocessed solution obtained by Cockburn, Luskin, Shu and Süli [54] for adaptivity purposes remains a challenging open problem.

Multiresolution analysis. The incorporation of multiresolution analysis into the finite element method is an exciting undertaking (which has to be differentiated from the standard finite element *hp*-refinement). One of the main difficulties is devising of wavelets satisfying boundary conditions, but with the use of the DG method this is no longer a requirement. As a consequence, the use of wavelet-based discontinuous Galerkin methods constitutes a possible breakthrough in this direction. In this volume, Coult [61] provides a most needed introduction to the subject.

Relation of the LDG method with other methods.

- **Mixed methods for elliptic equations.** When applied to elliptic equations, the LDG method can be considered to be a mixed method. However, it was originally devised by using discretization techniques closer to convective problems rather that to elliptic ones. This is reflected, for example, in the fact that error estimates for LDG methods can be obtained without having to deal (explicitly) with the classical inf – sup condition! The relationship of the LDG method with standard and stabilized mixed methods (and their hybridization) for elliptic equations is still unexplored.
- **Penalty methods for elliptic equations.** The relation of the LDG method with the DG method of Baumann and with the interior penalty methods of Baker [15] (1977), Wheeler [176] (1978), Arnold [9] (1982), and Baker, Jureidini, and Karakashian [16] (1990), is another interesting open problem. In this volume, Arnold, Brezzi, Cockburn and Marini [8] propose a unified framework that includes almost all the numerical methods proposed for elliptic equations that use totally discontinuous finite element discretizations.
- **Upwind-Mixed Methods** The relation between the Upwind-Mixed Methods introduced by Dawson [63], [64], [65], [66], and [67] and the LDG methods remains unexplored. In this volume, a first step towards a thorough comparison of these methods is presented by Dawson, Aizinger, and Cockburn [68]. An interesting point is to find out if the use of discontinuous discretizations of second-order terms have any advantage over the classical mixed finite element approximations.
- **Streamline diffusion methods.** The relationship between DG methods and streamline-diffusion methods is quite close but has never been studied. For example, Cockburn and Gremaud [50] analyzed these two methods as applied to the nonlinear scalar conservation law with the same technique. See the work of Houston, Schwab and Süli [102] in this direction. Also, there is a close relationship between the generalized slope limiters that some DG methods use and the so-called shock-capturing terms embedded in the definition of the streamline diffusion methods; this relation is still unexplored. An effort in this direction is the paper in this volume by Hughes, Engel, Mazzei, and Larson [105] in which a comparison of discontinuous and continuous methods is offered.
- **The cell discretization method.** In this volume, Greenstadt [97] and Swann [164] review their work on the so-called *cell discretization* method. This is a very interesting method related to nonconforming methods for elliptic equations and possibly to some DG methods.

7.2 Conclusion

Let us conclude this review by saying that the future development of DG methods which will take place in the next few years is an exciting scientific undertaking. Witnesses to the rapid incorporation of the finite element

methodology in computational fluid dynamics are the books by Schwab [151] and Sherwin and Karniadakis [114]; see also the lecture notes of Cockburn [47], [48], Schwab [152], and Süli [159], [160].

Acknowledments. We would like to acknowledge the support of NSF, DOE and ARO for the organization of this First International Symposium on Discontinuous Galerkin Methods. We would also like to thank Francesco Bassi, Igor Lomtev, Stefano Rebay, and Tim Warburton for permitting us to use their plots; and to S.R. Allmaras, C. Baumann, J. Flaherty, R. Haber, C. Johnson, K.W. Morton, T. Oden, and E. Süli for their valuable feedback on the first version of this review. The research work of the first author is supported by NSF and the Minnesota Supercomputing Institute, that of the second author by AFOSR and DOE, and that of the third author by ARO, NSF, NASA, and AFOSR.

References

1. S. Adjerid, M. Aiffa, and J. E. Flaherty. Computational methods for singularly perturbed systems. In J. Cronin and R.E. O'Malley, editors, *Singular Perturbation Concepts of Differential Equations*, AMS Proceedings of Symposia in Applied Mathematics. AMS, 1998.
2. S. Adjerid, M. Aiffa, and J.E. Flaherty. High-order finite element methods for singularly-perturbed elliptic and parabolic problems. *SIAM J. Appl. Math.*, 55:520–543, 1995.
3. S. Adjerid, J.E. Flaherty, and L. Krivodonova. Superconvergence and a posteriori error estimation for continuous and discontinuous Galerkin methods applied to singularly perturbed parabolic and hyperbolic problems. in preparation.
4. D. Aharoni and A. Barak. Parallel iterative discontinuous Galerkin FEM. In *this volume*, 1999.
5. S.R. Allmaras. *A coupled Euler/Navier Stokes algorithm for 2-D unsteady transonic shock/boundary-layer interaction.* PhD thesis, Massachussetts Institute of Technology, 1989.
6. S.R. Allmaras and M.B.Giles. A second order flux split scheme for the unsteady 2-D Euler equations on arbitrary meshes. In *8th. AIAA Computational Fluid Dynamic Conference, Honolulu, Hawai, June 9-11*, 1987. AIAA 87-1119-CP.
7. S.R. Allmaras and M.B.Giles. A coupled Euler/Navier-Stokes algorithm for 2-D transonic flows. In *27th. Aerospace Sciences Meeting, Reno, Nevada, January 9-12*, 1989.
8. D. Arnold, F. Brezzi, B. Cockburn, and D. Marini. DG methods for elliptic problems. In *this volume*, 1999.
9. D.N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19:742–760, 1982.
10. H. Atkins. Steps toward a robust high-order simulation tool for aerospace applications. In *this volume*, 1999.
11. H.L. Atkins and C.-W. Shu. Quadrature-free implementation of discontinuous Galerkin methods for hyperbolic equations. *AIAA Journal*, 36:775–782, 1998.

12. S. Augoula and R. Abgrall. A discontinuous prjection algorith for Hamilton-Jacobi equations. In *this volume*, 1999.
13. F.P.T. Baaijiens, A.C.B. Bogaerds, and W.M.H. Verbeeten. Successes and failures of discontinuous Galerkin methods in viscoelastic fluid analysis. In *this volume*, 1999.
14. A. Bahhar, J. Baranger, and D. Sandri. Galerkin discontinuous approximation of the transport equation and viscoelastic fluid flow on quadrilaterals. *Numer. Methods Partial Differential Equations*, 14:97–114, 1998.
15. G.A. Baker. Finite element methods for elliptic equations using nonconforming elements. *Math. Comp.*, 31:45–59, 1977.
16. G.A. Baker, W.N. Jureidini, and O.A. Karakashian. Piecewise solenoidal vector fields and the Stokes problem. *SIAM J. Numer. Anal.*, 27:1466–1485, 1990.
17. P. Bar-Yoseph. Space-time discontinuous finite element approximations for multidimensional nonlinear hyperbolic systems. *Comput. Mech.*, 5:145–160, 1989.
18. P. Bar-Yoseph and D. Elata. An efficient L^2 Galerkin finite element method for multi-dimensional nonlinear hyperbolic systems. *Internat. J. Numer. Methods Engrg.*, 29:1229–1245, 1990.
19. J. Baranger and A. Machmoum. A "natural" norm for the discontinuous finite element characteristic method: the 1-D case. *RAIRO Modél. Math. Anal.Numér.*, 30:549–574, 1996.
20. J. Baranger and A. Machmoum. Existence of approximate solutions and error bounds for viscoelastic fluid flow: Characteristics method. *Comput. Methods Appl. Mech. Engrg.*, 148:39–52, 1997.
21. J. Baranger and D. Sandri. Finite element approximation of viscoelastic fluid flow: existence of approximate solutions and error bounds. I. Discontinuous constraints. *Numer. Math.*, 63:13–27, 1992.
22. J. Baranger and S. Wardi. Numerical analysis of a FEM for a transient viscoelastic flow. *Comput. Methods Appl. Mech. Engrg.*, 125:171–185, 1995.
23. T. Barth. Simplified DG methods for systems of conservation laws with convex extension. In *this volume*, 1999.
24. F. Bassi. A high-order discontinuous Galerkin method for compressible turbulent flow. In *this volume*, 1999.
25. F. Bassi and S. Rebay. A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations. *J. Comput. Phys.*, 131:267–279, 1997.
26. F. Bassi and S. Rebay. High-order accurate discontinuous finite element solution of the 2D Euler equations. *J. Comput. Phys.*, 138:251–285, 1997.
27. F. Bassi and S. Rebay. An implicit high-order discontinuous Galerkin method for the steady state compressible Navier–Stokes equations. In K.D. Papailiou, D. Tsahalis, D. Périaux, C. Hirsh, and M. Pandolfi, editors, *Computational Fluid Dynamics 98, Proceedings of the Fourth European Computational Fluid Dynamics Conference*, volume 2, pages 1227–1233, Athens, Greece, September 5–7 1998. John Wiley and Sons.
28. F. Bassi, S. Rebay, G. Mariotti, S. Pedinotti, and M. Savini. A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows. In R. Decuypere and G. Dibelius, editors, *2nd European Conference on Turbomachinery Fluid Dynamics and Thermodynamics*, pages 99–108, Antwerpen, Belgium, March 5–7 1997. Technologisch Instituut.

29. C.E. Baumann and J.T. Oden. A discontinuous *hp* finite element method for convection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.* in press, special issue on Spectral, Spectral Element, and *hp* Methods in CFD, edited by G.E. Karniadakis, M. Ainsworth and C. Bernardi.
30. C.E. Baumann and J.T. Oden. A discontinuous *hp* finite element method for the Navier-Stokes equations. In *10th. International Conference on Finite Element in Fluids*, 1998.
31. C.E. Baumann and J.T. Oden. A discontinuous *hp* finite element method for the solution of the Euler equation of gas dynamics. In *10th. International Conference on Finite Element in Fluids*, 1998.
32. K.S. Bey and J.T. Oden. A Runge-Kutta discontinuous Galerkin finite element method for high speed flows. In *10th. AIAA Computational Fluid Dynamics Conference, Honolulu, Hawaii, June 24-27*, 1991.
33. K.S. Bey and J.T. Oden. *hp*-version discontinuous Galerkin methods for hyperbolic conservation laws. *Comput. Methods Appl. Mech. Engrg.*, 133:259–286, 1996.
34. K.S. Bey, J.T. Oden, and A. Patra. A parallel *hp*-adaptive discontinuous Galerkin method for hyperbolic conservation laws. *Appl. Numer. Math.*, 20:321–286, 1996.
35. K.S. Bey, A. Patra, and J.T. Oden. *hp*-version discontinuous Galerkin methods for hyperbolic conservation laws: A parallel strategy. *Internat. J. Numer. Methods Engrg.*, 38:3889–3908, 1995.
36. R. Biswas, K.D. Devine, and J. Flaherty. Parallel, adaptive finite element methods for conservation laws. *Appl. Numer. Math.*, 14:255–283, 1994.
37. K. Bottcher and R. Rannacher. Adaptive error control in solving ordinary differential equations by the discontinuous Galerkin method. Technical report, University of Heidelberg, 1996.
38. F. Brezzi, D. Marini, P. Pietra, and A. Russo. Discontinuous finite elements for diffusion problems. *Numerical Methods for Partial Differential Equations*, 1999. submitted.
39. W. Cai. Mixed high-order basis functions for electromagnetic scattering of curved surfaces. In *this volume*, 1999.
40. F.L. Carranza, B. Fang, and R.B. Haber. An adaptive discontinuous Galerkin model for coupled viscoplastic crack growth and chemical transport. In *this volume*, 1999.
41. P. Castillo. An optimal error estimate for the local discontinuous Galerkin method. In *this volume*, 1999.
42. G. Chavent and B. Cockburn. The local projection P^0 P^1-discontinuous-Galerkin finite element method for scalar conservation laws. *RAIRO Modél. Math. Anal.Numér.*, 23:565–592, 1989.
43. G. Chavent and J. Jaffré. *Mathematical Models and Finite Elements for Reservoir Simulation*, volume 17 of *Studies in Mathematics and its Applications*. North-Holland, Amsterdam, 1986.
44. G. Chavent and G. Salzano. A finite element method for the 1D water flooding problem with gravity. *J. Comput. Phys.*, 45:307–344, 1982.
45. Z. Chen, B. Cockburn, C. Gardner, and J. Jerome. Quantum hydrodynamic simulation of hysteresis in the resonant tunneling diode. *J. Comput. Phys.*, 117:274–280, 1995.

46. Z. Chen, B. Cockburn, J. Jerome, and C.-W. Shu. Mixed-RKDG finite element methods for the 2-D hydrodynamic model for semiconductor device simulation. *VLSI Design*, 3:145–158, 1995.
47. B. Cockburn. An introduction to the discontinuous Galerkin method for convection-dominated problems. In A. Quarteroni, editor, *Advanced numerical approximation of nonlinear hyperbolic equations*, volume 1697 of *Lecture Notes in Mathematics; subseries Fondazione C.I.M.E., Firenze*, pages 151–268. Springer Verlag, 1998.
48. B. Cockburn. Discontinuous Galerkin methods for convection-dominated problems. In T. Barth and H. Deconink, editors, *High-Order Methods for Computational Physics*, volume 9 of *Lecture Notes in Computational Science and Engineering*, pages 69–224. Springer Verlag, 1999.
49. B. Cockburn. A simple introduction to error estimation for nonnlinear hyperbolic conservation laws. Some ideas, techniques, and promising results. In *Proceedings of the 1998 EPSRC Summer School in Numerical Analysis,SSCM*, volume 26 of *The Graduate Student's Guide to Numerical Analysis*, pages 1–46. Springer-Verlag, 1999.
50. B. Cockburn and P.A. Gremaud. Error estimates for finite element methods for nonlinear conservation laws. *SIAM J. Numer. Anal.*, 33:522–554, 1996.
51. B. Cockburn, S. Hou, and C.W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: The multidimensional case. *Math. Comp.*, 54:545–581, 1990.
52. B. Cockburn, J. Jerome, and C.-W. Shu. The utility of modeling and simulation in determining performance and symmetry properties of semiconductors. In *this volume*, 1999.
53. B. Cockburn, S.Y. Lin, and C.W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: One dimensional systems. *J. Comput. Phys.*, 84:90–113, 1989.
54. B. Cockburn, M. Luskin, C.-W. Shu, and E. Süli. Enhanced accuracy by post-processing for finite element methods for hyperbolic equations. in preparation.
55. B. Cockburn, M. Luskin, C.-W. Shu, and E. Süli. Postprocessing of Galerkin methods for hyperbolic problems. In *this volume*, 1999.
56. B. Cockburn and C.W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws II: General framework. *Math. Comp.*, 52:411–435, 1989.
57. B. Cockburn and C.W. Shu. The P^1-RKDG method for two-dimensional Euler equations of gas dynamics. Technical Report 91-32, ICASE, 1991.
58. B. Cockburn and C.W. Shu. The Runge-Kutta local projection P^1-discontinuous Galerkin method for scalar conservation laws. *RAIRO Modél. Math. Anal.Numér.*, 25:337–361, 1991.
59. B. Cockburn and C.W. Shu. The local discontinuous Galerkin finite element method for convection-diffusion systems. *SIAM J. Numer. Anal.*, 35:2440–2463, 1998.
60. B. Cockburn and C.W. Shu. The Runge-Kutta discontinuous Galerkin finite element method for conservation laws V: Multidimensional systems. *J. Comput. Phys.*, 141:199–224, 1998.
61. N. Coult. Wavelet-based discontinuous Galerkin methods. In *this volume*, page pages, 1999.

62. R.B. Dahlburg and J.M. Picone. Evolution of the Orszag-Tang vortex system in a compressible medium. I. Initial average subsonic flow. *Phys. Fluids B*, 1(11):2153–2171, 1989.
63. C.N. Dawson. Godunov-mixed methods for advection-diffusion equations in one space dimension. *SIAM J. Numer. Anal.*, 28:1282–1309, 1991.
64. C.N. Dawson. Godunov-mixed methods for advection-diffusion equations in multidimensions. *SIAM J. Numer. Anal.*, 30:1315–1332, 1993.
65. C.N. Dawson. High resolution upwind-mixed finite element methods for advection-diffusion equations with variable time-stepping. *Numerical Methods for Partial Differential Equations*, 11:525–538, 1995.
66. C.N. Dawson. Analysis of an upwind-mixed finite element method for nonlinear contiminant transport problems. *SIAM J. Numer. Anal.*, 35:1709–1724, 1998.
67. C.N. Dawson and V. Aizinger. Upwing-mixed methods for transport equations. *Comp. Geo.* to appear.
68. C.N. Dawson, V. Aizinger, and B. Cockburn. The Local Discontinuous Galerkin method for contaminant transport problems. In *this volume*, 1999.
69. H.L. deCougny, K.D. Devine, J.E. Flaherty, R.M. Loy, C. Ozturan, and M.S. Shephard. Load balancing for the parallel adaptive solution of partial differential equations. *Appl. Numer. Math.*, 16:157–182, 1994.
70. M. Delfour, W. Hager, and F. Trochu. Discontinuous Galerkin methods for ordinary differential equations. *Math. Comp.*, 36:455–473, 1981.
71. M. Delfour and F. Trochu. Discontinuous Galerkin methods for the approximation of optimal control problems governed by hereditary differential systems. In A. Ruberti, editor, *Distributed Parameter Systems: Modelling and Identification*, pages 256–271. Springer Verlag, 1978.
72. B. Depres. Discontinuous Galerkin method for the numerical solution of euler equations in axisymmetric geometry. In *this volume*, 1999.
73. K.D. Devine and J.E. Flaherty. Parallel adaptive *hp*-refinement techniques for conservation laws. *Appl. Numer. Math.*, 20:367–386, 1996.
74. K.D. Devine, J.E. Flaherty, R.M. Loy, and S.R. Wheat. Parallel partitioning strategies for the adaptive solution of conservation laws. In I Babuška, W.D. Henshaw, J.E. Hopcroft, J.E. Oliger, and T. Tezduyar, editors, *Modeling, mesh generation, and adaptive numerical methods for partial differential equations*, volume 75, pages 215–242, 1995.
75. K.D. Devine, J.E. Flaherty, S.R. Wheat, and A.B. Maccabe. A massively parallel adaptive finite element method with dynamic load balancing. In *Proceedings Supercomputing'93*, pages 2–11, 1993.
76. M. Dubiner. Spectral methods on triangles and other domains. *J. Sci. Comp.*, 6:345–390, 1991.
77. K. Eriksson and C. Johnson. Error estimates and automatic time step control for nonlinear parabolic problems. *SIAM J. Numer. Anal.*, 24:12–23, 1987.
78. K. Eriksson and C. Johnson. Adaptive finite element methods for parabolic problems I: A linear model problem. *SIAM J. Numer. Anal.*, 28:43–77, 1991.
79. K. Eriksson and C. Johnson. Adaptive finite element methods for parabolic problems II: Optimal error estimates in $l_\infty l_2$ and $l_\infty l_\infty$. *SIAM J. Numer. Anal.*, 32:706–740, 1995.
80. K. Eriksson and C. Johnson. Adaptive finite element methods for parabolic problems IV: A nonlinear model problem. *SIAM J. Numer. Anal.*, 32:1729–1749, 1995.

81. K. Eriksson and C. Johnson. Adaptive finite element methods for parabolic problems V: Long time integration. *SIAM J. Numer. Anal.*, 32:1750–1762, 1995.
82. K. Eriksson, C. Johnson, and V. Thomée. Time discretization of parabolic problems by the discontinuous Galerkin method. *RAIRO, Anal. Numér.*, 19:611–643, 1985.
83. D. Estep. A posteriori error bounds and global error control for approximation of ordinary differential equations. *SIAM J. Numer. Anal.*, 32:1–48, 1995.
84. D. Estep and D. French. Global error control for the continuous Galerkin finite element method for ordinary differential equations. *RAIRO, Anal. Numér.*, 28:815–852, 1994.
85. D.J. Estep and R.W. Freund. Using Krylov-subspace iterations in discontinuous Galerkin methods for nonlinear reaction-diffusion systems. In *this volume*, 1999.
86. R. Falk. Analysis of finite element methods for linear hyperbolic problems. In B. Cockburn, G.E. Karniadakis, and C.-W. Shu, editors, *First International Symposium on Discontinuous Galerkin Methods*, volume 33 of *Lecture Notes in Computational Science and Engineering*. Springer Verlag, May 1999.
87. R.S. Falk and G.R. Richter. Explicit finite element methods for symmetric hyperbolic equations. *SIAM J. Numer. Anal.* to appear.
88. J. Flaherty, R.M. Loy, M.S. Shephard, and J. Teresco. Software for the parallel adaptive solution of conservation laws by a discontinuous Galerkin method. In *this volume*, 1999.
89. J.E. Flaherty, R.M. Loy, C. Özturan, M.S. Shephard, B.K. Szymanski, J.D. Teresco, and L.H. Ziantz. Parallel structures and dynamic load balancing for adaptive finite element computation. *Appl. Numer. Math.*, 26:241–265, 1998.
90. J.E. Flaherty, R.M. Loy, M.S. Shephard, M.L. Simone, B.K. Szymanski, J.D. Teresco, and L.H. Ziantz. Distributed octree data structures and local refinement method for the parallel solution of three-dimensional conservation laws. In M.W. Bern, J.E. Flaherty, and M. Luskin, editors, *Grid Generation and Adaptive Algorithms*, volume 113 of *The IMA Volumes in Mathematics and its Applications*, pages 113–134, Minneapolis, 1999. Institute for Mathematics and its Applications, Springer.
91. J.E. Flaherty, R.M. Loy, M.S. Shephard, B.K. Szymanski, J.D. Teresco, and L.H. Ziantz. Adaptive local refinement with octree load-balancing for the parallel solution of three-dimensional conservation laws. *J. Parallel and Dist. Comput.*, 47:139–152, 1997.
92. A. Fortin, A. Béliveau, M.C. Heuzey, and A. Lioret. Ten years using discontinuous Galerkin methods for polymer processing problems. In *this volume*, 1999.
93. M. Fortin and A. Fortin. New approach for the finite element method simulation of viscoelastic flows. *J. Non-Newt. Fluid Mech.*, 32:295–310, 1989.
94. GAMM Workshop, December 4-6 1985, Nice, France. Numerical simulation of compressible Navier-Stokes equations - external 2D flows around a NACA0012 airfoil. In *Ed. INRIA, Centre de Rocquefort, de Rennes et de Sophia-Antipolis*, 1986.
95. I.G. Giannakouros. *Spectral element/Flux-Corrected methods for unsteady compressible viscous flows*. PhD thesis, Princeton University, Dept. of Mechanical and Aerospace Engineering, 1994.

96. J. Goodman and R. LeVeque. On the accuracy of stable schemes for 2D scalar conservation laws. *Math. Comp.*, 45:15–21, 1985.
97. J. Greenstadt. An abridged history of cell discretization. In *this volume*, 1999.
98. P.-A. Gremaud. Simulation of granular flows. In *this volume*, 1999.
99. D.W. Halt. *A compact higher order Euler solver for unstructured grids.* PhD thesis, Washington University, 1992.
100. D.W. Halt and R.K. Agarwall. A compact higher order characteristic-based Euler solver for unstructured grids. In *September*, 1991. AIAA 91-3234.
101. D.W. Halt and R.K. Agarwall. A compact higher order Euler solver for unstructured grids with curved boundaries. In *June*, 1992. AIAA 92-2696.
102. P. Houston, C. Schwab, and E. Süli. Stabilized *hp*-finite element methods for hyperbolic problems. *SIAM J. Numer. Anal.* to appear.
103. C. Hu, O. Lepsky, and C.-W. Shu. The effect of the lest square procedure for discontinuous Galerkin methods for Hamilton-Jacobi equations. In *this volume*, 1999.
104. C. Hu and C.-W. Shu. A discontinuous Galerkin finite element method for Hamilton-Jacobi equations. *SIAM J. Sci. Comput.* to appear.
105. T. Hughes, G. Engel, L. Mazzei, and M. Larson. A comparison of discontinuous and continuous Galerkin methods. In *this volume*, 1999.
106. Hulbert and Hughes. Space-time finite element methods for second-order hyperbolic equations. *Comput. Methods Appl. Mech. Engrg.*, 84:327–348, 1990.
107. B. L. Hulme. One-step piecewise polynomial Galerkin methods for initial value problems. *Math. Comp.*, 26:415–426, 1972.
108. B. L. Hulme. One-step piecewise polynomial Galerkin methods for initial value problems. *Math. Comp.*, 26:881–891, 1972.
109. J. Jaffré, C. Johnson, and A. Szepessy. Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws. *Mathematical Models & Methods in Applied Sciences*, 5:367–386, 1995.
110. P. Jamet. Galerkin-type approximations which are discontinuous in time for parabolic equations in a variable domain. *SIAM J. Numer. Anal.*, 15:912–928, 1978.
111. G. Jiang and C.-W. Shu. On cell entropy inequality for discontinuous Galerkin methods. *Math. Comp.*, 62:531–538, 1994.
112. C. Johnson. Error estimates and adaptive time-step control for a class of one-step methods for stiff ordinary differential equations. *SIAM J. Numer. Anal.*, 25:908–926, 1988.
113. C. Johnson and J. Pitkäranta. An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comp.*, 46:1–26, 1986.
114. G.E. Karniadakis and S.J. Sherwin. *Spectral/hp Element Methods in CFD.* Oxford University Press, 1999.
115. G. Karypis and V. Kumar. *METIS Unstructured graph partitioning and sparse matrix ordering system version 2.0.* Dept. of Computer Science, University of Minnesota, 1995.
116. D.A. Kopriva, S.L. Woodruff, and M.Y. Hussaini. Discontinuous spectral element approximation of Maxwell's equations. In *this volume*, 1999.
117. P. LeSaint and P.A. Raviart. On a finite element method for solving the neutron transport equation. In C. de Boor, editor, *Mathematical aspects of finite elements in partial differential equations*, pages 89–145. Academic Press, 1974.

118. Q. Lin. Full convergence for hyperbolic finite elements. In *this volume*, 1999.
119. Q. Lin, N. Yan, and A.-H. Zhou. An optimal error estimate of the discontinuous Galerkin method. *Journal of Engineering Mathematics*, 13:101–105, 1996.
120. Q. Lin and A. Zhou. A rectangle test for the first order hyperbolic equation. *Proc. Sys. Sci. & Sys., Great Wall Culture Publ. Co., Hong Kong*, pages 234–235, 1991.
121. Q. Lin and A.-H. Zhou. Convergence of the discontinuous Galerkin method for a scalar hyperbolic equation. *Acta Math. Sci.*, 13:207–210, 1993.
122. J.-G. Liu and C.-W. Shu. Numerical results on a high-order discontinuosu Galerkin method for 2D incompressible flows. In *this volume*, 1999.
123. I. Lomtev and G.E. Karniadakis. A discontinuous Galerkin method for the Navier-Stokes equations. *Int. J. Numer. Meth. Fluids*, 29:587–603, 1999.
124. I. Lomtev, R.M. Kirby, and G.E. Karniadakis. A discontinuous Galerkin ALE method for compressible viscous flow in moving domains. In *this volume*, 1999.
125. I. Lomtev, C.W. Quillen, and G.E. Karniadakis. Spectral/*hp* methods for viscous compressible flows on unstructured 2D meshes. *J. Comput. Phys.*, 144:325–357, 1998.
126. R. B. Lowrie, P. L. Roe, and B. van Leer. Space-time methods for hyperbolic conservation laws. In *Barriers and Challenges in Computational Fluid Dynamics*, volume 6 of *ICASE/LaRC Interdisciplinary Series in Science and Engineering*, pages 79–98. Kluwer, 1998.
127. R.B. Lowrie. *Compact higher-order numerical methods for hyperbolic conservation laws.* PhD thesis, University of Michigan, 1996.
128. R.B. Lowrie and J. Morel. Discontinuous Galerkin for hyperbolic systems with stiff relaxation. In *this volume*, 1999.
129. R.B. Lowrie, P.L. Roe, and B. van Leer. A space-time discontinuous Galerkin method for the time accurate numerical solution of hyperbolic conservation laws. 1995. AIAA 95-1658.
130. L. Machiels. A posteriori finite element output bounds of discontinuous Galerkin discretizations of parabolic problems. In *this volume*, 1999.
131. Y. Maday and E.M. Ronquist. Optimal error analysis of spectral methods with emphasis on non-constant coefficients and deformed geometries. In C. Canuto and A. Quarteroni, editors, *Spectral and high order methods for partial differential equations (Como, Italy, 1989)*, pages 91–115. North-Holland, 1990.
132. X. Makridakis and I. Babuška. On the stability of the discontinuous Galerkin method for the heat equation. *SIAM J. Numer. Anal.*, 34:389–401, 1997.
133. J.T. Oden, Ivo Babuška, and C.E. Baumann. A discontinuous *hp* finite element method for diffusion problems. *J. Comput. Phys.*, 146:491–519, 1998.
134. J.T. Oden and L.C. Wellford, Jr. Discontinuous finite element approximations for the analysis of acceleration waves in eslastic solids. *The Mathematics of finite element methods and applications II (J.R. Whiteman, Ed.) Academic Press, London*, pages 269–284, 1976.
135. T.J. Oden and C.E. Baumann. A conservative discontinuous Galerkin method for convection-diffusion and Navier-Stokes problems. In *this volume*, 1999.
136. S.A. Orszag and C. Tang. Small-scale structure of two-dimensional magneto-hydrodynamic turbulence. *J. Fluid Mech.*, 90(1):129–143, 1979.
137. S. Osher and J. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulation. *J. Comput. Phys.*, 79:12–49, 1988.

138. C. Ozturan, H.L. deCougny, M.S. Shephard, and J.E. Flaherty. Parallel adaptive mesh refinement and redistribution on distributed memory computers. *Comput. Methods Appl. Mech. Engrg.*, 119:123–137, 1994.
139. J. Peiro, J. Peraire, and K. Morgan. *Felisa System- Reference Manual.* Dept. of Aeronautics, Imperial College, 1994.
140. T. Peterson. A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation. *SIAM J. Numer. Anal.*, 28:133–140, 1991.
141. K.G. Powell. An approximate Riemann solver for magnetohydrodynamics (that works in more than one dimension). Technical Report ICASE Report 94-24, ICASE, NASA Langley, 1994.
142. M.K. Prasad, J.L. Milovich, A.I. Shestakov, D.S. Kershaw, and J.J. Shaw. 3D unstructures mesh ALE hydrodynamics with the upwind discontinuous Galerkin method. In *this volume*, 1999.
143. P. Rasetarinera, M.Y. Hussaini, and F.Q. Hu. Recent results in wave propagation analysis of the discontinuous Galerkin method. In *this volume*, 1999.
144. S. Rebay. GMRES for discontinuous Galerkin solution of the compressible Navier-Stokes equations. In *this volume*, 1999.
145. W.H. Reed and T.R. Hill. Triangular mesh methods for the neutron transport equation. Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, 1973.
146. G.R. Richter. An optimal-order error estimate for the discontinuous Galerkin method. *Math. Comp.*, 50:75–88, 1988.
147. G.R. Richter. The discontinuous Galerkin method with diffusion. *Math. Comp.*, 58:631–643, 1992.
148. G.R. Richter. An explicit finite element method for the wave equation. *Applied Numerical Mathematics*, 16:65–80, 1994.
149. G.R. Richter. Explicit finite element methods for linear hyperbolic systems. In *this volume*, 1999.
150. B. Rivière and M.F. Wheeler. A discontinuous Galerkin method applied to nonlinear parabolic equations. In *this volume*, 1999.
151. Ch. Schwab. *p- and hp-finite element methods: Theory and applications in solid and fluid mechanics.* Oxford University Press, Oxford, 1998.
152. Ch. Schwab. *hp*-FEM for fluid flow. In T. Barth and H. Deconink, editors, *High-Order Methods for Computational Physics*, volume 9 of *Lecture Notes in Computational Science and Engineering*, pages 325–414. Springer Verlag, 1999.
153. D. Schwanenberg and J. Kongeter. A discontinuous Galerkin method for the shalow water equations with source terms. In *this volume*, 1999.
154. S.J. Sherwin. Numerical phase properties analysis of the continuous and discontinuous Galerkin methods. In *this volume*, 1999.
155. C.-W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *J. Comput. Phys.*, 77:439–471, 1988.
156. C.-W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock capturing schemes, II. *J. Comput. Phys.*, 83:32–78, 1989.
157. N. Sobh, J. Huang, L. Yin, R.B. Haber, and D.A. Tortorelli. A discontinuous Galerkin model for precipitate nucleation and growth in aluminum alloy quench processes. *Internat. J. Numer. Methods Engrg.* to appear.
158. T. Strouboulis and J.T. Oden. A posteriori error estimation of finite element approximations in fluid mechanics. *Comput. Methods Appl. Mech. Engrg.*, 78:201–242, 1990.

159. E. Süli. A posteriori error analysis and global error control for adaptive finite element approximations of hyperbolic problems. In D.F. Griffiths and G.A. Watson, editors, *Numerical Analysis 1995*, volume 344 of *Pitman Lecture Notes in Mathematics Series*, pages 196–190, 1996.
160. E. Süli. A posteriori error analysis and adaptivity for finite element approximations of hyperbolic problems. In D. Kröner, M. Ohlberger, and C. Rhode, editors, *An introduction to recent developments in theory and numerics for conservation laws*, volume 5 of *Lecture Notes in Computational Sciences and Engineering*, pages 123–194. Springer, 1999.
161. E. Süli and P. Houston. Finite element methods for hyperbolic problems: A posteriori error analysis and adaptivity. In I.S. Duff and G.A. Watson, editors, *The State of the Art in Numerical Analysis*, pages 441–471. Clarendon Press, Oxford, 1997.
162. E. Süli, Ch. Schwab, and P. Houston. *hp*-DGFEM for partial differential equations with non-negative characteristic form. In *this volume*, 1999.
163. J. Sun, M.D. Smith, R.C. Armstrong, and R. Brown. Finite element method for viscoelastic flows based on the discrete adaptive viscoelastic stress splitting and the discontinuous Galerkin method. Technical report, Dept. Chemical Engineering, MIT, 1998.
164. H. Swann. The cell discretization algorithm: An overview. In *this volume*, 1999.
165. J.D. Teresco, M.W. Beall, J.E. Flaherty, and M.S.Shephard. A hierarchical partition model for adaptive finite element computation. *Comput. Methods Appl. Mech. Engrg.* submitted.
166. E. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics.* Springer Verlag, 1997.
167. H. van der Ven and J.J.W. van der Vegt. Accuracy, resolution, and computational complexity of a discontinuous Galerkin finite element method. In *this volume*, 1999.
168. B. van Leer. Towards the ultimate conservation difference scheme, II. *J. Comput. Phys.*, 14:361–376, 1974.
169. T. Warburton. Application of the discontinuous Galerkin method to Maxwell's equations using unstructured polymorphic *hp*-finite elements. In *this volume*, 1999.
170. T.C. Warburton. *Spectral/hp methods on polymorphic multi-domains: Algorithms and Applications.* PhD thesis, Brown University, 1998.
171. T.C. Warburton and G.E. Karniadakis. A discontinuous Galerkin method for the viscous MHD equations. *J. Comput. Phys.*, 152:1–34, 1999.
172. T.C. Warburton, I. Lomtev, R.M. Kirby, and G.E. Karniadakis. A discontinuous Galerkin method for the Navier-Stokes equations in hybrid grids. In M. Hafez and J.C. Heinrich, editors, *10th. International Conference on Finite Elements in Fluids, Tucson, Arizona*, 1998.
173. L.C. Wellford, Jr. and J.T. Oden. A theory of discontinuous finite element approximations for the analysis of shock waves in nonlinear elastic materials. *J. Comput. Phys.*, 19:179–210, 1975.
174. L.C. Wellford, Jr. and J.T. Oden. A theory of discontinuous finite element approximations for of shock waves in nonlinear elastic solids: Variational theory. *Comput. Methods Appl. Mech. Engrg.*, 8:1–16, 1976.

175. L.C. Wellford, Jr. and J.T. Oden. A theory of discontinuous finite element approximations for of shock waves in nonlinear elastic solids: Accuracy and convergence. *Comput. Methods Appl. Mech. Engrg.*, 8:17–36, 1976.
176. M.F. Wheeler. An elliptic collocation-finite element method with interior penalties. *SIAM J. Numer. Anal.*, 15:152–161, 1978.
177. M. Wierse. A new theoretically motivated higher order upwind scheme on unstructured grids of simplices. *Adv. Comput. Math.*, 7:303–335, 1997.
178. P. Woodward and P. Colella. The numerical simulation of two-dimensional fluid flow with strong shocks. *J. Comput. Phys.*, 54:115–173, 1984.
179. L. Yin, A. Acharya, N. Sobh, R.B.Haber, and D.A.Tortorelli. A space-time discontinuous Galerkin method for elastodynamic analysis. In *this volume*, 1999.
180. A. Zhou and Q. Lin. Optimal and superconvergence estimates of the finite element method for a scalar hyperbolic equation. *Acta Math. Sci.*, 14:90–94, 1994.

Part II

Invited Papers

Steps Toward a Robust High-Order Simulation Tool for Aerospace Applications

Harold L. Atkins

NASA Langley Research Center
Hampton, VA 23681

Abstract. The discontinuous Galerkin method is seemingly immune to many of the problems that commonly plague high-order finite-difference methods, and as such, has the potential to bring the robustness of low-order methods and the efficiency of high-order methods to bear on a broad class of engineering problems. However the dependence of the method on numerical quadrature has significantly increased the cost of the method and limited the use of the method to element shapes for which quadrature formulas are readily available. A quadrature-free formulation has been proposed that allows the discontinuous Galerkin method to be implemented for any element shape and for polynomial basis functions of any degree.

1 Introduction

A wide variety of high-order methods have long been used for the detailed analysis of physical problems. Like most good experiments, such detailed analyses are usually performed for a highly simplified case in order to create a controlled environment. The controlled environment allows a better understanding of the physical mechanisms at work, and allows cause – effect relationships to be clearly identified. However, in too many cases, the primary driver on the road to simplification is the need to eliminate influences that may adversely affect the numerical accuracy of the simulation or to contain the cost of the calculation.

In the aerospace engineering community, the majority of experiments, both conventional and numerical, are performed to evaluate the performance of a particular vehicle or vehicle component. In this situation, there is a limit to the degree of simplification that can be made. Although high-order methods offer the potential of providing accurate and inexpensive solutions, low-order methods are most often used. These real world problems usually introduce features that render high-order methods ineffective if not totally unusable. More often then not, the difficulty is traced to the grid on which the computation is performed; however, boundary conditions can also be a factor. Although low-order methods are less accurate, they are robust and reliable. Engineers needing results are more willing to use a low-order method on the finest mesh they can afford, than use a high-order method on a coarser mesh and risk not getting any usable results.

The discontinuous Galerkin method has only recently been applied in the computational fluid dynamics community[1–4]; however, the method has shown itself to be immune to many of the problems that plague traditional high-order finite-difference and finite-volume methods. In particular, the discontinuous Galerkin method is readily applied to unstructured grids, which are easily generated about even the most complex geometries. As a bonus, the discontinuous Galerkin method has been shown to be insensitive to the smoothness of the mesh. Finally, the compact formulation of the discontinuous Galerkin can be applied in the neighborhood of a boundary without modification, which greatly increases the robustness and accuracy of any boundary condition implementation. As such, the discontinuous Galerkin method has the potential to bring the efficiency and accuracy of high-order methods enjoyed by the research community to the broader engineering community where robustness is a necessity.

Although the discontinuous Galerkin method is less susceptible to problems that commonly plague finite-difference schemes, it is not without a few weaknesses. The method has been labeled as expensive, and although in principle it is readily applied to any element shape, it is most often used with quadrilateral and hexahedral elements. Both problems arise from the use of numerical quadrature for the evaluation of integrals contained in the formulation. In an effort to remove these restrictions and provide a more efficient implementation, Atkins et.al.[4] introduced the quadrature free form of the discontinuous Galerkin method. This article reviews the quadrature-free formulation and implementation, and provides examples of its use for linear wave propagation.

1.1 Notation and Numerical Formulation

Given an equation of the form

$$\frac{\partial Q}{\partial t} + \nabla \cdot \boldsymbol{F}(Q) = S(Q) \ , \tag{1}$$

we write the method for a single arbitrary element Ω in terms of a coordinate system (ξ, η, ζ) that is local to that element. The discontinuous Galerkin method is obtained by approximating the solution in Ω in terms of an appropriate local set of basis functions $Q \approx V = \sum_{k=1}^{N} v_k b_k$, where $\{b_k \mid 1 \leq k \leq N\}$, and by performing a local integral projection of equation (1) onto each basis function in the set. The coefficients of the approximate solution v_k are the new unknowns, and the local integral projection generates a set of equations governing these unknowns. The projection equations are expressed in the weak form

$$\int_{\Omega} \left(b_k \frac{\partial V}{\partial t} - \nabla b_k \cdot \mathbf{J}^{-1} \boldsymbol{F}(V) \right) J d\Omega + \sum_{e=1}^{E} \int_{\partial \Omega_e} F_e^{\mathrm{R}}(\bar{V}, \bar{W}) ds = \int_{\Omega} b_k S(V) J d\Omega \tag{2}$$

where E denotes the number of edges, $\bar{V}$ denotes the trace of V on edge e, $\bar{W}$ denotes the trace of the approximate solution in the neighboring element on edge e, $F_e^{\mathrm{R}}(\bar{V}, \bar{W})$ denotes an approximate Riemann flux,

$$\mathbf{J} = \frac{\partial(x, y, z)}{\partial(\xi, \eta, \zeta)}, \quad \text{and} \quad J = |\mathbf{J}| \ .$$

The term edge will be used to refer to any segment of an element boundary that is shared with a neighboring element or with the physical boundary of the domain. In the present work, the basis functions are polynomials of the form $\xi^l \eta^m \zeta^n$ for $l+m+n \leq p$, and the approximate Riemann flux is modeled by a simple Lax-Friedrichs flux of the form $F^{\mathrm{R}}(V, W) \equiv [F(V) + F(W) - \lambda(W - V)]/2$ where λ is greater than the maximum absolute eigenvalue of the Jacobian $\partial F/\partial V$.

Implementation by Numerical Quadrature. The discrete form of equation (2) is usually obtained by evaluating the integrals using numerical quadrature formulas of the required order[5]. Although this approach is simple and straightforward, the associated computational cost is high, and the lack of suitable quadrature formulas has restricted practical applications to simple elements or relatively low order solution approximations. Numerical quadrature is most efficient when the unknown variables are stored at the quadrature points (e.g., a collocation method); however, in the implementation of the discontinuous Galerkin method, the common practice is to store the coefficients of the solution expansion, as described above. As such, an order-N operation is required at each of the N_q quadrature points simply to obtain the data required to evaluate the numerical quadrature.

Optimal quadrature formulas are not generally available for arbitrary elements shapes of arbitrary order, and this has restricted the application of the discontinuous Galerkin method. Tensor products of one-dimensional quadrature formulas can be used to integrate quadrilateral and hexahedral elements to any required degree. However, as seen in table 1, the number of terms in the quadrature summation N_q exceeds the number of unknowns N by a considerable margin. For instance, for $p = 4$ in three dimensions $N_q/N > 3.5$.

Table 1. Variation of N_q/N with p

p	0	2	4	10	∞
N_q/N for 2D	1	1.5	1.666	1.833	2.0
N_q/N for 3D	1	2.7	3.571	4.654	6.0

Dubiner[6] introduced a procedure in which triangles and tetrahedrons are mapped into quadrilaterals and hexahedrons such that tensor products of one-dimensional quadrature formulas can be applied. This approach was further extended and applied to finite-element formulations by Sherwin and Karniadakis[7]. Aside from Dubiner's approach, there is not general procedure for deriving quadrature formulas of arbitrary order for elements such as triangles or tetrahedrons. For these general elements, near optimal quadrature formulas have been computed numerically and tabulated for only a limited range of p. This has restricted most implementations of the discontinuous Galerkin method to quadrilateral, hexahedral, or relatively low-order triangular elements.

Quadrature-Free Implementation. The quadrature-free approach[4] was developed to circumvent this difficulty and to allow the discontinuous Galerkin method to be easily and efficiently implemented on general unstructured grids to any order of accuracy. To implement the quadrature-free approach, the fluxes and sources are also written as an expansion in terms of the basis functions:

$$\boldsymbol{F}(Q) \approx \sum_{j=1}^{M} \boldsymbol{f}_j(V) b_j, \qquad S \approx \sum_{j=1}^{M} s_j b_j$$

When $\boldsymbol{F}(Q)$ is a linear function of Q, then $M = N$, and the expansion is trivial and exact. When the flux is a non-linear or a linear but non-constant coefficient function of Q, then the degree of the flux expansion must be at least $p+1$ and M will be greater than N. The same comment applies to the source term S except that the source expansion may be truncated to degree p.

Similar treatment of the approximate Riemann flux is only slightly more complex due to the fact that the solutions on either side of an interior edge are defined in terms different coordinate systems. However, this difficulty is easily resolved by rewriting the trace of V and W on edge e in terms of a coordinate system $(\bar{\xi}_e, \bar{\eta}_e)$ associated with the edge. That is, $\bar{V}_e = \sum_{k=1}^{N} \bar{v}_{e,k} \bar{b}_{e,k}$, and $\bar{W}_e = \sum_{k=1}^{N} \bar{w}_{e,k} \bar{b}_{e,k}$, where $\bar{b}_{e,k}$ denotes a basis function associated with the edge coordinate system on edge e. This, of course, is just a coordinate transformation, and the coefficients $\bar{v}_e$ are easily computed from v by a linear matrix operator $[\bar{v}_e] = \mathbf{T}_e [v]$. The trace of the flux can be computed either by taking the trace of the volume flux, $[\bar{f}_e] = \mathbf{T}_e \left[\mathbf{J}^{-1} J \boldsymbol{f}_j \cdot \boldsymbol{n}\right]$, or by recomputing the flux from the trace of the solution. Generally, the later is prefered for a linear problem. Now the approximate Riemann flux can be expanded in terms of $\bar{b}_{e,k}$ as

$$F_e^{\mathrm{R}} \equiv \sum_{k=1}^{M} f_{e,k}^{\mathrm{R}} \bar{b}_k = \sum_{k=1}^{M} \left[\bar{f}_{e,k}(\bar{v}) + \bar{f}_{e,k}(\bar{w}) - \lambda \left(\bar{w}_{e,k} - \bar{v}_{e,k}\right)\right] \bar{b}_{e,k}/2$$

without regard for the type of element or the orientation of the coordinate system of the adjacent elements.

Now that the functional forms of the solution, source, and fluxes are explicit, the integrals are analytically evaluated to give

$$\frac{\partial [v_k]}{\partial t} + \mathbf{A} \cdot \mathbf{J}^{-1} J [f_j] + \sum_{e=1}^{E} \mathbf{B}_e [f^{\mathrm{R}}_{e,j}] = [s_k], \tag{3}$$

where

$$\mathbf{A} = \mathbf{M}^{-1} \left[\int_{\Omega} b_j \nabla b_k d\Omega \right], \quad \mathbf{B} = \mathbf{M}^{-1} \left[\int_{\partial\Omega} b_k \bar{b}_j ds \right], \text{ and } \mathbf{M} = \left[\int_{\Omega} b_k b_j d\Omega \right].$$

The matrices $\mathbf{M}$, $\mathbf{A}$, and $\mathbf{B}_e$ depend only on the type of the computational element and the degree of the approximate solution p. Thus, the set of matrices associated with a particular type of computational element can be precomputed and applied in all elements that map to it. Finally, as a result of symmetry in the computational elements, the matrices $\mathbf{A}$, $\mathbf{B}_e$, and $\mathbf{T}_e$ are sparse such that the work required to evaluate the integrals is proportional to $N/2$ instead of 3 to 6 times N.

Discussion

Though clearly the quadrature-free approach is ideally suited for linear problems, the approach has been applied to non-linear problems. Previous work[8] suggest this approach may offer some advantages with regard to shock capturing; however, this is beyond the scope of this article. In the following sections, the quadrature-free approach applied to the solution of the linear Euler equations for acoustic wave propagation. The first section describes the treatment of curved wall; the second section presents a three-dimensional simulation using a tetrahedral mesh. Detailed validation studies can be found in references [4,8–10].

Modeling of Curved Walls. Curved walls can be modeled with at least second-order accuracy by straight line segments. However, a high-order treatment of a curved wall is easily supported by allowing the edge of the computational element to be curved as shown in figure 1. The matrices $\mathbf{A}$, $\mathbf{B}_e$, and $\mathbf{T}_e$ can still be evaluated exactly and in advance of the simulation as in the usual implementation; however, each element adjacent to a curved wall has a distinct set of matrices, and the $\mathbf{A}$ and $\mathbf{B}_e$ matrices are full. Experience indicates that the number of curved sided elements required is a small percentage of the total number of elements, thus the additional storage is not significant.

Figure 2 shows two solutions in which an acoustic pulse, originating from a point 6 radii from a cylinder, has passed over the cylinder to produce

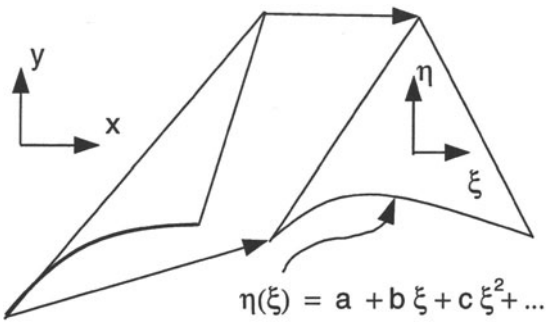
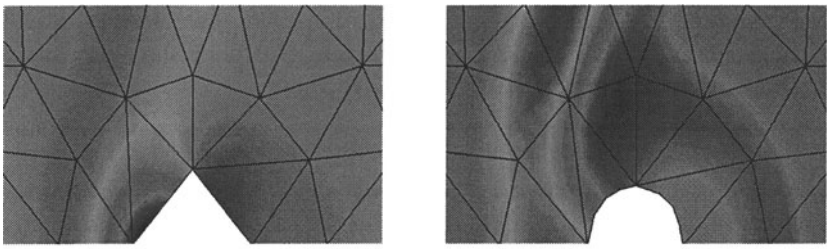


Fig. 1. Mapping for curved wall element.

a reflection. In the extreme case shown, the half cylinder is modeled with only two elements. In figure 2(b), the curved sides are approximated by cubic polynomials, and the form of the reflection is clearly improved. Figure 3 shows the convergence of the solution as the average mesh spacing ΔS is reduced. The L_2-norm of the difference in pressure, relative to a reference solution computed on a fine grid ($\Delta S = 0.0498$), is measured at a large number of points uniformly distributed in the region $0.63 < r < 2.0$, $0 < \theta < \pi/2$. The case with the cubic approximation for the wall maintains a fifth-order rate of convergence over the range of grids tested. The rate of convergence for the case with the linear approximation for the wall drops to less than third order as the mesh is refined.



a. Linear wall segments. b. Cubic wall segments.

Fig. 2. Reflection of cylindrical pressure pulse off of solid cylinder.

Three Dimensional Simulations. A computation is presented for the propagation and reflection of an acoustic pulse from a blended-wing-body configuration. A tetrahedron mesh having 78048 elements was generated by

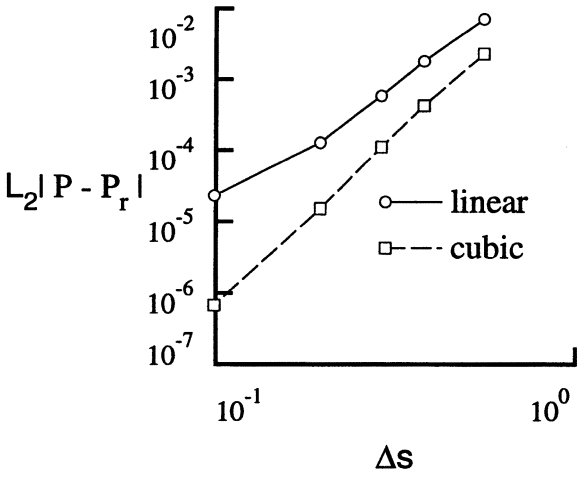


Fig. 3. Convergence of solution error with grid refinement.

AFLR3[11]. The simulation was performed using fifth-order elements ($p = 4$). Figure 4 shows the surface mesh and several planes and a line on which results were examined. A compact initial pressure disturbance of the form

$$p = \begin{cases} \cos^2(\pi R/8) & \text{for } R \leq 4 \\ 0 & \text{for } R > 4 \end{cases},$$

where $R = \sqrt{(x-140)^2 + (y+10)^2 + (z-5)^2}$, creates a transient wave that is propagated to a time of $t = 160$. The surface mesh was generated with a target edge length of five, so the width of the initial pulse is less than two edge lengths. Figure 5(a) shows the solution along line "A" at time $t = 0$, 50, 100, and 150 with an appropriate scaling. The final two sample times are in close agreement which indicates the wave is propagating at the correct speed and decaying at the correct rate. This is consistent with mesh refinement studies[10] on model problems in which it is shown that a resolution of two edge lengths per wave length is sufficient to propagate a wave 100 wavelengths with less than 3% error. Figure 5(b) shows the solution on a spanwise cut through the wing and winglet at time $t = 140$. Though qualitative, the waves propagate and reflect in the expected manner.

Summary

The quadrature-free approach allows the discontinuous Galerkin method to be implemented for any element shape, and for polynomial basis functions of any degree. In three dimensions and for $p \geq 4$, the work required to evaluate the integrals is 7 to 12 times less than that required by tensor products of

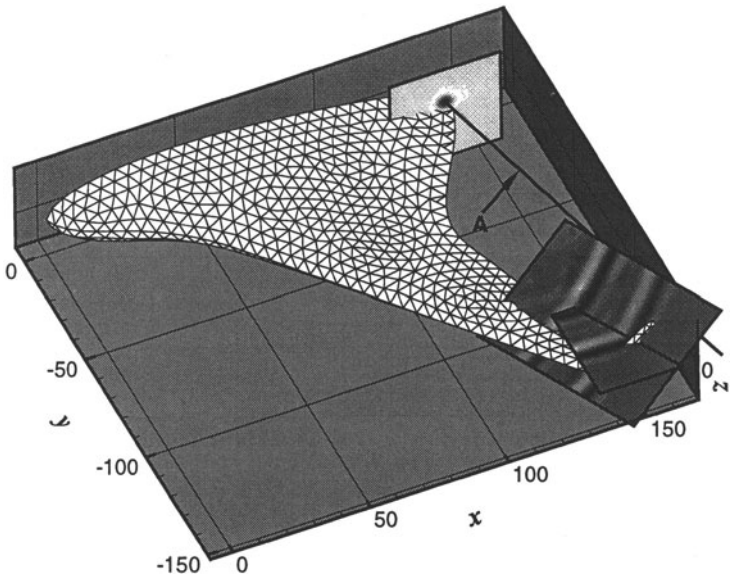


Fig. 4. Generic blended-wing-body configuration with planes and lines on which data is examined.

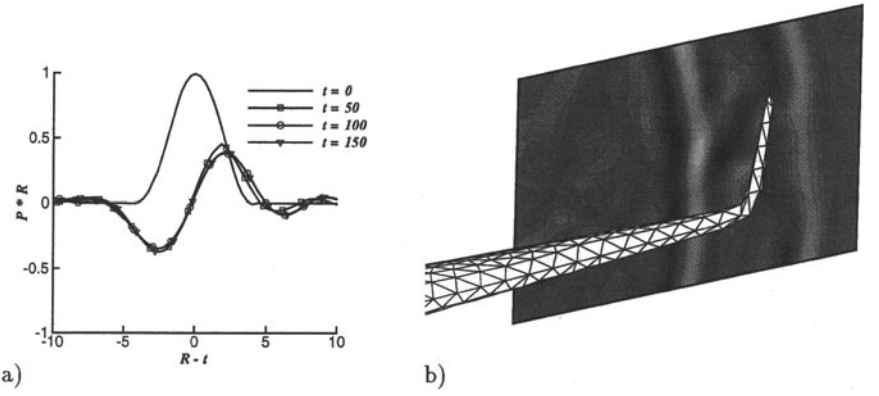


Fig. 5. Perturbation pressure: (a) on line "A" scaled to show the $1/r$ decay, (b) on spanwise cut through wing and winglet at $t = 140$.

one-dimensional quadratures. Though ideally suited for linear problems, the method has been applied to non-linear problems, including those with discontinuities. The approach has been used to perform time accurate simulations of acoustic propagation and scatter about full scale aircraft configurations using general unstructured grids.

References

1. Bassi, F. and Rebay, S., "Accurate 2D Euler Computations by Means of a High-Order Discontinuous Finite Element Method," *Proceedings of the Conference Held in Bangalore, India*, Lecture Notes in Physics, Springer, October 1995, pp. 234–240.
2. Bassi, F. and Rebay, S., "Discontinuous Finite Element High Order Accruate Numerical Solution of the Compressible Navier-Stokes Equations," *Numerical Methods for Fluid Dynamics*, Clarendon, Oxford, England, UK, 1995, pp. 295–302.
3. Lowrie, R. B., Roe, P. L., and van Leer, B., "A Space-Time Discontinuous Galerkin Method for the Time-Accurate Numerical Solution of Hyperbolic Conservation Laws," AIAA Paper-95-1658, 1995.
4. Atkins, H. L. and Shu, C.-W., "Quadrature-Free Implementation of Discontinuous Galerkin Method for Hyperbolic Equations," *AIAA Journal*, Vol. 36, No. 5, 1998, pp. 775–782.
5. Cockburn, B., Hou, S., and Shu, C.-W., "TVB Runge-Kutta Local Projection Discontinuous Galerkin Finite Element Method for Conservation Laws IV: The MultiDimensional Case," *Mathematics of Computation*, Vol. 54, No. 190, 1990, pp. 545–581.
6. Dubiner, M., "Spectral Methods on Triangles and Other Domains," *Journal of Scientific Computing*, Vol. 6, 1991, pp. 345.
7. Sherwin, S. J. and Karniadakis, G. E., "A New Triangular and Tetrahedral Basis for High-Order (*hp*) Finite Element Methods," *International Journal for Numerical Methods in Engineering*, Vol. 38, 1999, pp. 3775 – 3802.
8. Atkins, H. L., "Local Analysis of Shock Capturing Using Discontinuous Galerkin Methodology," AIAA Paper 97-2032, 1997, 13th AIAA Computational Fluid Dynamics Conference, Snowmass Village, Colorado, June 29-July 2.
9. Atkins, H. L., "Continued Development of the Discontinuous Galerkin Method for Computational Aeroacoustic Applications," AIAA Paper-97-1581, 1997, Third Joint CEAS/AIAA Aeroacoustics Conference, May 12-14.
10. Atkins, H. L. and Lockard, D. P., "A High-Order Method Using Unstructured Grids for the Aeroacoustic Analysis of Realistic Aircraft Configurations," AIAA Paper-99-1945, 1999, Fifth Joint AIAA/CEAS Aeroacoustics Conference, May 10-12.
11. Marcum, D. L., "Generation of Unstructured Grids for Viscous Flow Applications," AIAA Paper-95-0212, 1995.

Simplified Discontinuous Galerkin Methods for Systems of Conservation Laws with Convex Extension

Timothy J. Barth

NASA Ames Research Center, Information Sciences Directorate, Moffett Field, CA 94035, USA
barth@nas.nasa.gov

Abstract. Simplified forms of the space-time discontinuous Galerkin (DG) and discontinuous Galerkin least-squares (DGLS) finite element method are developed and analyzed. The new formulations exploit simplifying properties of entropy endowed conservation law systems while retaining the favorable energy properties associated with symmetric variable formulations.

Introduction

The high-order accurate numerical solution of systems of nonlinear conservation laws remains a computationally expensive endeavor. This article considers simplified forms of the Discontinuous Galerkin (DG) and Discontinuous Galerkin least-squares (DGLS) finite element methods tailored to systems of first-order nonlinear conservation laws with convex entropy extension. Central to the development is the Eigenvalue Scaling Theorem which characterizes right symmetrizers of entropy endowed systems of conservation laws in terms of scaled eigenvectors of the corresponding flux Jacobian matrices. This yields a simplification of the DG and DGLS methods without sacrificing the pleasing energy properties of symmetric variable formulations. The next section briefly reviews a number of results in symmetrization theory discussed in detail in Barth [2,1].

1 Brief Review of Symmetrization and the Eigenvector Scaling Theorem

Consider a system of m coupled first-order differential equations in d space coordinates and time which represents a conservation law process. Let $\boldsymbol{u}(x,t) : \mathbb{R}^d \times \mathbb{R}^+ \mapsto \mathbb{R}^m$ denote the dependent solution variables and $\boldsymbol{f}(\boldsymbol{u}) : \mathbb{R}^m \mapsto \mathbb{R}^{m\times d}$ the flux vector. The prototype conservation system is then given by

$$\boldsymbol{u}_{,t} + \boldsymbol{f}^i_{,x_i} = 0 \tag{1.1}$$

with implied summation on the index i. Additionally, the system is assumed to possess an scalar entropy extension. Let $U(\boldsymbol{u}) : \mathbb{R}^m \mapsto \mathbb{R}$ denote an

entropy function and $F(\boldsymbol{u}) : \mathbb{R}^m \mapsto \mathbb{R}^d$ the entropy flux such that in addition to (1.1) the following inequality holds

$$U_{,t} + F^i_{,x_i} \leq 0 \tag{1.2}$$

with equality for smooth solutions. In symmetrization theory for first-order conservation laws, one seeks a mapping $\boldsymbol{u}(\boldsymbol{v}) : \mathbb{R}^m \mapsto \mathbb{R}^m$ applied to (1.1) so that when transformed

$$\boldsymbol{u}_{,\boldsymbol{v}}\boldsymbol{v}_{,t} + \boldsymbol{f}^i_{,\boldsymbol{v}}\boldsymbol{v}_{,x_i} = 0 \tag{1.3}$$

the matrix $\boldsymbol{u}_{,\boldsymbol{v}}$ is symmetric positive definite (SPD) and the matrices $\boldsymbol{f}^i_{,\boldsymbol{v}}$ are symmetric. Clearly, if functions $\mathcal{U}(\boldsymbol{v}), \mathcal{F}^i(\boldsymbol{v}) : \mathbb{R}^m \mapsto \mathbb{R}$ can be found so that

$$\boldsymbol{u}^T = \mathcal{U}_{,\boldsymbol{v}}, \quad (\boldsymbol{f}^i)^T = \mathcal{F}^i_{,\boldsymbol{v}} \tag{1.4}$$

then the matrices

$$\boldsymbol{u}_{,\boldsymbol{v}} = \mathcal{U}_{,\boldsymbol{v},\boldsymbol{v}}, \quad \boldsymbol{f}^i_{,\boldsymbol{v}} = \mathcal{F}^i_{,\boldsymbol{v},\boldsymbol{v}} \tag{1.5}$$

are symmetric. Since $\boldsymbol{v}$ is not yet known, little progress has been made. Introducing the following duality relationships

$$U(\boldsymbol{u}) = \boldsymbol{v}^T(\boldsymbol{u})\,\boldsymbol{u} - \mathcal{U}(\boldsymbol{v}(\boldsymbol{u})) \tag{1.6}$$

$$F^i(\boldsymbol{u}) = \boldsymbol{v}^T(\boldsymbol{u})\,\boldsymbol{f}^i(\boldsymbol{u}) - \mathcal{F}^i(\boldsymbol{v}(\boldsymbol{u})) \tag{1.7}$$

followed by differentiation yields

$$U_{,\boldsymbol{u}} = \boldsymbol{v}^T + \boldsymbol{u}^T\boldsymbol{v}_{,\boldsymbol{u}} - \mathcal{U}_{,\boldsymbol{v}}\boldsymbol{v}_{,\boldsymbol{u}} = \boldsymbol{v}^T \tag{1.8}$$

$$F^i_{,\boldsymbol{u}} = \boldsymbol{v}^T\boldsymbol{f}^i_{,\boldsymbol{u}} + (\boldsymbol{f}^i)^T\boldsymbol{v}_{,\boldsymbol{u}} - \mathcal{F}_{,\boldsymbol{v}}\boldsymbol{v}_{,\boldsymbol{u}} = \boldsymbol{v}^T\boldsymbol{f}^i_{,\boldsymbol{u}} \ . \tag{1.9}$$

Equation (1.8) gives an explicit expression for the entropy variables $\boldsymbol{v}$ in terms of derivatives of the entropy function $U(\boldsymbol{u})$

$$\boldsymbol{v}^T = U_{,\boldsymbol{u}} \ . \tag{1.10}$$

Finally, we require convexity of $U(\boldsymbol{u})$ which insures positive definiteness of $\boldsymbol{v}_{,\boldsymbol{u}}$ and $\boldsymbol{u}_{,\boldsymbol{v}}$ and implies hyperbolicity of (1.1) [5,9], viz., that the linear combination $\boldsymbol{f}_{,\boldsymbol{u}}(\boldsymbol{n}) = n_i\,\boldsymbol{f}^i_{,\boldsymbol{u}}$ has real eigenvalues and a complete set of real-valued eigenvectors for all nonzero $\boldsymbol{n} \in \mathbb{R}^d$. This result follows immediately from the identity

$$(\boldsymbol{u}_{,\boldsymbol{v}})^{-1/2}\boldsymbol{f}_{,\boldsymbol{u}}(\boldsymbol{n})(\boldsymbol{u}_{,\boldsymbol{v}})^{1/2} = \underbrace{(\boldsymbol{u}_{,\boldsymbol{v}})^{-1/2}\boldsymbol{f}_{,\boldsymbol{v}}(\boldsymbol{n})(\boldsymbol{u}_{,\boldsymbol{v}})^{-1/2}}_{\text{symm}}$$

since $\boldsymbol{f}_{,\boldsymbol{u}}(\boldsymbol{n})$ is similar to a symmetric matrix.

1.1 The Eigenvector Scaling Theorem

Next, we consider an important algebraic property of right symmetrizable systems which is used later in the implementation of the DG and DGLS schemes. Simplifying upon the previous notation, let $\tilde{A}_0 = \boldsymbol{u}_{,\boldsymbol{v}}$, $A_i = \boldsymbol{f}^i_{,\boldsymbol{v}}$, $\tilde{A}_i = A_i \tilde{A}_0$ and rewrite (1.3)

$$\tilde{A}_0 \boldsymbol{v}_{,t} + \tilde{A}_i \boldsymbol{v}_{,x_i} = 0 \ . \tag{1.11}$$

The following theorem states a property of the symmetric matrix $\tilde{A}_i$ symmetrized via the symmetric positive definite matrix $\tilde{A}_0$.

Theorem 1.1 (Eigenvector Scaling). *Let $A \in \mathbb{R}^{n\times n}$ be an arbitrary diagonalizable matrix and S the set of all right symmetrizers:*

$$S = \{B \in \mathbb{R}^{n\times n} \mid B \ \text{SPD}, \ AB \ \text{symmetric}\}.$$

Further, let $R \in \mathbb{R}^{n\times n}$ denote the right eigenvector matrix which diagonalizes A

$$A = R\Lambda R^{-1}$$

with r distinct eigenvalues, $\Lambda = \text{Diag}(\lambda_1 I_{m_1\times m_1}, \lambda_2 I_{m_2\times m_2}, \ldots, \lambda_r I_{m_r\times m_r})$. Then for each $B \in S$ there exists a symmetric block diagonal matrix $T = \text{Diag}(T_{m_1\times m_1}, T_{m_2\times m_2}, \ldots, T_{m_r\times m_r})$ that block scales columns of R, $\tilde{R} = RT$, such that

$$B = \tilde{R}\tilde{R}^T, \quad A = \tilde{R}\Lambda\tilde{R}^{-1}$$

which imply that

$$AB = \tilde{R}\Lambda\tilde{R}^T.$$

Proof. Omitted, see [2]. □

Note that this last formula states a congruence relationship since $\tilde{R}$ is not generally orthonormal and Λ does not represent the eigenvalues of AB. The Eigenvalue Scaling Theorem is a variant of the well known theory developed for the commuting matrix equation $AX - XA = 0$, $A, X \in \mathbb{R}^{n\times n}$, see for example Gantmacher [6]. Examples of the Eigenvector Scaling Theorem for the Euler and magnetohydrodynamic equations are given in [2].

1.2 Generalized Matrix Functions with Respect to the Riemannian Matrix $\tilde{A}_0$

For use in later developments, it is useful to define the following generic matrix function $f(\tilde{A})$ with respect to the Riemannian matrix $\tilde{A}_0$

$$f_{\tilde{A}_0}(\tilde{A}) \equiv \tilde{A}_0 f(\tilde{A}_0^{-1}\tilde{A}) \ . \tag{1.12}$$

This definition reflects the following steps: (1) multiplication of the system (1.11) by $\tilde{A}_0^{-1}$ in order to restore a Euclidean metric, (2) invocation of the

matrix function on the matrix product $\tilde{A}_0^{-1}\tilde{A}_i$, (3) multiplication of the result by $\tilde{A}_0$ to restore the original metric matrix. Proposition 1.2 shows that this generalized matrix function is symmetric and has a rather simple construction for symmetrizable systems by exploiting the Eigenvalue Scaling Theorem.

Proposition 1.2. *Barth [2,1]. Let $\tilde{A}_0$ denote the SPD right symmetrizer of A such that $\tilde{A} = A\tilde{A}_0$ $\tilde{A}_0 = \tilde{R}\tilde{R}^T$, and $A = \tilde{R}\Lambda\tilde{R}^{-1}$. The generalized matrix function $f_{\tilde{A}_0}(\tilde{A})$ is symmetric and defined canonically in terms of entropy scaled eigenvectors as*

$$f_{\tilde{A}_0}(\tilde{A}) = \tilde{R}f(\Lambda)\tilde{R}^T \ . \tag{1.13}$$

In later sections, the generalized matrix absolute value function $|\tilde{A}|_{\tilde{A}_0}$ will be required

$$|\tilde{A}|_{\tilde{A}_0} = \tilde{R}|\Lambda|\tilde{R}^T \ . \tag{1.14}$$

This matrix absolute value function has a natural generalization to $\mathbb{R}^d$ using an L_p-like norm definition

$$|\tilde{A}|_{p,\tilde{A}_0} = \left(\sum_{i=1}^{d} |A_i|^p\right)^{1/p} \tilde{A}_0 \tag{1.15}$$

which has a particularly simple form when $p = 1$ which is used later in the least-squares term appearing in the DGLS method

$$|\tilde{A}|_{1,\tilde{A}_0} = \sum_{i=1}^{d} \tilde{R}_i \, |\Lambda_i| \, \tilde{R}_i^T \ . \tag{1.16}$$

2 Simplified DG, DGLS, and GLS Finite Element Methods

Let Ω denote a spatial domain composed of nonoverlapping elements T_i, $\Omega = \cup T_i$, $T_i \cap T_j = \emptyset$, $i \neq j$ and $I^n =]t^n, t^{n+1}[$ the n-th time interval. It is useful to also define the element set $\mathcal{T} = \{T_1, T_2, \ldots, T_{|\mathcal{T}|}\}$ and edge set $\mathcal{E} = \{e_1, e_2, \ldots, e_{|\mathcal{E}|}\}$. To simplify the exposition, consider a single variational formulation with weakly enforced boundary conditions. By choosing the correct space of functions (discontinuous or continuous) and omitting the least-squares variational term, one can switch from the DGLS formulation to the DG or GLS formulations. In the GLS formulation [7,10], functions are continuous in space and discontinuous in time

$$\mathcal{V}^h = \left\{ v^h \,|\, v^h \in \left(C^0(\Omega \times I^n)\right)^m, v^h_{|T\times I^n} \in \left(\mathcal{P}_k(T \times I^n)\right)^m \right\}$$

where v denotes the entropy variables for the system. In the DG and DGLS formulations [8,3], functions are discontinuous in space and time, i.e.

$$\mathcal{V}^h = \left\{ v^h \,|\, v^h_{|T\times I^n} \in \left(\mathcal{P}_k(T \times I^n)\right)^m \right\} \ .$$

Consider the prototype hyperbolic system for the space-time domain $\Omega \times [0,T]$ with boundary data $\boldsymbol{g}$ imposed on Γ via admissibility condition

$$\begin{aligned} \boldsymbol{u}_{,t} + \boldsymbol{f}^i_{,x_i} &= 0 \quad \text{in } \Omega \\ A^-(\boldsymbol{n})\,(\boldsymbol{g} - \boldsymbol{u}) &= 0 \quad \text{on } \Gamma \end{aligned} \tag{2.1}$$

or in symmetric quasi-linear form for smooth solutions

$$\begin{aligned} \tilde{A}_0 \boldsymbol{v}_{,t} + \tilde{A}_i \boldsymbol{v}_{,x_i} &= 0 \quad \text{in } \Omega \\ \tilde{A}^-(\boldsymbol{n})\,(\tilde{\boldsymbol{g}} - \boldsymbol{v}) &= 0 \quad \text{on } \Gamma \end{aligned} \tag{2.2}$$

with $A(\boldsymbol{n}) = \boldsymbol{n}_i A_i$ and $\tilde{A}(\boldsymbol{n}) = \boldsymbol{n}_i \tilde{A}_i$. The combined GLS and DG schemes are defined by the following stabilized variational formulation:

Find $\boldsymbol{v}^h \in \mathcal{V}^h$ such that for all $\boldsymbol{w}^h \in \mathcal{V}^h$

$$B(\boldsymbol{v}^h, \boldsymbol{w}^h)_{\text{GAL}} + B(\boldsymbol{v}^h, \boldsymbol{w}^h)_{\text{LS}} + B(\boldsymbol{v}^h, \boldsymbol{w}^h)_{\text{BC}} = 0 \tag{2.3}$$

$$\begin{aligned} B(\boldsymbol{v}, \boldsymbol{w})_{\text{GAL}} &= \int_{I^n}\int_{\Omega} (-\boldsymbol{u}(\boldsymbol{v}) \cdot \boldsymbol{w}_{,t} - \boldsymbol{f}^i(\boldsymbol{v}) \cdot \boldsymbol{w}_{,x_i})\, dx\, dt \\ &\quad + \int_{\Omega} \left(\boldsymbol{w}(t^{n+1}_-) \cdot \boldsymbol{u}(\boldsymbol{v}(t^{n+1}_-)) - \boldsymbol{w}(t^n_+) \cdot \boldsymbol{u}(\boldsymbol{v}(t^n_-)) \right) dx \\ &\quad + \int_{I^n} \sum_{e \in \mathcal{E}} \int_e (\boldsymbol{w}(x_-) - \boldsymbol{w}(x_+)) \cdot \boldsymbol{h}(\boldsymbol{v}(x_-), \boldsymbol{v}(x_+); \boldsymbol{n})\, dx\, dt \\ B(\boldsymbol{v}, \boldsymbol{w})_{\text{LS}} &= \int_{I^n} \sum_{T \in \mathcal{T}} \int_T \left(\tilde{A}_0 \boldsymbol{w}_{,t} + \tilde{A}_i \boldsymbol{w}_{,x_i} \right) \cdot \boldsymbol{\tau} \left(\tilde{A}_0 \boldsymbol{v}_{,t} + \tilde{A}_i \boldsymbol{v}_{,x_i} \right) dx\, dt \\ B(\boldsymbol{v}, \boldsymbol{w})_{\text{BC}} &= \int_{I^n}\int_{\Gamma} \boldsymbol{w} \cdot \boldsymbol{h}(\boldsymbol{v}, \tilde{\boldsymbol{g}}; \boldsymbol{n})\, dx\, dt \end{aligned}$$

where $\boldsymbol{h}$ denotes a numerical flux function and $\boldsymbol{\tau}$ a small $m \times m$ SPD matrix for the least-squares term. For theoretical and practical reasons, two numerical flux functions are considered. Both are of the form

$$\boldsymbol{h}(\boldsymbol{v}_-, \boldsymbol{v}_+; \boldsymbol{n}) = \frac{1}{2}\left(\boldsymbol{f}(\boldsymbol{v}_-; \boldsymbol{n}) + \boldsymbol{f}(\boldsymbol{v}_+; \boldsymbol{n})\right) - \frac{1}{2}\boldsymbol{h}^d(\boldsymbol{v}_-, \boldsymbol{v}_+; \boldsymbol{n}) \tag{2.4}$$

and consistent with the true flux in the sense that $\boldsymbol{h}(\boldsymbol{v}, \boldsymbol{v}; \boldsymbol{n}) = \boldsymbol{f}(\boldsymbol{v}; \boldsymbol{n})$.

1. **Symmetric Mean-Value Flux**. This flux is motivated from the nonlinear energy theory of Sect. 2.2. Define the parameterization $\overline{\boldsymbol{v}}(\theta) \equiv \boldsymbol{v}(x_-) + \theta\,[\boldsymbol{v}]^{x_+}_{x_-}$. The symmetric mean-value flux is then given by

$$\boldsymbol{h}^d_{\text{SMV}}(\boldsymbol{v}_-, \boldsymbol{v}_+; \boldsymbol{n}) = |\tilde{A}(\boldsymbol{v}_-, \boldsymbol{v}_+; \boldsymbol{n})|_{\text{SMV}}\,[\boldsymbol{v}]^{x_+}_{x_-}$$

with

$$|\tilde{A}(\boldsymbol{v}_-, \boldsymbol{v}_+; \boldsymbol{n})|_{\text{SMV}} \equiv \int_0^1 |\tilde{A}(\overline{\boldsymbol{v}}(\theta); \boldsymbol{n})|_{\tilde{A}_0}\, d\theta\ . \tag{2.5}$$

By construction, the matrix $|\tilde{A}(\boldsymbol{v}_-, \boldsymbol{v}_+; \boldsymbol{n})|_{\text{SMV}}$ is symmetric positive semi-definite. Using this form of flux dissipation (2.5), nonlinear entropy

norm stability of the DG, DGLS, and GLS formulations can be shown as discussed in Sect. 2.2. In addition, let

$$\tilde{A}(\boldsymbol{n})_{\text{SMV}} = \int_0^1 \tilde{A}(\overline{\boldsymbol{v}}(\theta); \boldsymbol{n})\, d\theta \tag{2.6}$$

denote the integral mean-value matrix for which the following useful property exists

$$[\boldsymbol{f}(\boldsymbol{n})]_{x_-}^{x_+} = \tilde{A}(\boldsymbol{n})_{\text{SMV}}\, [\boldsymbol{v}]_{x_-}^{x_+} \tag{2.7}$$

which is a necessary ingredient for optimal discontinuity resolution. To prove stability of other (more practical) forms of flux dissipation, one formally needs only show that the new form is more energy dissipative than the symmetric mean-value form in the following sense:

$$[\boldsymbol{v}]_{x_-}^{x_+} \cdot \boldsymbol{h}_{\text{SMV}}^d \leq [\boldsymbol{v}]_{x_-}^{x_+} \cdot \boldsymbol{h}^d \ .$$

2. **Discrete Symmetric Mean-Value Flux.** The discrete symmetric mean-value flux function replaces the state-space path integration in (2.5) by N point numerical quadrature

$$\boldsymbol{h}_{\text{DSMV}}^d(\boldsymbol{v}_-, \boldsymbol{v}_+; \boldsymbol{n}) = |\tilde{A}(\boldsymbol{v}_-, \boldsymbol{v}_+; \boldsymbol{n})|_{\text{DSMV}}\, [\boldsymbol{v}]_{x_-}^{x_+}$$

with

$$|\tilde{A}(\boldsymbol{v}_-, \boldsymbol{v}_+; \boldsymbol{n})|_{\text{DSMV}}\, [\boldsymbol{v}]_{x_-}^{x_+} \equiv \sum_{q=1}^{N} w_q\, |\tilde{A}(\overline{\boldsymbol{v}}(\theta_q); \boldsymbol{n})|_{\tilde{A}_0}\ [\boldsymbol{v}]_{x_-}^{x_+} \tag{2.8}$$

where w_q and θ_q denotes the quadrature weights and positions. In forming this flux, recall from the Eigenvalue Scaling Theorem that $|\tilde{A}|_{\tilde{A}_0} = \tilde{R}\,|\Lambda|\,\tilde{R}^T$. This flux function is of practical interest since it is easily formed and has a relatively straightforward Jacobian linearization as will be shown later. The absolute value in this equation renders the state-space path integration from $\boldsymbol{v}_-$ to $\boldsymbol{v}_+$ slope discontinuous whenever entries in Λ change sign. In this case, to retain accuracy of the numerical quadrature at $n \leq m$ possible points of slope discontinuity, the path integration is further subdivided into subintervals, e.g. $[\boldsymbol{v}_-, \boldsymbol{v}_+] = [\boldsymbol{v}_-, \boldsymbol{v}_1^*] \cup [\boldsymbol{v}_1^*, \boldsymbol{v}_2^*] \cup \ldots \cup [\boldsymbol{v}_n, \boldsymbol{v}_+]$ where $\boldsymbol{v}_i^* \equiv \boldsymbol{v}(\theta_i)$ is a location θ_i such that an entry of Λ vanishes. In practice, satisfactory results [1] have been obtained using 2-point Gaussian quadrature rules (which integrate cubic polynomials exactly).

2.1 Linear Energy Analysis

Due to length constraints of this article, we simply restate some relevant theorems given in Barth [2,1] concerning energy boundedness of variational form (2.3) for systems of hyperbolic equations.

Theorem 2.1. Global Energy Stability (Linear Hyperbolic System). *The variational formulation (2.3) for linear hyperbolic systems is energy stable (modulo data $\tilde{g}$) with the following global energy balance:*

$$\sum_{n=0}^{N-1}\left(\|[\![v]\!]_{t_-^n}^{t_+^n}\|^2_{\tilde{A}_0,\Omega} + 2\|\tilde{A}_0 v_{,t} + \tilde{A}_i v_{,x_i}\|^2_{\mathcal{T},\Omega\times I^n} + \sum_{e\in\mathcal{E}}\langle [\![v]\!]^{x_+}_{x_-}\rangle^2_{|\tilde{A}|,e\times I^n}\right)$$
$$+ \sum_{n=0}^{N-1}\langle v\rangle^2_{|\tilde{A}|,\Gamma\times I^n} + \|v(t_-^N)\|^2_{\tilde{A}_0,\Omega} = \|v(t_-^0)\|^2_{\tilde{A}_0,\Omega} + \sum_{n=0}^{N-1} 2\,\langle v,\tilde{g}\rangle_{(-\tilde{A}^-),\Gamma\times I^n} \ .$$

Proof. Omitted, see [2,1]. □

This energy balance equation formally bounds the energy at time t_-^N in terms of initial data and inflow boundary data. Next, we consider the full nonlinear system of conservation laws.

2.2 Nonlinear Energy Analysis

Before presenting the nonlinear energy result, we prove a general lemma and consequential corollaries concerning entropy function/flux jump identities at space-time slab interfaces. Note that throughout this section, we utilize the state-space parameterization

$$\overline{v}(\theta) \equiv v(x_-) + \theta\,[\![v]\!]^{x_+}_{x_-}$$

(similarly across time slab interfaces) for use in state-space path integrations and the interface averaging operator

$$\langle\!\langle a\rangle\!\rangle^{x_+}_{x_-} \equiv \frac{a(x_-)+a(x_+)}{2} \ .$$

Lemma 2.2. Interface Jump Identities. *Let $Z(u), \mathcal{Z}(v) : \mathbb{R}^m \mapsto \mathbb{R}$ be twice differentiable functions of their argument satisfying the duality relationship*

$$Z(u) + \mathcal{Z}(v) = \mathcal{Z}_{,v}\,v \ . \tag{2.9}$$

The following jump identities hold across interfaces

$$[\![Z]\!]^{x_+}_{x_-} - [\![\mathcal{Z}_{,v}]\!]^{x_+}_{x_-}\,v(x_+) + \int_0^1 (1-\theta)[\![v]\!]^{x_+}_{x_-}\cdot\,\mathcal{Z}_{,v,v}(\overline{v}(\theta))\,[\![v]\!]^{x_+}_{x_-}\,d\theta = 0 \tag{2.10a}$$

$$[\![Z]\!]^{x_+}_{x_-} - [\![\mathcal{Z}_{,v}]\!]^{x_+}_{x_-}\,v(x_-) - \int_0^1 \theta\,[\![v]\!]^{x_+}_{x_-}\cdot\,\mathcal{Z}_{,v,v}(\overline{v}(\theta))\,[\![v]\!]^{x_+}_{x_-}\,d\theta = 0 \ . \tag{2.10b}$$

Proof. Recall the following forms of Taylor series with integral remainder

$$[\![\mathcal{Z}]\!]^{x_+}_{x_-} - \mathcal{Z}_{,v}(x_+)\,[\![v]\!]^{x_+}_{x_-} + \int_0^1 \theta\,[\![v]\!]^{x_+}_{x_-}\cdot\,\mathcal{Z}_{,v,v}\,(\overline{v}(\theta))\,[\![v]\!]^{x_+}_{x_-}\,d\theta = 0 \tag{2.11a}$$

$$[\![\mathcal{Z}]\!]^{x_+}_{x_-} - \mathcal{Z}_{,v}(x_-)\,[\![v]\!]^{x_+}_{x_-} - \int_0^1 (1-\theta)\,[\![v]\!]^{x_+}_{x_-}\cdot\,\mathcal{Z}_{,v,v}\,(\overline{v}(\theta))\,[\![v]\!]^{x_+}_{x_-}\,d\theta = 0 \tag{2.11b}$$

and the jump form of (2.9)

$$[\mathcal{Z}]_{x_-}^{x_+} + [Z]_{x_-}^{x_+} = \langle\!\langle \mathcal{Z}_{,\boldsymbol{v}} \rangle\!\rangle_{x_-}^{x_+} [\boldsymbol{v}]_{x_-}^{x_+} + \langle\!\langle \boldsymbol{v} \rangle\!\rangle_{x_-}^{x_+} [\mathcal{Z}_{,\boldsymbol{v}}]_{x_-}^{x_+} \quad . \tag{2.12}$$

Combining (2.11a), (2.11b) and (2.12) yields

$$[Z]_{x_-}^{x_+} - \langle\!\langle \boldsymbol{v} \rangle\!\rangle_{x_-}^{x_+} [\mathcal{Z}_{,\boldsymbol{v}}]_{x_-}^{x_+} + \frac{1}{2} \int_0^1 (1-2\theta)\, [\boldsymbol{v}]_{x_-}^{x_+} \cdot \mathcal{Z}_{,\boldsymbol{v},\boldsymbol{v}}\, (\overline{\boldsymbol{v}}(\theta))\, [\boldsymbol{v}]_{x_-}^{x_+}\, d\theta = 0. \tag{2.13}$$

Finally, algebraically manipulating this form together with the mean-value identity

$$[\mathcal{Z}_{,\boldsymbol{v}}]_{x_-}^{x_+} = \int_0^1 \mathcal{Z}_{,\boldsymbol{v},\boldsymbol{v}}(\overline{\boldsymbol{v}}(\theta))\, [\boldsymbol{v}]_{x_-}^{x_+}\, d\theta \tag{2.14}$$

produces the stated lemma. □

Corollary 2.3. Temporal Space-Time Slab Interface Identity. *Let $t_\pm$ denote a temporal space-time slab interface. The following entropy function jump identity holds across time slab interfaces*

$$\int_\Omega \left([U]_{t_-}^{t_+} - \boldsymbol{v}^T(t_+)\, [\boldsymbol{u}]_{t_-}^{t_+} \right) dx + \frac{1}{2} |||\, [\boldsymbol{v}]_{t_-}^{t_+} \,|||_{\tilde{A}_0,\Omega}^2 = 0 \tag{2.15}$$

where

$$|||\, [\boldsymbol{v}]_{t_-}^{t_+} \,|||_{\tilde{A}_0,\Omega}^2 \equiv \int_\Omega \int_0^1 2(1-\theta)\, [\boldsymbol{v}]_{t_-}^{t_+} \cdot \tilde{A}_0(\overline{\boldsymbol{v}}(\theta))\, [\boldsymbol{v}]_{t_-}^{t_+}\, d\theta\, dx \geq 0 \quad . \tag{2.16}$$

Proof. Set $Z = U$, $\mathcal{Z} = \mathcal{U}$ with $\mathcal{U}_{,\boldsymbol{v}} = \boldsymbol{u}^T$ and $\mathcal{U}_{,\boldsymbol{v},\boldsymbol{v}} = \tilde{A}_0$ in Lemma 2.2 and replace $x_\pm$ with $t_\pm$ in (2.10a), see [10] for an alternative form. □

Corollary 2.4. Spatial Space-Time Slab Interface Identity. *Let $x_\pm$ denote a spatial element interface. The following entropy jump identity holds across spatial element interfaces*

$$[F^i]_{x_-}^{x_+} - \langle\!\langle \boldsymbol{v}^T \rangle\!\rangle_{x_-}^{x_+} [\boldsymbol{f}^i]_{x_-}^{x_+} + \frac{1}{2} \int_0^1 (1-2\theta)\, [\boldsymbol{v}]_{x_-}^{x_+} \cdot \tilde{A}_i(\overline{\boldsymbol{v}}(\theta))\, [\boldsymbol{v}]_{x_-}^{x_+}\, d\theta = 0 \quad . \tag{2.17}$$

Proof. Set $Z = F^i$, $\mathcal{Z} = \mathcal{F}^i$, $i = 1, \ldots, d$ with $\mathcal{F}^i_{,\boldsymbol{v}} = (\boldsymbol{f}^i)^T$ and $\mathcal{F}^i_{,\boldsymbol{v},\boldsymbol{v}} = \tilde{A}_i$ in (2.13) of Lemma 2.2. □

Note that in actual numerical calculations, it is desirable to use the variational form given by (2.3) since integration by parts has been used to insure exact discrete conservation even with inexact numerical quadrature of the various integrals. For analysis purposes, however, it is desirable to use the following equivalent non-integrated-by-parts formulation:

Find $\boldsymbol{v}^h \in \mathcal{V}^h$ such that for all $\boldsymbol{w}^h \in \mathcal{V}^h$

$$B(\boldsymbol{v}^h, \boldsymbol{w}^h)_{\text{GAL}} + B(\boldsymbol{v}^h, \boldsymbol{w}^h)_{\text{LS}} + B(\boldsymbol{v}^h, \boldsymbol{w}^h)_{\text{BC}} = 0 \tag{2.18}$$

$$
\begin{aligned}
B(\boldsymbol{v},\boldsymbol{w})_{\mathrm{GAL}} &= \int_{I^n}\int_{\Omega} \boldsymbol{w}\cdot\left(\boldsymbol{u}_{,t}+\boldsymbol{f}^i_{,x_i}(\boldsymbol{v})\right)\,dx\,dt \\
&+ \int_{\Omega} \boldsymbol{w}(t^n_+)\cdot[\boldsymbol{u}]^{t^n_+}_{t^n_-}\,dx \\
&+ \int_{I^n}\sum_{e\in\mathcal{E}}\int_e \frac{1}{2}\,[\boldsymbol{w}]^{x_+}_{x_-}\cdot\boldsymbol{h}^d(\boldsymbol{v}(x_-),\boldsymbol{v}(x_+);\boldsymbol{n})\,dx\,dt \\
&+ \int_{I^n}\sum_{e\in\mathcal{E}}\int_e \langle\!\langle \boldsymbol{w}\rangle\!\rangle^{x_+}_{x_-}\cdot[\boldsymbol{f}(\boldsymbol{v};\boldsymbol{n})]^{x_+}_{x_-}\,dx\,dt \\
B(\boldsymbol{v},\boldsymbol{w})_{\mathrm{LS}} &= \int_{I^n}\sum_{T\in\mathcal{T}}\int_T \left(\tilde{A}_0\boldsymbol{w}_{,t}+\tilde{A}_i\boldsymbol{w}_{,x_i}\right)\cdot\boldsymbol{\tau}\left(\tilde{A}_0\boldsymbol{v}_{,t}+\tilde{A}_i\boldsymbol{v}_{,x_i}\right)\,dx\,dt \\
B(\boldsymbol{v},\boldsymbol{w})_{\mathrm{BC}} &= \int_{I^n}\int_{\Gamma} \boldsymbol{w}\cdot\frac{1}{2}\left(\boldsymbol{f}(\tilde{\boldsymbol{g}};\boldsymbol{n})-\boldsymbol{f}(\boldsymbol{v};\boldsymbol{n})-\boldsymbol{h}^d(\boldsymbol{v},\tilde{\boldsymbol{g}};\boldsymbol{n})\right)\,dx\,dt
\end{aligned}
$$

where $\boldsymbol{h}^d$ denotes the flux dissipation term incorporated into the total numerical flux.

Theorem 2.5. Global Entropy Norm Stability (Nonlinear Hyperbolic System). *The variational formulation (2.3) for nonlinear systems of conservation laws with convex entropy extension and symmetric mean-value flux dissipation*

$$
\boldsymbol{h}^d_{\mathrm{SMV}}(\boldsymbol{v}_-,\boldsymbol{v}_+;\boldsymbol{n}) = |A|_{\mathrm{SMV}}\,[\boldsymbol{v}]^{x_+}_{x_-}\quad,\quad |A|_{\mathrm{SMV}}\equiv\int_0^1 |\tilde{A}(\overline{\boldsymbol{v}}(\theta);\boldsymbol{n})|_{\tilde{A}_0}\,d\theta
$$

is entropy norm stable (modulo data $\tilde{\boldsymbol{g}}$) with the following global balance:

$$
\begin{aligned}
&\sum_{n=0}^{N-1}\left(|||\,[\boldsymbol{v}]^{t^n_+}_{t^n_-}\,|||^2_{\tilde{A}_0,\Omega}+2\|\tilde{A}_0\boldsymbol{v}_{,t}+\tilde{A}_i\boldsymbol{v}_{,x_i}\|^2_{\tau,\Omega\times I^n}+\sum_{e\in\mathcal{E}}\langle[\boldsymbol{v}]^{x_+}_{x_-}\rangle^2_{|\underline{\tilde{A}}|,e\times I^n}\right) \\
&+\sum_{n=0}^{N-1}\langle\boldsymbol{v}\rangle^2_{|\underline{\tilde{A}}|,\Gamma\times I^n}+\int_{\Omega}2\,U(t^N_-)\,dx=\int_{\Omega}2\,U(t^0_-)\,dx+\sum_{n=0}^{N-1}2\left(\langle\boldsymbol{v},\tilde{\boldsymbol{g}}\rangle_{(-\underline{\tilde{A}}^-),\Gamma\times I^n}+G^n_{\Gamma}(\tilde{\boldsymbol{g}},\boldsymbol{v})\right)
\end{aligned}
$$

with

$$
|\underline{\tilde{A}}(\boldsymbol{n})| = \int_0^1 2\,(1-\theta)\,\left(\tilde{A}^+(\overline{\boldsymbol{v}}(\theta);\boldsymbol{n})_{\tilde{A}_0}-\tilde{A}^-(\overline{\boldsymbol{v}}(1-\theta);\boldsymbol{n})_{\tilde{A}_0}\right)\,d\theta
$$

and

$$
G^n_{\Gamma}(\tilde{\boldsymbol{g}},\boldsymbol{v}) = \int_{I^n}\int_{\Gamma}\left(F(\tilde{\boldsymbol{g}};\boldsymbol{n})-\int_0^1\theta\;\tilde{\boldsymbol{g}}\cdot A(\overline{\boldsymbol{v}}(\theta);\boldsymbol{n})\;\tilde{\boldsymbol{g}}\,d\theta\right)\,dx\,dt\ .
$$

Proof. Construct the energy balance for the interval $[t^N_-,t^0_-]=\cup_{n=0}^{N-1}I^n$ by setting $\boldsymbol{w}=\boldsymbol{v}$ and evaluating the various integrals. Consider the time derivative integral

$$
\int_{\Omega}\int_{I^n}\boldsymbol{v}^T\boldsymbol{u}_{,t}\,dt\,dx=\int_{\Omega}\int_{I^n}U_{,t}\,dt\,dx=\int_{\Omega}\left([U]^{t^{n+1}_-}_{t^n_-}-[U]^{t^n_+}_{t^n_-}\right)\,dx
$$

and combine with the jump integral across time slabs. ¿From Corollary 2.3

$$\int_\Omega\int_{I^n} \boldsymbol{v}^T \boldsymbol{u}_{,t}\, dt\, dx + \int_\Omega \boldsymbol{v}^T(t_+^n)[\boldsymbol{u}]_{t_-^n}^{t_+^n}\, dx = \int_\Omega [U]_{t_-^n}^{t_-^{n+1}}\, dx + \frac{1}{2}|||\,[\boldsymbol{v}]_{t_-^n}^{t_+^n}\,|||^2_{\tilde{A}_0,\Omega}\ .$$

When summed over all time slabs, the first term on the right-hand-side of this equation vanishes except for initial and final time slab contributions. Next, consider the spatial operator term and apply the divergence theorem

$$\begin{aligned}\int_{I^n}\int_\Omega \boldsymbol{v}^T \boldsymbol{f}^i_{,x_i}\, dx\, dt &= \int_{I^n}\int_\Omega F^i_{,x_i}\, dx\, dt \\ &= \int_{I^n}\sum_{e\in\mathcal{E}}\int_e -[F(\boldsymbol{v};\boldsymbol{n})]_{x_-}^{x_+}\, dx\, dt + \int_{I^n}\int_\Gamma F(\boldsymbol{v};\boldsymbol{n})\, dx\, dt\end{aligned}$$

where $F(\boldsymbol{v};\boldsymbol{n}) = \boldsymbol{n}_i\, F^i(\boldsymbol{v})$. ¿From Corollary 2.4 and the definition of $|\underline{\tilde{A}}|$, it follows that

$$\begin{aligned}\int_{I^n}\int_\Omega \boldsymbol{v}^T \boldsymbol{f}^i_{,x_i}\, dx\, dt + \int_{I^n}\sum_{e\in\mathcal{E}}\int_e \Big(\langle\!\langle \boldsymbol{v}\rangle\!\rangle_{x_-}^{x_+}\cdot[\boldsymbol{f}(\boldsymbol{n})]_{x_-}^{x_+} + \frac{1}{2}[\boldsymbol{v}]_{x_-}^{x_+}\cdot \boldsymbol{h}^d_{\mathrm{SMV}}\Big)\, dx\, dt \\ = \sum_{e\in\mathcal{E}}\frac{1}{2}\langle[\boldsymbol{v}]_{x_-}^{x_+}\rangle^2_{|\underline{\tilde{A}}|,e\times I^n} + \int_{I^n}\int_\Gamma F(\boldsymbol{v};\boldsymbol{n})\, dx\, dt\ .\end{aligned}$$

In summary, collecting terms we have

$$\begin{aligned}B(\boldsymbol{v},\boldsymbol{v})_{\mathrm{GAL}} = \int_\Omega [U]_{t_-^n}^{t_-^{n+1}}\, dx + \frac{1}{2}|||\,[\boldsymbol{v}]_{t_-^n}^{t_+^n}\,|||^2_{\tilde{A}_0,\Omega} + \sum_{e\in\mathcal{E}}\frac{1}{2}\langle[\boldsymbol{v}]_{x_-}^{x_+}\rangle^2_{|\underline{\tilde{A}}|,e\times I^n} \\ + \int_{I^n}\int_\Gamma F(\boldsymbol{v};\boldsymbol{n})\, dx\, dt\ .\end{aligned}$$

The least-squares integral produces a pure quadratic form without modification

$$B(\boldsymbol{v},\boldsymbol{v})_{\mathrm{LS}} = \|\tilde{A}_0\boldsymbol{v}_{,t} + \tilde{A}_i\boldsymbol{v}_{,x_i}\|^2_{\boldsymbol{\tau},\Omega\times I^n}\ .$$

Finally, consider the boundary condition terms and apply Corollary 2.4

$$\begin{aligned}B(\boldsymbol{v},\boldsymbol{v})_{\mathrm{BC}} &= \int_{I^n}\int_\Gamma \Big(\frac{1}{2}\boldsymbol{v}\cdot(\boldsymbol{f}(\tilde{\boldsymbol{g}};\boldsymbol{n}) - \boldsymbol{f}(\boldsymbol{v};\boldsymbol{n})) - \frac{1}{2}\boldsymbol{v}\cdot|A(\boldsymbol{n})|_{\mathrm{SMV}}(\tilde{\boldsymbol{g}}-\boldsymbol{v})\Big)\, dx\, dt \\ &= \int_{I^n}\int_\Gamma \Big(F(\tilde{\boldsymbol{g}};\boldsymbol{n}) - F(\boldsymbol{v};\boldsymbol{n}) + \frac{1}{2}\int_0^1 (1-2\,\theta)(\tilde{\boldsymbol{g}}-\boldsymbol{v})\cdot\tilde{A}(\overline{\boldsymbol{v}}(\theta);\boldsymbol{n})\,(\tilde{\boldsymbol{g}}-\boldsymbol{v})\, d\theta \\ &\qquad -\frac{1}{2}\boldsymbol{g}\cdot(\boldsymbol{f}(\tilde{\boldsymbol{g}};\boldsymbol{n}) - \boldsymbol{f}(\boldsymbol{v};\boldsymbol{n})) - \frac{1}{2}\boldsymbol{v}\cdot|A(\boldsymbol{n})|_{\mathrm{SMV}}(\tilde{\boldsymbol{g}}-\boldsymbol{v})\Big)\, dx\, dt \\ &= \int_{I^n}\int_\Gamma \Big(F(\tilde{\boldsymbol{g}};\boldsymbol{n}) - F(\boldsymbol{v};\boldsymbol{n}) - \int_0^1 \theta\ \tilde{\boldsymbol{g}}\cdot\tilde{A}(\overline{\boldsymbol{v}}(\theta))\,\tilde{\boldsymbol{g}}\, d\theta \\ &\qquad +\frac{1}{2}\boldsymbol{v}\cdot|\underline{\tilde{A}}(\boldsymbol{n})|\,\boldsymbol{v} - \boldsymbol{v}\cdot\underline{\tilde{A}}^-(\boldsymbol{n})\,\tilde{\boldsymbol{g}}\Big)\, dx\, dt\ .\end{aligned}$$

Combining the above results, summing over time slabs, and multiplication by two yields an entropy norm balance equation (2.19) which bounds the global entropy norm of the system at the final time T in terms of the initial data and boundary data $\tilde{\boldsymbol{g}}$. □

Remark 2.6. Note that when the $\tilde{A}_i$ matrices are assumed constant, $\boldsymbol{f}^i = \tilde{A}_i \boldsymbol{v}$ and $F^i = \frac{1}{2}\boldsymbol{v} \cdot \tilde{A}_i \boldsymbol{v}$ so that the additional term $G_I^n(\tilde{\boldsymbol{g}}, \boldsymbol{v}; \boldsymbol{n})$ vanishes identically and a one-to-one correspondence of terms between (2.9) and (2.5) is achieved.

2.3 A Simplified DG Method in Symmetric Form

DG Flux Formulas. Simplification of the discontinuous Galerkin method follows by choosing the discrete symmetric mean-value flux function proposed earlier, i.e.

$$\boldsymbol{h}_{\mathrm{DSMV}}(\boldsymbol{v}_-, \boldsymbol{v}_+; \boldsymbol{n}) = \frac{1}{2}\left(\boldsymbol{f}(\boldsymbol{v}_-; \boldsymbol{n}) + \boldsymbol{f}(\boldsymbol{v}_+; \boldsymbol{n})\right) - \frac{1}{2}\sum_{q=1}^{N} w_q |\tilde{A}(\overline{\boldsymbol{v}}(\theta_q)); \boldsymbol{n})|_{\tilde{A}_0} \, [\boldsymbol{v}]_{x_-}^{x_+}$$

with $|\tilde{A}|_{\tilde{A}_0} = \tilde{R}\,|\Lambda|\,\tilde{R}^T$, $\overline{\boldsymbol{v}}(\theta) = \boldsymbol{v}(x_-) + \theta\,[\boldsymbol{v}]_{x_-}^{x_+}$. By using sufficient order numerical quadrature and subdivision of the state-space path integration at points of non-differentiability, the $\boldsymbol{h}^d_{\mathrm{DSMV}}$ flux can be made arbitrarily close to $\boldsymbol{h}^d_{\mathrm{SMV}}$ for which nonlinear stability in the DG method follows from the analysis of Sect. 2.2. Suppose that elements of $|\tilde{A}(\overline{\boldsymbol{v}}(\theta_q)); \boldsymbol{n})|_{\tilde{A}_0}$ remain bounded for $\theta \in [0, 1]$ independent of N. Using N point Gaussian quadrature

$$\|\boldsymbol{h}^d_{\mathrm{DSMV}} - \boldsymbol{h}^d_{\mathrm{SMV}}\|_2 = O([\boldsymbol{v}]^{2N+1}) \ .$$

Next, we consider single-point quadrature formulas.

Theorem 2.7. Discrete Symmetric Mean-Value Flux. *Let $\boldsymbol{v}_*$ be a state such that*

$$[\boldsymbol{v}]_{x_-}^{x_+} \cdot |\tilde{A}(\boldsymbol{v}_*; \boldsymbol{n})|\, [\boldsymbol{v}]_{x_-}^{x_+} = \sup_{0 \le \theta \le 1} [\boldsymbol{v}]_{x_-}^{x_+} \cdot |\tilde{A}(\overline{\boldsymbol{v}}(\theta); \boldsymbol{n})|_{\tilde{A}_0} \, [\boldsymbol{v}]_{x_-}^{x_+} \ .$$

The variational formulation (2.3) with numerical flux function

$$\boldsymbol{h}_{\mathrm{DSMV}*}(\boldsymbol{v}_-, \boldsymbol{v}_+; \boldsymbol{n}) = \frac{1}{2}\left(\boldsymbol{f}(\boldsymbol{v}_-; \boldsymbol{n}) + \boldsymbol{f}(\boldsymbol{v}_+; \boldsymbol{n})\right) - \frac{1}{2}|\tilde{A}(\boldsymbol{v}_*); \boldsymbol{n})|_{\tilde{A}_0} \, [\boldsymbol{v}]_{x_-}^{x_+} \tag{2.19}$$

is energy bounded in the sense of Theorem 2.5.

Proof. It is sufficient to show that the given flux dissipation

$$\boldsymbol{h}^d_{\mathrm{DSMV}*} = |\tilde{A}(\boldsymbol{v}_*; \boldsymbol{n})|_{\tilde{A}_0} \, [\boldsymbol{v}]_{x_-}^{x_+}$$

exceeds the symmetric mean-value value flux dissipation. This is reflected by the algebraic condition

$$[\boldsymbol{v}]_{x_-}^{x_+} \cdot \boldsymbol{h}^d_{\mathrm{SMV}} \le [\boldsymbol{v}]_{x_-}^{x_+} \cdot \boldsymbol{h}^d_{\mathrm{DSMV}*} \ .$$

¿From the symmetric mean-value flux definition

$$\begin{aligned} [\boldsymbol{v}]_{x_-}^{x_+} \cdot \boldsymbol{h}^d_{\mathrm{SMV}} &= [\boldsymbol{v}]_{x_-}^{x_+} \cdot \int_0^1 |\tilde{A}(\overline{\boldsymbol{v}}(\theta); \boldsymbol{n})|_{\tilde{A}_0} \, d\theta \, [\boldsymbol{v}]_{x_-}^{x_+} \\ &\le \sup_{0 \le \theta \le 1} [\boldsymbol{v}]_{x_-}^{x_+} \cdot |\tilde{A}(\overline{\boldsymbol{v}}(\theta); \boldsymbol{n})|_{\tilde{A}_0} \, [\boldsymbol{v}]_{x_-}^{x_+} \end{aligned}$$

$$= [\boldsymbol{v}]_{x_-}^{x_+} \cdot |\tilde{A}(\boldsymbol{v}_*;\boldsymbol{n})|_{\tilde{A}_0} [\boldsymbol{v}]_{x_-}^{x_+}$$
$$= [\boldsymbol{v}]_{x_-}^{x_+} \cdot \boldsymbol{h}_{\mathrm{DSMV}}^{d} \ .$$

This establishes nonlinear stability of the DG method using the simplified flux function. □

Remark 2.8. Unfortunately, the state $\boldsymbol{v}_*$ is not generally known in closed form. Cockburn and Shu [4] have shown impressive results using the simpler Lax-Friedrichs flux. It is straightforward to derive a corresponding "symmetric Lax-Friedrichs" numerical flux function

$$\boldsymbol{h}_{\mathrm{SLF}}(\boldsymbol{v}_-,\boldsymbol{v}_+;\boldsymbol{n}) = \frac{1}{2}\left(\boldsymbol{f}(\boldsymbol{v}_-;\boldsymbol{n}) + \boldsymbol{f}(\boldsymbol{v}_+;\boldsymbol{n})\right) - \frac{1}{2}\lambda_{\max}\tilde{A}_0(\boldsymbol{v}_*)\,[\boldsymbol{v}]_{x_-}^{x_+}$$

with $\lambda_{\max} = \sup_{0\le\theta\le 1}\left(\max_{1\le i\le m}\left(\Lambda_{ii}(\overline{\boldsymbol{v}}(\theta))\right)\right)$. Nonlinear entropy norm stability follows starting from Theorem 2.7

$$\begin{aligned}
[\boldsymbol{v}]_{x_-}^{x_+} \cdot \boldsymbol{h}_{\mathrm{SMV}}^{d} &\le [\boldsymbol{v}]_{x_-}^{x_+} \cdot |\tilde{A}(\boldsymbol{v}_*;\boldsymbol{n})|_{\tilde{A}_0} [\boldsymbol{v}]_{x_-}^{x_+} \\
&= [\boldsymbol{v}]_{x_-}^{x_+} \cdot \tilde{R}(\boldsymbol{v}_*;\boldsymbol{n})\,|\Lambda(\boldsymbol{v}_*;\boldsymbol{n})|\,\tilde{R}^T(\boldsymbol{v}_*;\boldsymbol{n})\,[\boldsymbol{v}]_{x_-}^{x_+} \\
&\le \sup_{0\le\theta\le 1}\left(\max_{1\le i\le m}\left(\Lambda_{ii}(\overline{\boldsymbol{v}}(\theta))\right)\right)[\boldsymbol{v}]_{x_-}^{x_+} \cdot \tilde{R}(\boldsymbol{v}_*;\boldsymbol{n})\,\tilde{R}^T(\boldsymbol{v}_*;\boldsymbol{n})\,[\boldsymbol{v}]_{x_-}^{x_+} \\
&= \lambda_{\max}\,[\boldsymbol{v}]_{x_-}^{x_+} \cdot \tilde{A}_0(\boldsymbol{v}_*;\boldsymbol{n})\,[\boldsymbol{v}]_{x_-}^{x_+} \ .
\end{aligned}$$

Finally, for systems such as the Euler equations of gas dynamics that exhibit the property $\frac{\partial^4 \mathcal{U}}{\partial v_i v_j v_k v_l} z_i z_j z_k z_l > 0$, $|\boldsymbol{z}| \neq 0$, we have

$$[\boldsymbol{v}]_{x_-}^{x_+} \cdot \tilde{A}_0(\boldsymbol{v}_*;\boldsymbol{n})\,[\boldsymbol{v}]_{x_-}^{x_+} \le \max\left([\boldsymbol{v}]_{x_-}^{x_+} \cdot \tilde{A}_0(\overline{\boldsymbol{v}}(0);\boldsymbol{n})\,[\boldsymbol{v}]_{x_-}^{x_+}, [\boldsymbol{v}]_{x_-}^{x_+} \cdot \tilde{A}_0(\overline{\boldsymbol{v}}(1);\boldsymbol{n})\,[\boldsymbol{v}]_{x_-}^{x_+}\right)$$

thereby avoiding the need for calculating $\boldsymbol{v}_*$ altogether, see [1] for details.

DG Jacobian Derivatives. Using the discrete mean-value fluxes, it becomes straightforward to compute Jacobian derivatives of various terms. For example, to compute derivatives of $|\tilde{A}|_{\tilde{A}_0}$ with respect to a vector $\boldsymbol{w}$, chain-rule differentiation is used

$$\frac{\partial |\tilde{A}(\boldsymbol{n})|_{\tilde{A}_0}}{\partial \boldsymbol{w}} = \frac{\partial \tilde{R}(\boldsymbol{n})}{\partial \boldsymbol{w}}|\Lambda(\boldsymbol{n})|\tilde{R}^T(\boldsymbol{n}) + \tilde{R}(\boldsymbol{n})\frac{\partial |\Lambda(\boldsymbol{n})|}{\partial \boldsymbol{w}}\tilde{R}^T(\boldsymbol{n}) + \tilde{R}(\boldsymbol{n})|\Lambda(\boldsymbol{n})|\frac{\partial \tilde{R}^T(\boldsymbol{n})}{\partial \boldsymbol{w}}.$$

Note that a high degree of computational efficiency can be achieved in the calculation of these Jacobian terms by exploiting the transpose symmetry of intermediate products.

2.4 Simplified Least-Squares Stabilization in Symmetric Form

Consider an isoparametric element mapping $\boldsymbol{\xi} \mapsto \mathbf{x}$ from a unit element space $\boldsymbol{\xi}$ to a physical space $\mathbf{x}$. In the papers by Hughes and Mallet [7] and Shakib [10], they proposed the following form for τ appearing in (2.3) on a mapped element

$$\tau_p = |\tilde{B}|_{p,\tilde{A}_0}^{-1}, \quad |\tilde{B}|_{p,\tilde{A}_0} = \left(\sum_{i=0}^{d}|B^i|^p\right)^{1/p}\tilde{A}_0, \quad B^i = \sum_{j=0}^{d}\xi_{,x_j}A_j \ . \tag{2.20}$$

Equation (2.20) is of the same form given earlier in (1.15). In standard implementations of least-squares stabilization, $p = 2$ is used. In light of the Eigenvector Scaling Theorem 1.1, it is useful to revisit the derivation of τ with $p = 1$. Let $\tilde{B}^i = B^i \tilde{A}_0$, from (1.16) it follows that

$$\tau_1 = |\tilde{B}|^{-1}_{1,\tilde{A}_0} = \left[|\nabla\xi^0|\tilde{A}_0 + \sum_{i=1}^{d} |\nabla\xi^i| \, \tilde{R}(n^i) \, |\Lambda(n^i)| \, \tilde{R}^T(n^i) \right]^{-1}$$

using the entropy scaled eigenvectors $\tilde{R}(n^i)$ of $\tilde{B}^i$. This represents a substantial simplification of the τ matrix calculation.

3 Concluding Remarks

Simplified forms of the DG, DGLS, and GLS schemes have been presented and analyzed for first-order systems of conservation laws with convex entropy extension. Numerical examples are given in [1] using linear, quadratic, and cubic element approximation.

References

1. T. J. Barth. Simplified discontinuous Galerkin methods for systems of conservation laws with convex extension. *Math. Comp.*, submitted 1999.
2. T.J. Barth. Numerical methods for gasdynamic systems on unstructured meshes. In Kröner, Ohlberger, and Rohde, editors, *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws*, volume 5 of *Lecture Notes in Computational Science and Engineering*, pages 195–285. Springer-Verlag, Heidelberg, 1998.
3. B. Cockburn, S. Hou, and C.W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: The multidimensional case. *Math. Comp.*, 54:545–581, 1990.
4. B. Cockburn and C.W. Shu. The Runge-Kutta discontinuous Galerkin method for conservation laws V: Multidimensional systems. *J. Comput. Phys.*, 141:199-224, 1998.
5. K. O. Friedrichs and P. D. Lax. Systems of conservation laws with convex extension. *Proc. Nat. Acad. Sci. USA*, 68(8):1686–1688, 1971.
6. F. R. Gantmacher. *Matrix Theory.* Chelsea Publishing Company, New York, N.Y., 1959.
7. T. J. R. Hughes and M. Mallet. A new finite element formulation for CFD: III. the generalized streamline operator for multidimensional advective-diffusive systems. *Comp. Meth. Appl. Mech. Engrg.*, 58:305–328, 1986.
8. C. Johnson and J. Pitkäranta. An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comp.*, 46:1–26, 1986.
9. M. S. Mock. Systems of conservation laws of mixed type. *J. Diff. Eqns.*, 37:70–88, 1980.
10. F. Shakib. *Finite Element Analysis of the Compressible Euler and Navier-Stokes Equations.* PhD thesis, Stanford University, Department of Mechanical Engineering, 1988.

A High Order Discontinuous Galerkin Method for Compressible Turbulent Flows

F. Bassi[1] and S. Rebay[2]

[1] Dipartimento di Energetica
Università degli Studi di Ancona
Via Brecce Bianche, 60100 Ancona, Italy
[2] Dipartimento di Ingegneria Meccanica
Università degli Studi di Brescia
Via Branze 38, 25123 Brescia, Italy

Abstract

An implicit high order accurate Discontinuous Galerkin method for the numerical solution of the compressible Favre–Reynolds Averaged Navier–Stokes equations is presented. The method is characterized by a highly compact discretization support even for higher order approximations and this feature can be exploited in the development of implicit integration schemes. Turbulence effects are accounted for by means of the low-Reynolds k-ω turbulence model. A non-standard implementation of the model, whereby the logarithm of ω rather than ω itself is used as unknown, has been found very useful to enhance the stability of the method especially for the higher (third and fourth) order approximations. We present computational results of the transitional flow over a flat plate and of the turbulent flow through a turbine vane with wall heat transfer.

1 Introduction

During the last few years several authors have contributed to the theoretical development and to the application of the Discontinuous Galerkin (DG) method, see e.g. [11,12,3,2,4,7,6]. The growing interest for the DG method is due to its several attractive features. The DG method combines two distinctive characteristics of the finite volume and of the finite element methods, the physics of wave propagation being accounted for by means of Riemann solvers and accuracy being obtained by means of high order polynomial approximations within elements. The method is therefore ideally suited to compute high order accurate solution of the Euler or the Navier–Stokes equations on general unstructured grids. Thanks to the locality and the compactness of the discretization the degrees of freedom associated to a generic element are coupled only with the degrees of freedom associated with neighboring elements. In the case of triangular (tetrahedral) elements, this means that coupling

is introduced only among four (five) elements, respectively. This compactness results in very sparse matrices which are very convenient for implicit integration schemes (especially in 3D).

In this work, we solve the compressible Favre–Reynolds Averaged Navier–Stokes (RANS) equations with the k-ω closure by means of the DG method. The extension of the method described in [4] to the case of the RANS equations is relatively straightforward, but, to improve the stability of the method, we have found useful to adopt the non-standard implementation of the k-ω model described in the next section.

The performance of the method is displayed by computing the transitional flow over a flat plate and through a turbine vane.

2 Governing Equations

The complete set of the RANS and k-ω equations can be written as

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x_j}(\rho u_j) = 0 \tag{1}$$

$$\frac{\partial}{\partial t}(\rho e_0) + \frac{\partial}{\partial x_j}(\rho u_j h_0) = \frac{\partial}{\partial x_j}\left[u_i \widehat{\tau}_{ij} - q_j\right] - \tau_{ij}\frac{\partial u_i}{\partial x_j} + \beta^* \rho \overline{k} e^{\widetilde{\omega}} \tag{2}$$

$$\frac{\partial}{\partial t}(\rho u_i) + \frac{\partial}{\partial x_j}(\rho u_j u_i) = -\frac{\partial p}{\partial x_i} + \frac{\partial \widehat{\tau}_{ji}}{\partial x_j} \tag{3}$$

$$\frac{\partial}{\partial t}(\rho k) + \frac{\partial}{\partial x_j}(\rho u_j k) = \tau_{ij}\frac{\partial u_i}{\partial x_j} - \beta^* \rho \overline{k} e^{\widetilde{\omega}} + \frac{\partial}{\partial x_j}\left[(\mu + \sigma^* \overline{\mu}_t)\frac{\partial k}{\partial x_j}\right] \tag{4}$$

$$\frac{\partial}{\partial t}(\rho \widetilde{\omega}) + \frac{\partial}{\partial x_j}(\rho u_j \widetilde{\omega}) = \frac{\alpha}{\overline{k}}\tau_{ij}\frac{\partial u_i}{\partial x_j} - \beta \rho e^{\widetilde{\omega}} + (\mu + \sigma \overline{\mu}_t)\frac{\partial \widetilde{\omega}}{\partial x_k}\frac{\partial \widetilde{\omega}}{\partial x_k} + \frac{\partial}{\partial x_j}\left[(\mu + \sigma \overline{\mu}_t)\frac{\partial \widetilde{\omega}}{\partial x_j}\right] \tag{5}$$

$$p = (\gamma - 1)\rho\left(e_0 - u_k u_k/2\right) \tag{6}$$

$$q_j = -\left(\frac{\mu}{\mathrm{Pr}} + \frac{\overline{\mu}_t}{\mathrm{Pr}_t}\right)\frac{\partial h}{\partial x_j} \tag{7}$$

$$\tau_{ij} = 2\overline{\mu}_t\left[\frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) - \frac{1}{3}\frac{\partial u_k}{\partial x_k}\delta_{ij}\right] - \frac{2}{3}\rho \overline{k}\delta_{ij} \tag{8}$$

$$\widehat{\tau}_{ij} = 2\mu \left[\frac{1}{2} \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) - \frac{1}{3} \frac{\partial u_k}{\partial x_k} \delta_{ij} \right] + \tau_{ij} \tag{9}$$

$$\overline{k} = \max(0, k) \qquad \overline{\mu}_t = \alpha^* \rho \overline{k} e^{-\widetilde{\omega}} \tag{10}$$

where γ, Pr and Pr_t are the ratio of gas specific heats, the molecular and turbulent Prandtl numbers which are constant for perfect gases. The value of the k-ω closure parameters α, α^*, β, β^*, σ, σ^* are those of the low-Reynolds k-ω model, see [18].

The above equations present some notable differences in comparison with the usual formulations of the RANS equations with k-ω closure. The equation for the turbulence specific dissipation rate ω has been replaced by the equation for $\widetilde{\omega} = \log\omega$. The use of the logarithm of turbulence variables, introduced by Ilinca and Pelletier in [15] in the context of the k-ϵ model using wall functions, is advantageous at least for two reasons. First of all the turbulence variables are guaranteed to be positive. Secondly, the near wall distribution of the logarithmic turbulence variables is much more smooth than that of the turbulence variables themselves. Owing to the fact that the k-ω model used in this work is integrated down to the wall without using wall functions, we have found useful to introduce the logarithm of ω but not the logarithm of k (which has to satisfy the condition $k = 0$ at solid walls). The positivity of k has been enforced through the definition of the "limited" value $\overline{k}$, which guarantees the positivity of $\overline{\mu}_t$, see Eq. (10), and affects the source terms in Eqs. (2), (4), (5). Notice that the turbulent kinetic energy k appearing in the differential operators of Eq. (4) is not limited, and could therefore take negative values. In practice we have noticed that negative values of k may occur during the pseudo time evolution of the solution or may even be present in a converged steady state solution. In our experience, however, the occurrence of negative k values can be eliminated by refining the computational grid and/or increasing the polynomial approximation degree. As a final comment on the Equations, notice that a source term is present in the mean-flow energy equation because the total energy e_0 and the total enthalpy h_0 do not include the turbulent kinetic energy, see e.g. [13].

3 DG Space Discretization

The complete set of the RANS and k-ω turbulence model equations can be written in compact form as

$$\frac{\partial u}{\partial t} + \nabla \cdot \boldsymbol{f}_c(u) + \nabla \cdot \boldsymbol{f}_v(u, \nabla u) + s(u, \nabla u) = 0,$$

where $u \in \mathbb{R}^{d+2}$ and $s \in \mathbb{R}^{d+2}$ denote the vectors of the conservative variables and of the source terms, $\boldsymbol{f}_c$ and $\boldsymbol{f}_v \in \mathbb{R}^{d+2} \otimes \mathbb{R}^d$ denote the inviscid and

viscous flux functions, respectively, and d is the space dimension. The entries of u, s, $\boldsymbol{f}_c$ and $\boldsymbol{f}_v$ can be found by comparison with Eqs. (1–9).

The weighted residual formulation of the RANS equations is

$$\int_\Omega \phi \frac{\partial u}{\partial t}\, d\Omega + \int_\Omega \phi \nabla \cdot \boldsymbol{f}(u, \nabla u)\, d\Omega + \int_\Omega \phi s(u, \nabla u)\, d\Omega = 0, \tag{11}$$

where $\boldsymbol{f}$ is the sum of the inviscid and of the viscous fluxes, ϕ denotes an arbitrary test function and Ω is the domain in which the solution is to be computed.

Eq. (11) is discretized by subdividing Ω into a set of elements $\{E\}$ and by restricting u and ϕ to be polynomial functions inside each element. No global continuity is enforced on u and ϕ, which are therefore discontinuous at element interfaces. By splitting the integral over Ω into the sum of integrals over the elements E and by performing an integration by parts, we obtain the weak formulation of Eq. (11). If we consider, for each element E, functions ϕ which are zero outside E, we obtain the elemental equation

$$\int_E \phi \frac{\partial u}{\partial t}\, d\Omega + \oint_{\partial E} \phi \boldsymbol{f}(u, \nabla u) \cdot \boldsymbol{n}\, d\sigma - \int_E \nabla \phi \cdot \boldsymbol{f}(u, \nabla u)\, d\Omega + \int_E \phi s(u, \nabla u)\, d\Omega = 0. \tag{12}$$

Due to the discontinuous function approximation, the flux function in the boundary integral of Eq. (12) is not uniquely defined. To ensure a coupling between neighboring elements, which otherwise would be completely missing, the physical normal flux $\boldsymbol{f}(u, \nabla u) \cdot \boldsymbol{n}$ is replaced by a numerical flux $h(u, \nabla u, u^+, \nabla u^+, \boldsymbol{n})$ which depends on the internal $(\cdot)$ and external $(\cdot)^+$ interface quantities and on the normal unit vector $\boldsymbol{n}$, pointing outward from E.

The numerical flux function $h_c(u, u^+, \boldsymbol{n})$ for the inviscid part of the RANS equations is completely analogous to that commonly employed in upwind finite volume methods. In our computations we have used the van Leer vector split flux as modified by Hänel [14].

The treatment of the viscous flux in the context of the DG method has been addressed by introducing the auxiliary variable $\boldsymbol{\theta} = \nabla u$ and by discretizing the following system of two first order equations

$$\begin{cases} \boldsymbol{\theta} = \nabla u \\ \partial_t u + \nabla \cdot \boldsymbol{f}_c + \nabla \cdot \boldsymbol{f}_v + s = 0 \end{cases} \tag{13}$$

by means of DG techniques similar to those already developed for hyperbolic systems of conservation laws. By applying the DG discretization to the first equation of the system we obtain

$$\int_E \phi \boldsymbol{\theta}\, d\Omega - \int_E \phi \nabla u\, d\Omega = \oint_{\partial E} \phi (u_0 - u) \boldsymbol{n}\, d\sigma \tag{14}$$

where $u_0 = u_b$ if $S \cap \partial\Omega \neq 0$ (u_b denotes the boundary data), and $u_0 = (u + u^+)/2$ if $S \cap \partial\Omega = 0$. Eq. (14) suggested us to introduce the function $\boldsymbol{R}$ defined as

$$\int_E \phi \boldsymbol{R}\, d\Omega = \oint_{\partial E} \phi(u_0 - u)\boldsymbol{n}\, d\sigma. \tag{15}$$

Comparing Eqs. (14) and (15) we obtain, at the discrete level,

$$\boldsymbol{\theta} = \boldsymbol{\nabla} u + \boldsymbol{R} \tag{16}$$

and $\boldsymbol{\theta}$ can thus be interpreted as a "modified" gradient that takes into account the effect of interface discontinuities represented by $\boldsymbol{R}$. Considering Eq. (16) the DG discretization of the second equation of system (13) can be written as

$$\int_E \phi \frac{\partial u}{\partial t}\, d\Omega + \oint_{\partial E} \phi h(u, \boldsymbol{\nabla} u + \boldsymbol{R}, u^+, \boldsymbol{\nabla} u^+ + \boldsymbol{R}^+, \boldsymbol{n})\, d\sigma - \int_E \boldsymbol{\nabla}\phi \cdot \boldsymbol{f}(u, \boldsymbol{\nabla} u + \boldsymbol{R})\, d\Omega + \int_E \phi s(u, \boldsymbol{\nabla} u + \boldsymbol{R}) = 0. \tag{17}$$

Eq. (15) and Eq. (17) can be regarded as a system of two equations in the unknowns u and $\boldsymbol{R}$ which discretize the viscous contribution to the RANS equations in mixed form, see e.g. [2]. Unfortunately, this formulation can be shown to be singular in some model problems and, moreover, displays an unsatisfactory convergence rate for polynomial approximations of odd order, see e.g. [5,4].

A cure to this problem, proposed in [5,4] and theoretically justified by Brezzi and coworkers [9,8,7], has been found by replacing the functions $\boldsymbol{R}$ in the contour integral of Eq. (17) with "face" contributions $\boldsymbol{r}_e$ defined as

$$\int_E \phi \boldsymbol{r}_e\, d\Omega = \int_e \phi(u_0 - u)\boldsymbol{n}\, d\sigma \quad \forall e \in \partial E. \tag{18}$$

Notice that the following relation between the functions $\boldsymbol{R}$ and $\boldsymbol{r}_e$ holds

$$\boldsymbol{R} = \sum_{e \in \partial E} \boldsymbol{r}_e. \tag{19}$$

With this modification, Eq. (17) becomes

$$\int_E \phi \frac{\partial u}{\partial t}\, d\Omega + \sum_{e \in \partial E} \int_e \phi h(u, \boldsymbol{\nabla} u + \boldsymbol{r}_e, u^+, \boldsymbol{\nabla} u^+ + \boldsymbol{r}_e^+, \boldsymbol{n})\, d\sigma - \int_E \boldsymbol{\nabla}\phi \cdot \boldsymbol{f}(u, \boldsymbol{\nabla} u + \boldsymbol{R})\, d\Omega + \int_E \phi s(u, \boldsymbol{\nabla} u + \boldsymbol{R}) = 0. \tag{20}$$

The numerical flux function $h(u, \boldsymbol{\nabla} u + \boldsymbol{r}_e, u^+, \boldsymbol{\nabla} u^+ + \boldsymbol{r}_e^+, \boldsymbol{n})$ is the sum of the inviscid numerical flux function $h_c(u, u^+, \boldsymbol{n})$ and of a viscous numerical

flux function h_v given by the average value of the viscous fluxes associated to an interface, i.e.

$$h_v(u, \nabla u + r_e, u^+, \nabla u^+ + r_e^+, n) = \frac{1}{2}\left[f_v(u, \nabla u + r_e) + f_v(u^+, \nabla u^+ + r_e^+)\right] \cdot n. \quad (21)$$

A very interesting feature of the outlined viscous flux discretization scheme is that it couples only the unknowns already coupled by the inviscid flux discretization scheme, irrespective of the degree of polynomial approximation of the solution. This feature is obviously very attractive for an implicit implementation of the method.

When the boundary ∂E of an element is part of σ, the numerical flux function appearing in Eq. (20) must be chosen in order to be consistent with the boundary conditions of the problem. We will denote the inviscid contribution by $h_c^*(u_b)$ and the viscous contribution by $h_v^*(u_b, \nabla u_b)$, where u_b and ∇u_b are "boundary states" computed according to the boundary conditions of the problem. The procedure to determine the values of u_b and of ∇u_b in the case of inflow, outflow and wall (with prescribed temperature) boundaries can be found in [3,5]. Following Wilcox [18], we use the surface roughness model to prescribe the wall value for ω, which is therefore computed as $\omega_w = S_R(k_R^+)u_\tau^2/\nu_w$, where k_R^+ is the nondimensional surface roughness (≤ 5 for an hydraulically smooth surface).

4 Implicit Time Integration

The fully coupled system of the space discretized RANS and k-ω equations is advanced in time with the backward Euler implicit time integration scheme. By denoting the global solution vector as $\boldsymbol{U}$ and the residual vector as $\boldsymbol{R}$, the semidiscrete equations can be written as

$$\boldsymbol{M}\frac{d\boldsymbol{U}}{dt} + \boldsymbol{R}(\boldsymbol{U}) = 0 \ , \quad (22)$$

where $\boldsymbol{M}$ is the block diagonal mass matrix. Linearizing the residual $\boldsymbol{R}^{n+1} = \boldsymbol{R}(\boldsymbol{U}^{n+1})$ in time, the backward Euler scheme is

$$\left[\frac{\boldsymbol{M}}{\Delta t} + \frac{\partial \boldsymbol{R}^n}{\partial \boldsymbol{U}}\right](\boldsymbol{U}^{n+1} - \boldsymbol{U}^n) = -\boldsymbol{R}^n. \quad (23)$$

Note that for very large time steps the scheme is equivalent to the solution of the system of non linear equations $\boldsymbol{R}(\boldsymbol{U}) = 0$ by means of Newton's method.

Eq. (23) implies that a linear system of algebraic equations $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} = 0$ must be solved at each time step. The matrix $\boldsymbol{A}$ can be regarded as a $n \times n$ sparse block matrix, n being the number of elements in the grid. Each block is an $m \times m$ matrix, m being the number of unknown fields (ρ, ρe_0, ρu,

ρv, ρk, $\rho\tilde{\omega}$) times the number of degrees of freedom used to represent each field within an element. Thanks to the inviscid and viscous flux discretization schemes outlined in the previous section, the number of nonzero blocks of a generic row i is equal to the number of elements surrounding element i plus one.

To solve Eq. (23) we have considered the preconditioned GMRES and have used the block diagonal part of $\boldsymbol{A}$, denoted by $\boldsymbol{D}$, as left preconditioner. This choice represents a reasonable compromise between efficiency and storage requirements. Eq. (23) can thus be written as

$$\left(\boldsymbol{I} + \boldsymbol{D}^{-1}\boldsymbol{N}\right)\boldsymbol{x} + \boldsymbol{D}^{-1}\boldsymbol{b} = 0,$$

where $\boldsymbol{N} = \boldsymbol{A} - \boldsymbol{D}$ is the block off-diagonal part of $\boldsymbol{A}$. In our implementation, the inviscid flux, viscous flux and source term Jacobian matrices needed to construct the elements of $\boldsymbol{D}$ and $\boldsymbol{N}$ are computed analitically and take into account the full dependence of the fluxes and of the source term on the unknown u, on its gradient $\boldsymbol{\nabla} u$ and on the functions $\boldsymbol{R}$ and $\boldsymbol{r}_e$.

The contributions to matrices $\boldsymbol{D}$ and $\boldsymbol{N}$ coming from the contour integral of the inviscid flux are relatively straightforward to compute if a flux vector splitting type numerical flux is employed. In this case we have

$$h_c(u, u^+, \boldsymbol{n}) = [\boldsymbol{f}_c^+(u) + \boldsymbol{f}_c^-(u^+)] \cdot \boldsymbol{n},$$

where $\boldsymbol{f}_c^+$ and $\boldsymbol{f}_c^-$ are the split positive and negative fluxes. Linearizing at time level n we obtain

$$h_c^{n+1} = h_c^n + A_n^+(u)^n \Delta u + A_n^-(u^+)^n \Delta u^+ \tag{24}$$

where $\Delta u = u^{n+1} - u^n$ and $A_n^{\pm}$ are the Jacobian matrices of the split fluxes in the normal direction $\boldsymbol{n} \cdot \partial \boldsymbol{f}_c^{\pm} / \partial u$.

The implicit treatment of the viscous flux introduces an additional difficulty essentially related to the interface discontinuity contributions to derivatives which must be considered in the evaluation of the viscous flux function. The linearized viscous flux at time level $n+1$ appearing in the volume integral is

$$\boldsymbol{f}_v^{n+1} = \boldsymbol{f}_v^n + \boldsymbol{B}(u, \boldsymbol{\nabla} u + \boldsymbol{R})\Delta u + \boldsymbol{C}(u)\Delta(\boldsymbol{\nabla} u + \boldsymbol{R}), \tag{25}$$

$\boldsymbol{B}$ and $\boldsymbol{C}$ denote the Jacobian matrices $\partial \boldsymbol{f}_v / \partial u$ and $\partial \boldsymbol{f}_v / \partial(\boldsymbol{\nabla} u + \boldsymbol{R})$, respectively. By using Eq. (15) the degrees of freedom of the function $\boldsymbol{R}$ can be expressed in terms of the degrees of freedom of functions u and u^+. This implies that the term $\Delta(\boldsymbol{\nabla} u + \boldsymbol{R})$ in Eq. (25) can be entirely expressed in terms of the original variables u.

Considering now the viscous numerical flux of contour integrals in the normal direction $\boldsymbol{n}$ linearized at time level n we have

$$h_v^{n+1} = h_v^n + \frac{1}{2}\left[\boldsymbol{B}(u, \boldsymbol{\nabla} u + \boldsymbol{r}_e)\Delta u + \boldsymbol{B}(u^+, \boldsymbol{\nabla} u^+ + \boldsymbol{r}_e^+)\Delta u^+\right] \cdot \boldsymbol{n} + \frac{1}{2}\left[\boldsymbol{C}(u)\Delta(\boldsymbol{\nabla} u + \boldsymbol{r}_e) + \boldsymbol{C}(u^+)\Delta(\boldsymbol{\nabla} u^+ + \boldsymbol{r}_e^+)\right] \cdot \boldsymbol{n}. \tag{26}$$

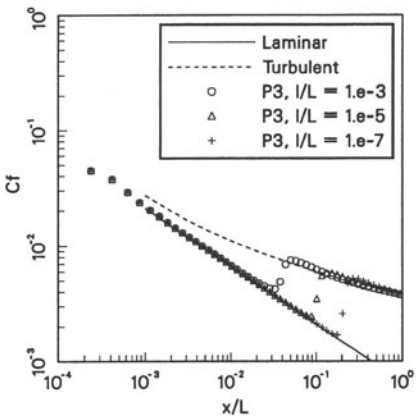


Fig. 1. Skin friction coefficient for $\ell = 10^{-3}L$, $10^{-5}L$, and $10^{-7}L$, P3 solutions.

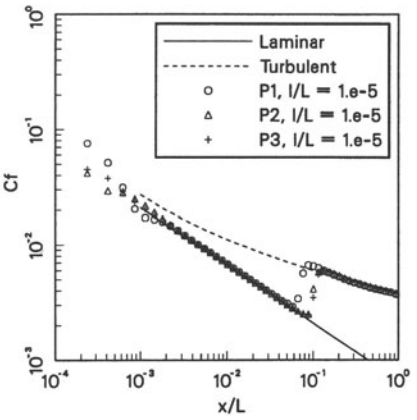


Fig. 2. Skin friction coefficient for $\ell = 10^{-5}L$, P1, P2, and P3 solutions.

The values of $\boldsymbol{r}_e$ and $\boldsymbol{r}_e^+$ appearing in Eq. (26) can be expressed in terms of u and u^+ in a way analogous to that described for Eq. (25).

The implicit treatment of boundary conditions is based on linearized relations which give the outer state u^+ and the outer gradient $\boldsymbol{\nabla} u^+ + \boldsymbol{r}_e^+$ as a function of the boundary data and of the internal state u and of the internal gradient $\boldsymbol{\nabla} u + \boldsymbol{r}_e$. These expressions are then introduced into the inviscid and viscous flux functions which therefore become functions of the internal state and of the boundary data only.

5 Numerical Results

The transitional flow over a flat plate has been used as a first test case for the proposed method. The Mach number ahead of the plate is $\mathrm{M}_{in} = 0.3$, and the Reynolds number based on the plate length is $\mathrm{Re}_L = 10^6$. The freestream turbulence intensity is $\mathrm{Tu}_{in} = 0.03$. The grid is the triangulation of a structured grid having 51 points in the streamwise direction and 17 points in the normal direction. The grid points are clustered both near the leading edge and near the wall. The wall spacing ranges from $y^+ = 3$ to 5. Three different inflow values of turbulence specific dissipation rate ω_{in} have been considered, and the computations have been performed with linear (P1), quadratic (P2), and cubic (P3) elements. The values of ω_{in} correspond to three different turbulent length scales $\ell = 10^{-3}L$, $10^{-5}L$, and $10^{-7}L$, where $\ell = \sqrt{k}/(\beta^*\omega) = k^{3/2}/\epsilon$. The symbols in the following plots represent average values on the edges lying on the wall. Fig. 1 shows the value of the skin friction coefficient along the plate computed with P3 elements and with increasing values of ω_{in}. The predicted onset of transition moves downstream

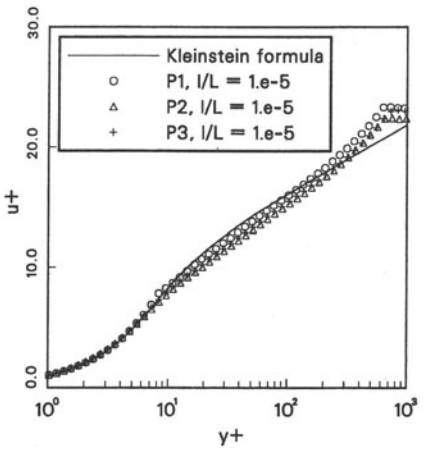


Fig. 3. Velocity profile for $x/L = 0.95$ and $\ell = 10^{-5}L$, P1, P2, and P3 solutions.

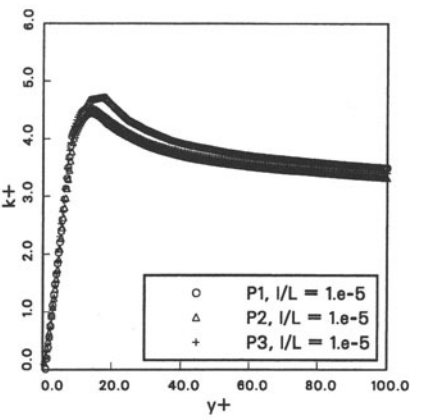


Fig. 4. Turbulent kinetic energy profile for $x/L = 0.95$ and $\ell = 10^{-5}L$, P1, P2, and P3 solutions.

with decreasing values of length scale ℓ, a result consistent with the expected behavior of the k-ω model with this level of freestream turbulence intensity (see e.g. Wilcox [18]). Fig. 2 shows the value of the skin friction coefficient for $\ell = 10^{-5}L$ computed with elements of increasing order of accuracy (P1, P2, and P3). It appears that the transition takes place a little too early for the P1 solution whilst it is virtually the same for the P2 and P3 solutions.

The near wall behavior of P1, P2 and P3 solutions is presented in terms of velocity and turbulence properties profiles. The quantities $u^+ = u/u_\tau$, $k^+ = k/u_\tau^2$, and $\epsilon^+ = \beta^* k^+ \omega^+$ (where $u_\tau = \sqrt{\tau_w/\rho_w}$ and $\omega^+ = \omega\nu_w/u_\tau^2$) are plotted as functions of $y^+ = yu_\tau/\nu_w$ at $x/L = 0.95$ and for $\ell = 10^{-5}L$. Notice that for these profile plots the symbols appearing in the Figures represent the high order polynomial solution inside the elements and not simply the element averages. Fig. 3 shows that all the computed velocity profiles compare fairly well with the composite velocity profile computed with the Kleinstein formula reported in [17]. The turbulent kinetic energy profiles, reported in Fig. 4, display plateau and peak values of k^+ which are in very good agreement with the most representative ones reported in [16]. The P1 solution, however, shows that linear elements are not accurate enough to represent the k^+ profile. Finally, Fig. 5 displays marked differences among the P1, P2 and P3 turbulence dissipation rate profiles in the near wall region. A comparison with high order computations on a more refined grid, not reported here, shows that only the P3 solution is adequate.

For the second test case we have considered the transonic turbine vane with wall heat transfer tested by Arts et al. [1]. We have selected the test case denoted as MUR228 which is characterized by $M_{2is} = 0.932$, $Re_{2is} = 0.595 \times 10^6$, and $Tu_{in} = 0.01$. In this test case the vane temperature is

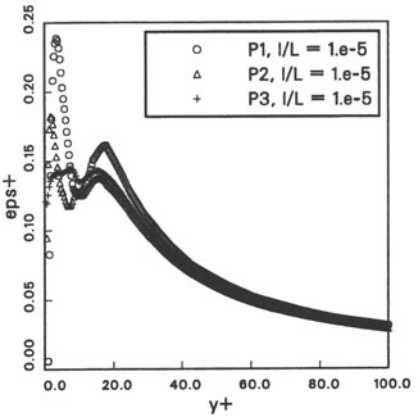


Fig. 5. Turbulence dissipation rate profile for $x/L = 0.95$ and $\ell = 10^{-5}L$, computed with P1, P2, and P3 elements.

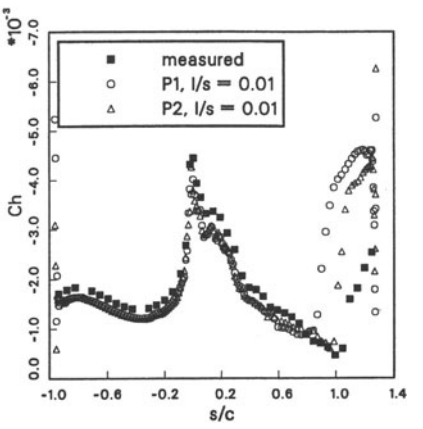


Fig. 6. Heat flux coefficient distribution for the LS89 turbine vane, P1 and P2 solutions.

$T_w = 302.85$ K and the inflow stagnation temperature is $T_{01} = 403.3$ K. The computational grid is the triangulation of a structured C-type grid having 193×17 points. There are 145 nodes on the blade. The wall spacing ranges from $y^+ = 3$ to 6. The test case has been run with P1 and P2 elements. The inflow value of turbulence specific dissipation rate corresponds to $\ell = 10^{-2}g$, where g is the blade spacing.

Fig. 6 shows the blade surface distribution of the heat transfer coefficient defined as $C_h = q_w/0.5(\rho v^3)_{2,is}$. According to the experiments, the computed flow field is close to a laminar state over most of the airfoil surface and the turbulence induces an abrupt increase of heat flux only in the rear part of the suction surface. However, the comparison between computational results and experimental data shows that the heat flux is underpredicted where the flow is close to a laminar one and that the transition on the suction surface is predicted a little too early. As observed by Chima in [10], both discrepancies can be ascribed to the turbulence model characteristics. Figs. 7 and 8 show the global view and the trailing edge detail of the Mach number and of the turbulent kinetic energy isolines computed with P2 elements. The flow appears to be unsteady behind the trailing edge and in fact the solution had to be computed in an unsteady fashion even though using the low order implicit backward Euler method.

6 Conclusions

In this paper we have presented an implicit high order DG method to compute turbulent compressible flows. The equation for ω in the k-ω turbulence model has been rewritten in terms of $\log\omega$ in order to improve the stability of

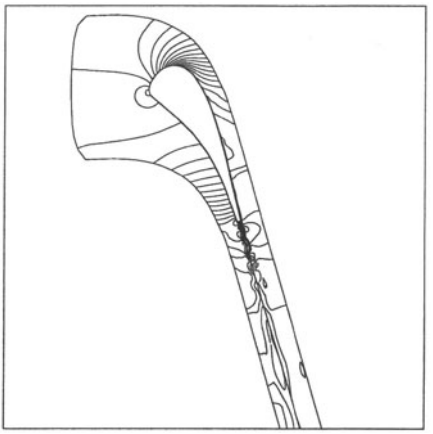


Fig. 7. Mach number isolines for the LS89 turbine vane, P2 solution.

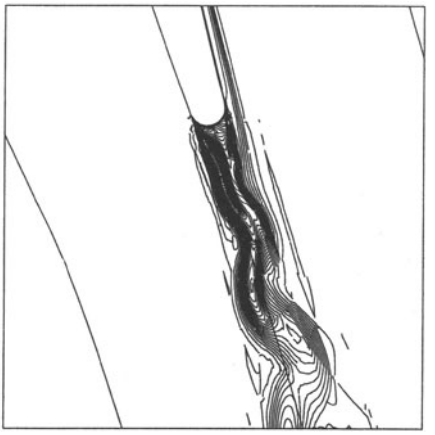


Fig. 8. TE detail of the turbulent kinetic energy isolines, P2 solution.

the method, especially with high order approximations. An implicit time discretization of the fully coupled RANS and k-ω equations has been found useful to overcome the stringent time step size limitations of explicit time integration schemes applied to the high order DG approximation.

Two test cases of transitional flow have been used to verify the code. As expected, numerical results improve by increasing the polynomial degree of approximate solution. The computed test cases show that the high order accuracy of the method allows to compute turbulent flows on relatively coarse grids.

Work is in progress to fully exploit the potentialities of the method by resorting to adaptive techniques.

Acknowledgement

We would like to thank Enel Spa Polo Termico for the financial support in the development of the 2d and 3d codes based on the DG method.

References

1. T. Arts, M. Lambert de Rouvroit, and A. W. Rutherford. Aero-thermal investigation of a highly loaded transonic linear turbine guide vane cascade. Technical Note 174, Von Karman Institute for Fluid Dynamics, September 1990.
2. F. Bassi and S. Rebay. A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier–Stokes equations. *J. Comput. Phys.*, 131:267–279, 1997.
3. F. Bassi and S. Rebay. High-order accurate discontinuous finite element solution of the 2D Euler equations. *J. Comput. Phys.*, 138:251–285, 1997.

4. F. Bassi and S. Rebay. An implicit high-order discontinuous galerkin method for the steady state compressible Navier–Stokes equations. In K. D. Papailiou, D. Tsahalis, D. Périaux, C. Hirsh, and M. Pandolfi, editors, *Computational Fluid Dynamics 98, Proceedings of the Fourth European Computational Fluid Dynamics Conference*, volume 2, pages 1227–1233, Athens, Greece, September 5–7 1998. John Wiley and Sons.
5. F. Bassi, S. Rebay, G. Mariotti, S. Pedinotti, and M. Savini. A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows. In R. Decuypere and G. Dibelius, editors, *2nd European Conference on Turbomachinery Fluid Dynamics and Thermodynamics*, pages 99–108, Antwerpen, Belgium, March 5–7 1997. Technologisch Instituut.
6. C. E. Baumann and J. T. Oden. A discontinuous hp finite element method for the Euler and Navier-Stokes equations. to appear on IJNME, 1998.
7. F. Brezzi, G. Manzini, D. Marini, P. Pietra, and A. Russo. Discontinuous finite elements for diffusion problems. Technical Report 1112, IAN-CNR, via Ferrata, 1, 1998. submitted for publication to Numer. Meth. PDEs.
8. F. Brezzi, G. Manzini, D. Marini, P. Pietra, and A. Russo. Discontinuous galerkin approximations for elliptic problems. Technical Report 1110, IAN-CNR, via Ferrata, 1, 1998. to appear in "Atti del Convegno in Memoria di F. Brioschi, Istituto Lombardo di Scienze e Lettere, Milano, 22/23 ottobre 1997".
9. F. Brezzi, G. Manzini, D. Marini, P. Pietra, and P. Russo. Analisi delle proprietà di elementi finiti di tipo discontinuo. Technical Report 1107, IAN-CNR, via Ferrata, 1, 1998. relazione finale del progetto ENEL-MIGALE.
10. R. V. Chima. A k-ω turbulence model for quasi three-dimensional turbomachinery flows. AIAA Paper 96-0248, ICASE, 1996. Also NASA TM-107051.
11. B. Cockburn, S. Hou, and C.-W. Shu. The Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: The multidimensional case. *Math. Comp.*, 54:454–581, 1990.
12. B. Cockburn and C.-W. Shu. The local discontinuous Galerkin method for time-dependent convection-diffusion systems. SIAM J. Numer. Anal. 35:2440-2463, 1998.
13. G. A. Gerolymos and I. Vallet. Implicit computation of the three-dimensional compressible Navier–Stokes equations. *AIAA Journal*, 34(7):1321–1330, July 1996.
14. D. Hanel, R. Schwane, and G. Seider. On the accuracy of upwind schemes for the solution of the Navier–Stokes equations. AIAA Paper 87-1105 CP, AIAA, July 1987. Proceedings of the AIAA 8th Computational Fluid Dynamics Conference.
15. F. Ilinca and D. Pelletier. Positivity preservation and adaptive solution for the k-ϵ model of turbulence. *AIAA Journal*, 36(1):44–50, January 1996.
16. V. C. Patel, W. Rodi, and G. Scheuerer. Turbulence models for near-wall and low Reynolds number flows: a review. *AIAA Journal*, 23(9):1308–1318, September 1985.
17. F. M. White. *Viscous fluid flow.* McGraw-Hill, 1974.
18. D. C. Wilcox. *Turbulence Modelling for CFD.* DCW industries Inc., La Cañada, CA 91011, USA, 1993.

Discontinuous Galerkin Methods for Elliptic Problems

Douglas N. Arnold[1], Franco Brezzi[2], Bernardo Cockburn[3], and Donatella Marini[2]

[1] Department of Mathematics, Penn State University, University Park, PA 16802, USA
[2] Dipartimento di Matematica and I.A.N.-C.N.R. Via Ferrata 1, 27100 Pavia, Italy
[3] School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455, USA

Abstract. We provide a common framework for the understanding, comparison, and analysis of several discontinuous Galerkin methods that have been proposed for the numerical treatment of elliptic problems. This class includes the recently introduced methods of Bassi and Rebay (together with the variants proposed by Brezzi, Manzini, Marini, Pietra and Russo), the local discontinuous Galerkin methods of Cockburn and Shu, and the method of Baumann and Oden. It also includes the so-called interior penalty methods developed some time ago by Douglas and Dupont, Wheeler, Baker, and Arnold among others.

1 Introduction

In 1973, Reed and Hill [21] introduced the first discontinuous Galerkin (DG) method for hyperbolic equations, and since that time there has been an active development of DG methods for hyperbolic and nearly hyperbolic problems, resulting in a variety of different methods. Also in the 1970's, but independently, Galerkin methods for elliptic and parabolic equations using discontinuous finite elements were proposed, and a number of variants introduced and studied. These were generally called *interior penalty* (IP) methods and their development remained independent of the development of the DG methods for hyperbolic equations. In this paper, we provide a common framework which includes nearly all the DG methods that have been proposed thus far.

We briefly review the development of penalty methods for elliptic and parabolic equations. Penalties were first introduced into the finite element method as a mean for imposing Dirichlet boundary conditions *weakly* rather than incorporating the boundary conditions into the finite element space. Let us begin by recalling Nitsche's method [19] for the model problem $-\Delta u = f$ in Ω, $u = 0$ on $\partial\Omega$. Clearly

$$\int_\Omega \nabla u \cdot \nabla v \, dx - \int_{\partial\Omega} \frac{\partial u}{\partial n} v \, ds = \int_\Omega f v \, dx,$$

for all sufficiently smooth test functions v. Since u vanishes on the boundary, we have as well that $B(u,v) = \int fv\,dx$, where

$$B(u,v) := \int_\Omega \nabla u \cdot \nabla v\,dx - \int_{\partial\Omega} \frac{\partial u}{\partial n} v\,ds - \int_{\partial\Omega} \frac{\partial v}{\partial n} u\,ds + \int_{\partial\Omega} \eta uv\,ds, \quad (1)$$

for any weighting function η. Nitsche's method then determines an approximate solution u_h in a finite element subspace of $H^1(\Omega)$ such that $B(u_h, v_h) = \int fv_h\,dx$ for all v_h in the same space. Note that the second term of the bilinear form B arose to ensure that the method is consistent. The third term was added so that the discrete problem is symmetric (and so the method is truly variational—the discrete solution minimizes $B(u,u)/2 - \int fu$ over the finite element space). Finally, the last term is the penalty term, which is necessary to guarantee stability. Nitsche proved that if η is taken as C/h where h is the element size and C is a sufficiently large constant, then the discrete solution converges to the exact solution with optimal order in H^1 and L^2.

A different penalty method for imposing Dirichlet boundary conditions is due to Babuška [2]. He does not include either the second or third term in (1), and uses as the penalty weight $h^{-\sigma}$ for some $\sigma \geq 0$. Because of the missing consistency term, his method, and its analysis, includes a consistency error. Another interesting possibility is to include all the terms in (1) but to reverse the sign of the third term in B. The bilinear form is then no longer symmetric, but it has a favorable coercivity property, namely, $B(u,u) \geq \int |\nabla u|^2$, no matter how $\eta \geq 0$ is chosen.

The IP methods arose from the observation that, just as Dirichlet boundary conditions could be imposed weakly instead of being built into the finite element space, so interelement continuity could be attained in a similar fashion. This makes it possible to use spaces of discontinuous piecewise polynomials for solving second order problems. The natural generalization of Nitsche's method to this context (in which there are consistency, symmetrization, and penalty terms on each edge, the latter penalizing the jump of the function across the edge) is stated in Wheeler's 1978 paper on IP collocation–finite element methods, [27], where it is attributed to a private communication of Douglas and Dupont. That method is analyzed in detail for linear and nonlinear elliptic and parabolic problems in the 1979 thesis of Arnold which is summarized in [1]. Interior penalties of this sort were also used by Baker [4] for imposing C^1 interelement continuity on C^0 elements for fourth order problems. In these, of course, it is the jump in the normal derivative that is penalized. In 1976, Douglas and Dupont [16] penalized the jump in the normal derivative of C^0 elements for second order elliptic and parabolic problems, with the goal of enforcing a degree of continuity in some sense intermediate between C^0 and C^1. Babuška and Zlámal [3], like Baker, used interior penalties to weakly impose C^1 continuity for fourth order problems, but their bilinear form is analogous to Babuška's finite element with penalty rather than to the bilinear form of Nitsche's method, i.e., it does not have the consistency and symmetry terms.

Not so much attention has been paid to IP methods since the early 1980's, although they have found a few new applications. In 1990, Baker, Jureidini and Karakashian [5] used interior penalties to enforce continuity on piecewise solenoidal vector fields for solving the Stokes equations. In the same year, Rusten, Vassilevski, and Winther [24] used an interior penalty method for second order elliptic problems as part of a preconditioner for mixed methods. Recently, Becker and Hansbo [10] used the IP approach as a way to enforce continuity across non-matching grids for domain decomposition.

On the other hand, DG methods for the numerical treatment of nonlinear hyperbolic systems experienced a vigorous development during the last ten years due to a strong interaction with the ideas of finite volumes methods for hyperbolic problems; see a review of this development in [14]. But the evolution of the DG methods did not stop there. The necessity of dealing with problems that, together with a dominant convective part, had a non-negligible diffusive part, prompted several authors to extend the DG methods to elliptic problems. Thus in 1997, Bassi and Rebay [6] introduced a DG method for the Navier-Stokes equations and in 1998, Cockburn and Shu [15] introduced the so-called local discontinuous Galerkin (LDG) methods by generalizing the original DG method of Bassi and Rebay. Around the same time, Oden and Bauman [8], [9] introduced another DG method for diffusion problems. Their approach uses a non-symmetric bilinear form, even for symmetric problems, analogous to the one obtained from Nitsche's penalty form by reversing the sign of the symmetrization term, as discussed earlier.

It was at this point that several authors were struck by the similarities between those recently introduced DG methods and the old IP methods and started to apply to the former the old techniques of analysis used on the latter. Thus, Brezzi et al. [12] studied several variations of the original method of Bassi and Rebay; Oden, Babuška and Baumann [20] studied the DG method of Baumann and Oden; Rivière and Wheeler [23], [22] analyzed several variations of the DG method of Baumann and Oden; and Süli, Schwab, and Houston [25], [26] synthesized the elliptic, parabolic, and hyperbolic theory by extending the analysis of DG methods to partial differential equations with non-negative characteristic form. Our long term goal is to follow this trend and produce a comprehensive study of the above mentioned methods as applied to elliptic problems. In this note, we recast *all* of the above mentioned methods within a single framework in order to lay down a basis for a better understanding of the connections among them, and, eventually, a unified analysis, that however we postpone to a subsequent paper.

An outline of the paper is as follows. For the sake of simplicity and clarity, we present our unified framework for the classical problem of the Laplacian with homogeneous Dirichlet boundary conditions. In § 2, we provide a general framework for its discretisation by means of DG methods. This framework is an extension of the approach used by Cockburn and Shu [15] to define the LDG methods and allows us to include methods that are not LDG methods, like the IP methods and the DG method of Baumann and Oden. In the

next two sections we verify that indeed many known methods fall within our framework, and we present a partial classification of methods. A table listing all these methods is included in the final section.

2 The general DG method for a model problem

For the sake of simplicity, we restrict ourselves to the following model problem:

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega,$$

where Ω is assumed to be a polygonal domain and f a given function in $L^2(\Omega)$. To obtain the weak formulation upon which the discretization is based, we rewrite the above problem as follows:

$$\sigma = \nabla u, \quad -\nabla \cdot \sigma = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega.$$

Let K be the closure of an open subset of Ω with a piecewise smooth boundary. If we multiply the above equations by test functions and integrate formally on K, we get

$$\int_K \sigma \cdot \tau \, dx = -\int_K u \nabla \cdot \tau \, dx + \int_{\partial K} u n_K \cdot \tau \, ds,$$
$$\int_K \sigma \cdot \nabla v \, dx = \int_K f v \, dx + \int_{\partial K} \sigma \cdot n_K v \, ds,$$

where n_K is the outward normal unit vector to ∂K. This is the weak formulation we sought. We are now ready to define the DG method.

We denote by $\mathcal{T}_h$ a triangulation of Ω in polygons K, and by $P(K)$ a finite dimensional space of smooth functions, typically polynomials, defined on the polygon K. This space will be used to approximate the variable u. We denote by $\Sigma(K)$ another finite dimensional space of smooth functions that we are going to use in order to approximate the auxiliary variable σ. Setting

$$V_h := \{ v \in L^2(\Omega) \quad | \quad v|_K \in P(K) \quad \forall K \in \mathcal{T}_h \},$$
$$\Sigma_h := \{ \tau \in (L^2(\Omega))^2 \quad | \quad \tau|_K \in \Sigma(K) \quad \forall K \in \mathcal{T}_h \},$$

and following Cockburn and Shu [15], we consider the following general weak formulation: Find $u_h \in V_h$ and $\sigma_h \in \Sigma_h$ such that $\forall K \in \mathcal{T}_h$ we have

$$\int_K \sigma_h \cdot \tau \, dx = -\int_K u_h \nabla \cdot \tau \, dx + \sum_{e \subset \partial K} \int_e h_u^{e,K} n_K \cdot \tau \, ds \quad \forall \tau \in \Sigma(K), \quad (2)$$

$$\int_K \sigma_h \cdot \nabla v \, dx = \int_K f v \, dx + \sum_{e \subset \partial K} \int_e h_\sigma^{e,K} \cdot n_K v \, ds \qquad \forall v \in P(K), \quad (3)$$

where the sums are taken over the edges e of the polygon K, and the *numerical fluxes* $h_\sigma^{e,K}$ and $h_u^{e,K}$ are approximations to $\sigma|_e = \nabla u|_e$ and to $u|_e$, respectively, on the edges of the triangulation. In order to complete the definition of a method we must provide the polynomial spaces $P(K)$ and $\Sigma(K)$ and the formula for the numerical fluxes $h_\sigma^{e,K}$ and $h_u^{e,K}$ in terms of σ_h and u_h. The choice of spaces will not play a large role in our study. For triangular elements, one could, for example, take $P(K)$ to consist of all polynomials of degree $p \geq 1$ and $\Sigma(K)$ to consist of all polynomial vector fields of degree $p-1$ or p. The choice of the consititutive relations defining the fluxes, on the other hand, will be crucial. The flux choices affect the stability and the accuracy of the method, as well as properties such as sparsity and symmetry of the stiffness matrix; cf. [15] and [13]. As we shall see, different choices for the fluxes will lead to the different methods that we are going to discuss.

Next, we discuss some basic properties that are shared by all the flux choices.

1. Locality. Let $K = K_1$ be an element in the triangulation, and let e be one of its edges. Assume first that e is an interior edge of our triangulation, so that there is a second element K_2 sharing the edge e with K_1. We then assume that $h_\sigma^{e,K}$ and $h_u^{e,K}$ depend on the restrictions $u_h|_{K_i}$ and $\sigma_h|_{K_i}$ of u_h and σ_h to K_i, $i = 1, 2$. More precisely, locality means that

$$h_\sigma^{e,K} = h_\sigma^{e,K}(u_h|_{K_1}, \sigma_h|_{K_1}, u_h|_{K_2}, \sigma_h|_{K_2}).$$

Actually, in all our examples, this fucntional dependence will have a special form in that both $h_\sigma^{e,K}$ and $h_u^{e,K}$ will depend only on the traces of $u_h|_{K_i}$, $\nabla u_h|_{K_i}$, and $\sigma_h|_{K_i}$ on the edge e. Since u_h, ∇u_h, and σ_h will, in general, be discontinuous across e, the trace of $u_h|_{K_1}$ on e will be different from the trace of $u_h|_{K_2}$ on e, and similarly ∇u_h and σ_h will each have two different traces on e. Thus $h_\sigma^{e,K}$ and $h_u^{e,K}$ will depend linearly on the six quantities

$$(u_h|_{K_1})|_e,\ (\nabla u_h|_{K_1})|_e,\ (\sigma_h|_{K_1})|_e,\ (u_h|_{K_2})|_e,\ (\nabla u_h|_{K_2})|_e,\ (\sigma_h|_{K_2})|_e.$$

In our particular case of a homogeneous Dirichlet problem, the fluxes on boundary edges will have the same functional dependence on these six traces, provided we interpret the traces coming from K_2 as follows: $(u_h|_{K_2})|_e = 0$, $(\nabla u_h|_{K_2})|_e = (\nabla u_h|_{K_1})|_e$, and $(\sigma_h|_{K_2})|_e = (\sigma_h|_{K_1})|_e$. Other boundary conditions can be handled easily as well, but, in order to keep the notation as simple as possible, we shall not discuss these here.

Finally, it is important to note that in all the methods we are going to analyze, $h_u^{e,K}$ will not depend on $\sigma_h|_{K_i}$ (nor on $\nabla u_h|_{K_i}$, but that will be less important). This, as we shall see, will allow us to eliminate the variable σ_h at the element level, often with a considerable computational saving.

2. Consistency. All the methods we consider are consistent in the sense that, in the functional form described above,

$$h_\sigma^{e,K}(u|_{K_1}, \nabla u|_{K_1}, u|_{K_2}, \nabla u|_{K_2}) = \nabla u|_e,$$
$$h_u^{e,K}(u|_{K_1}, \nabla u|_{K_1}, u|_{K_2}, \nabla u|_{K_2}) = u|_e,$$

whenever u is a smooth function satisfying the boundary conditions.

3. Conservation. All our methods satisfy

$$h_\sigma^{e,K_1} = h_\sigma^{e,K_2} \tag{4}$$

when e is an edge shared by elements K_1 and K_2, and so we may write simply h_σ^e. This is a conservation property: if S is the union of some collection of elements, then, taking v to be identically unity in (3) and adding over K contained in S we get

$$\int_S f\,dx + \sum_{e \subset \partial S} \int_e h_\sigma^e \cdot n\,ds = 0.$$

We close this section with several additional remarks concerning the above properties.

1. As we have seen, if $h_u^{e,K}$ does not depend on σ_h, then the auxiliary variable σ_h can be eliminated *locally* in terms of u_h and ∇u_h, using (2). When using triangles, the use of the orthonormal Dubiner basis [17] renders this elimination trivial. See also the extensions to 3D elements by Lomtev and Karniadakis [18].
2. In all the methods we consider, h_σ^e depends either on the traces of ∇u_h or on those of σ_h, but not on both. The former category, for which the stiffness matrix tends to be much sparser, includes the IP methods and the method of Baumann and Oden; we discuss this category of methods in § 4. The latter category, which we discuss in § 3, includes the LDG family of methods.
3. Most of the methods will satisfy, in addition to the conservation property (4), the analogous property $h_u^{e,K_1} = h_u^{e,K_2}$ (in which case we write h_u^e for $h_u^{e,K}$.) We shall refer to them as *completely conservative* methods. As we shall see, generally only completely conservative methods lead to a symmetric stiffness matrix after elimination of σ_h. Except for the methods of Baumann and Oden, and the so-called *pure penalty methods* discussed at the end of § 4, all the methods we consider are completely conservative.
4. We also note that, in view of (3), only the normal component $h_\sigma^{e,K} \cdot n_K$ of $h_\sigma^{e,K}$ enters the methods; its tangential component is irrelevant. In practice, the normal component will depend only on the normal traces.

3 Numerical fluxes independent of ∇u_h

In order to describe the flux functions for various methods we need to introduce some notation. Again let e be an edge shared by elements K_1 and K_2. Define also the normal vectors n_1 and n_2 on e pointing exterior to K_1 and

K_2, respectively. If v is a function on $K_1 \cup K_2$, but possibly discontinuous across e, let v_i denote $(v|_{K_i})|_e$, $i = 1, 2$. For a scalar function v we then define

$$\{v\} := \frac{1}{2}(v_1 + v_2), \qquad [\![v]\!] := v_1 n_1 + v_2 n_2.$$

If τ is a vector-valued function, we set

$$\{\tau\} := \frac{1}{2}(\tau_1 + \tau_2), \qquad [\![\tau]\!] := \tau_1 \cdot n_1 + \tau_2 \cdot n_2.$$

Notice that the jump $[\![v]\!]$ of the scalar function v is a vector parallel to n and that $[\![\tau]\!]$ is the jump of the normal component of the vector function τ- it is hence is a scalar quantity. The advantage of these definitions is that they do not depend on assigning an ordering to the elements K_i.

In this section, we consider the DG methods determined by the following choice of numerical fluxes:

$$h_\sigma^{e,K} = \{\sigma_h\} - \alpha^e([\![u_h]\!]) + \beta^e [\![\sigma_h]\!], \qquad h_u^{e,K} = \{u_h\} + \gamma^e \cdot [\![u_h]\!]. \tag{5}$$

Here β^e and γ^e are vector-valued functions on e. Often they are constant, and, indeed, for many methods they both vanish. The term $\alpha^e([\![u_h]\!])$ could simply be taken to be

$$\alpha^e([\![u_h]\!]) = \eta^e [\![u_h]\!] \tag{6}$$

for some constant (or function) η^e. Another possibility arises from the work of Bassi, Rebay, Mariotti, Pedinotti, and Savini [7]. Namely, we define the operator $r_e : L^1(e) \to \Sigma_h$ by

$$\int_\Omega r_e(q) \cdot \tau \, dx = -\int_e q \cdot \{\tau\} \, ds \qquad \forall \tau \in \Sigma_h, q \in L^1(e),$$

and set

$$\alpha^e([\![u_h]\!]) = \eta^e \{r_e([\![u_h]\!])\}. \tag{7}$$

First we rewrite the method by inserting the flux formulas (5) into the Galerkin equations (2)–(3) and adding over $K \in \mathcal{T}_h$. Denoting by $\mathcal{E}_h$ the set of all element edges, after simple algebraic manipulations we obtain

$$\int_\Omega \sigma_h \cdot \tau \, dx = \sum_K \int_K \nabla u_h \cdot \tau \, dx + \sum_{e \in \mathcal{E}_h} \int_e (\gamma^e \cdot [\![u_h]\!][\![\tau]\!] - [\![u_h]\!] \cdot \{\tau\}) \, ds, \tag{8}$$

$$\sum_K \int_K \sigma_h \cdot \nabla v \, dx = \int_\Omega f \, v \, dx + \sum_{e \in \mathcal{E}_h} \int_e \Big(\{\sigma_h\} - \alpha^e([\![u_h]\!]) + \beta^e [\![\sigma_h]\!] \Big) \cdot [\![v]\!] \, ds \tag{9}$$

for all $\tau \in \Sigma_h$, $v \in V_h$. If we take all the α^e, β^e, and γ^e to vanish, we recover the original DG method of Bassi and Rebay, cf. [6], formulae (13) and (15), and also [11], equations (12) and (14). This method can be unstable, at least for uniform meshes; see [11]. However, stability is achieved if α^e is a positive operator. Defining α^e by (7) with $\eta^e > 0$ (and still with β^e and γ^e zero) gives the variant of the method of Bassi and Rebay [6], as proposed by Brezzi et al. [12], formula (24). Defining α^e by (6), $\eta^e > 0$ gives the LDG methods (which allow general β^e and γ^e).

Next, we eliminate σ_h to rewrite the method in terms of u_h alone (this is usually the preferred implementation in practice). To do so, we define two operators, R and L. The operator $R : V_h \to \Sigma_h$ is given by $R(v) = \sum_{e\in\mathcal{E}_h} r_e([\![v]\!])$, or, equivalently,

$$\int_\Omega R(\varphi)\cdot\tau\,dx = -\sum_{e\in\mathcal{E}_h}\int_e [\![\varphi]\!]\{\tau\}\,ds \qquad \forall\tau\in\Sigma_h. \tag{10}$$

The operator $L : L^1(\bigcup \mathcal{E}_h) \to \Sigma_h$ is given by

$$\int_\Omega L(\varphi)\cdot\tau\,dx = \sum_{e\in\mathcal{E}_h}\int_e \varphi\,[\![\tau]\!]\,ds \qquad \forall\tau\in\Sigma_h. \tag{11}$$

Denoting by P_Σ the L^2-projection onto Σ_h, we can now rewrite equation (8) as

$$\sigma_h = P_\Sigma(\nabla u_h) + R(u_h) + L(\gamma\cdot[\![u_h]\!]), \tag{12}$$

and equation (9) as

$$\sum_K\int_K \sigma_h\cdot\nabla v\,dx = \int_\Omega f\,v\,dx + \int_\Omega \sigma_h\cdot(-R(v)+L(\beta\cdot[\![v]\!])) - \sum_{e\in\mathcal{E}_h}\int_e \alpha^e([\![u_h]\!])\cdot[\![v]\!]\,ds. \tag{13}$$

Here we mean by β and γ the functions on $\bigcup\mathcal{E}_h$ which are given by β^e and γ^e, respectively, on each edge e. Finally, inserting (12) in (13), we get

$$\sum_K\int_K \Big(P_\Sigma(\nabla u_h)+R(u_h)+L(\gamma\cdot[\![u_h]\!])\Big)\cdot\Big(\nabla v + R(v) - L(\beta\cdot[\![v]\!])\Big)\,dx + \sum_{e\in\mathcal{E}_h}\int_e \alpha^e([\![u_h]\!])\cdot[\![v]\!]\,ds = \int_\Omega f\,v\,dx. \tag{14}$$

Note that the second sum on the left-hand side of (14) is symmetric with respect to u_h and v. Indeed

$$\sum_{e\in\mathcal{E}_h}\int_e \alpha^e([\![u_h]\!])\cdot[\![v]\!]\,ds = \begin{cases} \displaystyle\sum_{e\in\mathcal{E}_h}\int_e \eta^e[\![u_h]\!]\cdot[\![v]\!]\,ds, & \text{if } \alpha^e \text{ is defined by (6),} \\ \displaystyle\sum_{e\in\mathcal{E}_h}\int_\Omega \eta^e r_e([\![u_h]\!])\cdot r_e([\![v]\!])\,dx, & \text{if } \alpha^e \text{ is defined by (7).} \end{cases}$$

It is thus clear that a symmetric stiffness matrix is obtained if we choose $\beta^e = -\gamma^e$ for all e. This choice was suggested by Cockburn and Shu [15] for the LDG methods.

In practice the inclusion $\nabla P(K) \subset \Sigma(K)$ generally holds. In that case the projection P_Σ is not needed in (14).

Finally, we remark that if the support of v is contained in a single element K, then the support of $R(v)$ will generally contain all the elements that contain an edge of K. Consequently the product $R(u_h)\cdot R(v)$ in (14) will generally have a big negative impact on the sparsity of the stiffness matrix. This problem is much less severe when the numerical fluxes are independent of σ_h.

4 Numerical fluxes independent of σ_h

First we consider, instead of (5), the following numerical fluxes:

$$h_\sigma^{e,K} = \{\nabla u_h\} - \alpha^e([\![u_h]\!]) + \beta^e[\![\nabla u_h]\!], \qquad h_u^{e,K} = \{u_h\} + \gamma^e[\![u_h]\!],$$

where β^e and γ^e are still vector-valued functions on e. Let us proceed now to the elimination of the variable σ_h as we did at the end of the previous section. By using the definitions of R and L, (10) and (11), respectively, a simple computation gives us that

$$\begin{aligned}\sum_K\int_K \big(P_\Sigma(\nabla u_h) + R(u_h) + L(\gamma\cdot[\![u_h]\!])\big)\cdot\nabla v \\ + \nabla u_h\cdot\big(R(v) - L(\beta\cdot[\![v]\!])\big)\,dx \\ + \sum_{e\in\mathcal{E}_h}\int_e \alpha^e([\![u_h]\!])\cdot[\![v]\!]\,dx = \int_\Omega f\,v\,dx.\end{aligned} \tag{15}$$

For $\beta = \gamma = 0$ and α chosen as in (6), we recover the old IP method of [16] and [1], while for $\beta = \gamma = 0$ and α as in (7) we recover the second formulation of the original DG method of Bassi and Rebay introduced by Bassi, Rebay, Mariotti, Pedinotti, and Savini [7]. As proven in [11], under rather general

assumptions, and for triangular elements, the scheme is stable and optimally convergent whenever $\eta^e > 3$, where the number 3 represents, in essence, the number of edges per element.

Notice that now the number of non-zero entries of the stiffness matrix is reduced to its minimum. This is due to the fact that the term $R(u_h) \cdot R(v)$ that appeared in (14) is not present anymore in (15).

We now consider another family of numerical fluxes. Let us choose:

$$h_\sigma^e = \zeta\{\nabla u_h\} - \alpha^e([\![u_h]\!]), \qquad h_u^{e,K} = \{u_h\} + \delta[\![u_h]\!] \cdot n_K, \tag{16}$$

where ζ and δ are real parameters. Different choices for these parameters will select different methods. We point out immediately that for $\delta \neq 0$ the corresponding methods will not be completely conservative, and for $\zeta \neq 1$ consistency will be violated.

Using (16) in (2)–(3) and proceeding in the elimination of σ_h as before, we get

$$\begin{aligned} \sum_K \int_K (P_\Sigma(\nabla u_h) \cdot \nabla v + (1-2\delta) R(u_h) \cdot \nabla v + \zeta \nabla u_h \cdot R(v))\, dx \\ + \sum_{e \in \mathcal{E}_h} \int_e \alpha^e([\![u_h]\!]) \cdot [\![v]\!]\, dx = \int_\Omega f\, v\, dx. \end{aligned} \tag{17}$$

For $\delta = 1$, $\zeta = 1$, $\alpha^e = 0$, and $\nabla P(K) \subset \Sigma(K)$ (so restricted to $\nabla P(K)$, P_Σ reduces to the inclusion operator and can be suppressed), this is exactly the DG method of Baumann and Oden. To see this, let us rewrite the above equation. We start by noting that

$$\int_\Omega \nabla u \cdot R(v)\, dx = -\sum_{e \in \mathcal{E}_h} \int_e [\![v]\!]\{\nabla u\}\, ds = -\sum_K \int_{\partial K} (\!| v |\!) \frac{\partial u}{\partial n_K}\, ds,$$

where we set, in each element K, for every $e \in \partial K$,

$$(\!| v |\!) = \frac{1}{2}(v^{int} - v^{ext})_e,$$

with obvious meaning of the symbols. With this notation and when $\nabla P(K) \subset \Sigma(K)$, the equation (17) can be rewritten as

$$\begin{aligned} \sum_K \left(\int_K \nabla u_h \cdot \nabla v\, dx + \int_{\partial K} ((2\delta - 1)(\!| u_h |\!) \frac{\partial v}{\partial n} - \zeta (\!| v |\!) \frac{\partial u_h}{\partial n}) \, ds \right) \\ + \sum_{e \in \mathcal{E}_h} \int_e \alpha^e([\![u_h]\!]) \cdot [\![v]\!]\, ds = \int_\Omega f\, v\, dx, \end{aligned}$$

which is nothing but the DG method of Baumann and Oden when $\delta = \zeta = 1$ and $\alpha^e = 0$, as claimed. This scheme has been analyzed by Oden, Babuška, and Baumann [20], and requires some extra assumptions to achieve stability,

e.g., polynomials of degree ≥ 2. The situation would clearly improve by taking α^e as in (6) or (7) with $\eta^e > 0$. This is also indicated by Süli, Schwab, and Houston in [25] and [26], where a full analysis of these methods (with $\delta = 0$ or 1, $\zeta = 1$ and α^e as in (6), $\eta^e > 0$) is performed.

On the other hand, by taking $\delta = 1/2$ and $\zeta = 0$ in (16), equation (17) becomes

$$\sum_K \int_K P_\Sigma(\nabla u_h) \cdot \nabla v \, dx + \sum_{e \in \mathcal{E}_h} \int_e \alpha^e([\![u_h]\!]) \cdot [\![v]\!] \, dx = \int_\Omega f \, v \, dx.$$

This, when $\nabla P(K) \subset \Sigma(K)$, can be seen as an extension of the Babuška-Zlámal IP method [3] to second order elliptic problems, when α^e is chosen as in (6). If instead, α^e is chosen as in (7), we obtain the penalty formulation proposed in [12]. Note that both methods are inconsistent, so that, in both cases, η^e has to go to $+\infty$ when the meshsize tends to zero, although with different speed for the two cases; for triangular grids, η^e should behave as $|e|^{-2p-1}$ in the former case, and as $|e|^{-2p}$ in the latter, where p is the degree of the polynomials in $P(K)$.

5 Concluding remarks

In this paper, we have proposed a unified framework to study a large class of DG methods for elliptic problems. This class includes the classical IP methods as well as practically all the recently introduced DG methods. The following table summarizes the flux choices needed to obtain the methods discussed; for all these methods $P(K)$ is a standard polynomial space and $\Sigma(K)$ is taken large enough to contain $\nabla P(K)$.

Method	$h_\sigma^{e,K}$	$h_u^{e,K}$
Bassi–Rebay 1	$\{\sigma_h\}$	$\{u_h\}$
Brezzi et al. 1	$\{\sigma_h\} - \eta^e\{r_e([\![u_h]\!])\}$	$\{u_h\}$
LDG	$\{\sigma_h\} - \eta^e[\![u_h]\!] + \beta^e[\![\sigma_h]\!]$	$\{u_h\} + \gamma^e[\![u_h]\!]$
IP	$\{\nabla u_h\} - \eta^e[\![u_h]\!]$	$\{u_h\}$
Bassi–Rebay 2	$\{\nabla u_h\} - \eta^e\{r_e([\![u_h]\!])\}$	$\{u_h\}$
Baumann–Oden	$\{\nabla u_h\}$	$\{u_h\} - [\![u_h]\!] \cdot n_K$
Babuška-Zlámal	$-\eta^e[\![u_h]\!]$	$u_h\|_K$
Brezzi et al. 2	$-\eta^e\{r_e([\![u_h]\!])\}$	$u_h\|_K$

We saw that this class subdivides naturally into completely conservative methods and partially conservative methods, on the one hand, and into methods whose fluxes are independent of σ_h and methods who aren't. We saw that completely conservative methods give rise to symmetric problems when the parameters of their numerical fluxes are suitably defined, and that partially conservative methods might give rise to non-symmetric methods. We also saw that DG methods whose numerical fluxes are independent of σ_h produce stiffness matrices with a remarkably smaller number of non-zero entries.

We believe that such a unified framework could facilitate the understanding of the various methods and their relationships, as well as a possible unified analysis of their convergence properties.

References

1. D.N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19:742–760, 1982.
2. I. Babuška. The finite element method with penalty. *Math. Comp.*, 27:221–228, 1973.
3. I. Babuška and M. Zlámal. Nonconforming elements in the finite element method with penalty. *SIAM J. Numer. Anal.*, 10:863–875, 1973.
4. G.A. Baker. Finite element methods for elliptic equations using nonconforming elements. *Math. Comp.*, 31:45–59, 1977.
5. G.A. Baker, W.N. Jureidini, and O.A. Karakashian. Piecewise solenoidal vector fields and the Stokes problem. *SIAM J. Numer. Anal.*, 27:1466–1485, 1990.
6. F. Bassi and S. Rebay. A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations. *J. Comput. Phys.*, 131:267–279, 1997.
7. F. Bassi, S. Rebay, G. Mariotti, S. Pedinotti, and M. Savini. A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows. In R. Decuypere and G. Dibelius, editors, *2nd European Conference on Turbomachinery Fluid Dynamics and Thermodynamics*, pages 99–108, Antwerpen, Belgium, March 5–7 1997. Technologisch Instituut.
8. C.E. Baumann and J.T. Oden. A discontinuous *hp* finite element method for convection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.* in press, special issue on Spectral, Spectral Element, and *hp* Methods in CFD, edited by G.E. Karniadakis, M. Ainsworth and C. Bernardi.
9. C.E. Baumann and J.T. Oden. A discontinuous *hp* finite element method for the Navier-Stokes equations. In *10th. International Conference on Finite Element in Fluids*, 1998.
10. R. Becker and P. Hansbo. A finite element method for domain decomposition with non-matching grids. Technical Report 3613, INRIA, 1999.
11. F. Brezzi, D. Marini, P. Pietra, and A. Russo. Discontinuous finite elements for diffusion problems. In *Atti Convegno in onore di F. Brioschi, Istituto Lombardo, Accademia di Scienze e Lettere*, 1997. to appear.
12. F. Brezzi, D. Marini, P. Pietra, and A. Russo. Discontinuous finite elements for diffusion problems. *Numerical Methods for Partial Differential Equations*, 1999. submitted.

13. P. Castillo. An optimal error estimate for the local discontinuous Galerkin method. In *this volume*, pages –, 1999.
14. B. Cockburn, G.E. Karniadakis, and C.-W. Shu. The development of discontinuous Galerkin methods. In *this volume*, pages –, 1999.
15. B. Cockburn and C.W. Shu. The local discontinuous Galerkin finite element method for convection-diffusion systems. *SIAM J. Numer. Anal.*, 35:2440–2463, 1998.
16. J. Douglas, Jr. and T. Dupont. *Interior penalty procedures for elliptic and parabolic Galerkin methods*, volume 58 of *Lecture Notes in Physics*. Springer-Verlag, Berlin, 1976.
17. M. Dubiner. Spectral methods on triangles and other domains. *J. Sci. Comp.*, 6:345–390, 1991.
18. I. Lomtev and G.E. Karniadakis. A discontinuous Galerkin method for the Navier-Stokes equations. *Int. J. Numer. Meth. Fluids*, 29:587–603, 1999.
19. J.A. Nitsche. Über ein Variationsprinzip zur Lösung Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unteworfen sind. *Abh. Math. Sem. Univ. Hamburg*, 36:9–15, 1971.
20. J.T. Oden, Ivo Babuška, and C.E. Baumann. A discontinuous *hp* finite element method for diffusion problems. *J. Comput. Phys.*, 146:491–519, 1998.
21. W.H. Reed and T.R. Hill. Triangular mesh methods for the neutron transport equation. Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, 1973.
22. B. Rivière and M.F. Wheeler. A discontinuous Galerkin method applied to nonlinear parabolic equations. In *this volume*, pages –, 1999.
23. B. Rivière and M.F. Wheeler. Part I. Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. Technical Report 99-09, TICAM, 1999.
24. T. Rusten, P.S. Vassilevski, and R. Winther. Interior penalty preconditioners for mixed finite element approximations of elliptic problems. *Math. Comp.*, 65:447–466, 1996.
25. E. Süli, Ch. Schwab, and P. Houston. *hp*-DGFEM for partial differential equations with non-negative characteristic form. In *this volume*, pages –, 1999.
26. E. Süli and Ch. Schwab and P. Houston. *hp*-finite element methods for hyperbolic problems. In J. R. Whiteman, editor, *The Mathematics of Finite Elements and Applications, MAFELAP X*. Springer Verlag, June 1999. to appear.
27. M.F. Wheeler. An elliptic collocation-finite element method with interior penalties. *SIAM J. Numer. Anal.*, 15:152–161, 1978.

Analysis of Finite Element Methods for Linear Hyperbolic Problems

Richard S. Falk*

Department of Mathematics
Rutgers University
Piscataway, NJ 08854
falk@math.rutgers.edu

Abstract. We summarize several techniques of analysis for finite element methods for linear hyperbolic problems, illustrating their key properties on the simplest model problem. These include the discontinuous Galerkin method, the continuous Galerkin methods on rectangles and triangles, and a nonconforming linear finite element on a special triangular mesh.

1 Introduction

Let Ω be a bounded polygonal domain and consider the simple linear hyperbolic problem:

$$\alpha \cdot \nabla u = f \quad \text{in } \Omega, \qquad u = g \quad \text{on } \Gamma_{in}(\Omega),$$

where $\alpha = (\alpha_1, \alpha_2)$ is a constant vector and $\Gamma_{in}(\Omega)$ is the portion of the boundary of Ω on which $\alpha \cdot \boldsymbol{n} < 0$, with $\boldsymbol{n}$ denoting the unit outward normal to $\partial\Omega$.

In this paper, we review several finite element methods proposed for this model problem, and discuss the key ingredients of their analysis. At the most basic level, all of the numerical analysis tries to follow in some way the basic conservation property of the homogeneous equation. That is, multiplying the homogeneous equation by u and integrating over a subdomain G, we have

$$0 = (\boldsymbol{\alpha} \cdot \nabla u, u)_G = \frac{1}{2} \int_G \boldsymbol{\alpha} \cdot \nabla(u^2) = \frac{1}{2} \int_{\partial G} u^2 \, \boldsymbol{\alpha} \cdot \boldsymbol{n}.$$

This may be written in the form

$$\frac{1}{2} \int_{\Gamma_{out}(G)} u^2 |\boldsymbol{\alpha} \cdot \boldsymbol{n}| = \frac{1}{2} \int_{\Gamma_{in}(G)} u^2 |\boldsymbol{\alpha} \cdot \boldsymbol{n}|$$

since $\boldsymbol{\alpha} \cdot \boldsymbol{n} \geq 0$ on $\Gamma_{out}(G)$ and $\boldsymbol{\alpha} \cdot \boldsymbol{n} \leq 0$ on $\Gamma_{in}(G)$.

* This work was supported by NSF grant DMS-9704556.

If we choose Ω so that it is the disjoint union of subdomains G_i and sum these identities, cancellation of integrals over the common boundaries leads to the conservation result:

$$\frac{1}{2}\int_{\Gamma_{out}(\Omega)} u^2|\boldsymbol{\alpha}\cdot\boldsymbol{n}| = \frac{1}{2}\int_{\Gamma_{in}(\Omega)} u^2|\boldsymbol{\alpha}\cdot\boldsymbol{n}|.$$

It is this type of analysis that we wish to follow at the discrete level to obtain stability and an error analysis of finite element approximation schemes.

2 The Discontinuous Galerkin Method

We begin with the method which is the subject of this conference, and whose analysis is the most familiar. Let τ_h denote a triangulation of Ω into triangles T of diameter $\leq h$ and $P_n(T)$ the space of polynomials of degree $\leq n$ on T. For each $T \in \tau_h$, the discontinuous Galerkin method is:

Find $u_h \in P_n(T)$ such that

$$(\boldsymbol{\alpha}\cdot\nabla u_h, v_h)_T - \int_{\Gamma_{in}(T)} J[u_h]\, v_h\, \boldsymbol{\alpha}\cdot\boldsymbol{n} = (f, v_h)_T \tag{1}$$

for all $v_h \in P_n(T)$, where $J[v] = v^+ - v^-$, with $v^{\pm}(x) = \lim_{\epsilon\to 0^{\pm}} v(x+\epsilon\boldsymbol{\alpha})$, $u_h^-(x) = g(x)$ if $x \in \Gamma_{in}(\Omega)$, and $(\cdot,\cdot)_T$ denotes the L^2 inner product over T.

For this method, one can follow the lead of the continuous problem and take the test function $v_h = u_h$. Then for the homogeneous problem $f = 0$,

$$(\boldsymbol{\alpha}\cdot\nabla u_h, u_h)_T - \int_{\Gamma_{in}(T)} J[u_h]\, u_h^+\, \boldsymbol{\alpha}\cdot\boldsymbol{n} = 0.$$

Integrating the first term by parts as before and recombining terms, one gets

$$\begin{aligned}\frac{1}{2}\int_{\Gamma_{out}(T)} (u_h^-)^2|\boldsymbol{\alpha}\cdot\boldsymbol{n}| &= \frac{1}{2}\int_{\Gamma_{in}(T)} (u_h^+)^2|\boldsymbol{\alpha}\cdot\boldsymbol{n}| - \int_{\Gamma_{in}(T)} [u_h^+ - u_h^-]u_h^+|\boldsymbol{\alpha}\cdot\boldsymbol{n}| \\ &= \frac{1}{2}\int_{\Gamma_{in}(T)} (u_h^-)^2|\boldsymbol{\alpha}\cdot\boldsymbol{n}| - \frac{1}{2}\int_{\Gamma_{in}(T)} (J[u_h])^2|\boldsymbol{\alpha}\cdot\boldsymbol{n}|.\end{aligned}$$

Summing over all triangles in the triangulation comprising Ω gives

$$\frac{1}{2}\int_{\Gamma_{out}(\Omega)} (u_h^-)^2|\boldsymbol{\alpha}\cdot\boldsymbol{n}| + \sum_T \frac{1}{2}\int_{\Gamma_{in}(T)} (J[u_h])^2|\boldsymbol{\alpha}\cdot\boldsymbol{n}| = \frac{1}{2}\int_{\Gamma_{in}(\Omega)} g^2|\boldsymbol{\alpha}\cdot\boldsymbol{n}|.$$

This identity is of course the basic one needed to establish stability of the method. The additional test function $\boldsymbol{\alpha}\cdot\nabla u_h$, used by Johnson and Pitkäranta [4] provides additional stability and leads to an improvement in the error estimates originally obtained by Lesaint and Raviart [5].

3 Winther's method

Next consider a method proposed by R. Winther, using a rectangular mesh. The approximate solution is sought in the space of continuous tensor product piecewise polynomials of degree $\leq n$ in each variable. On each rectangle R, the approximate solution $u_h \in \boldsymbol{Q}_n$ is determined by:

$$(\boldsymbol{\alpha} \cdot \nabla u_h, v_h)_R = (f, v_h)_R \qquad \text{for all } v_h \in \boldsymbol{Q}_{n-1},$$

where $\boldsymbol{Q}_n$ denotes the space of tensor products of polynomials of degree n in x and y. These equations must be solved in an order determined by the characteristic direction and it is assumed that u_h is already known on the inflow boundary of the rectangle R. In the simplest case $n = 1$, u_h is known at 3 of the rectangle vertices and the value at the fourth is determined by taking inner products against constants. In this case, the method is the box scheme or a simple finite volume method.

The analysis of such a method is not so obvious, since the test function $v_h = u_h$ is not allowed. Winther's idea is to get conservation of a quantity equivalent to L^2 conservation of u by choosing two test functions. The first is $(u_h)_{xy}$. Considering the homogeneous problem, and dropping the subscript h for the moment, we have for the rectangle R_{ij} with corners (ih, jh), $([i+1]h, jh)$, $(ih, [j+1]h)$, and $([i+1]h, [j+1]h)$,

$$\begin{aligned}(\alpha_1 u_x + \alpha_2 u_y, u_{xy})_{R_{ij}} &= \frac{1}{2}\int_{ih}^{[i+1]h}\int_{jh}^{[j+1]h} (\alpha_1[(u_x)^2]_y + \alpha_2[(u_y)^2]_x)\,dy\,dx \\ &= \frac{1}{2}\int_{ih}^{[i+1]h} \alpha_1\{[u_x(x,[j+1]h)]^2 - [u_x(x,jh)]^2\}dx \\ &\quad + \frac{1}{2}\int_{jh}^{[j+1]h} \alpha_2\{[u_y([i+1]h,y)]^2 - [u_y(ih,y)]^2\}dy \\ &= \frac{1}{2}\int_{\Gamma_{out}(R_{ij})} (u_\tau)^2|\boldsymbol{\alpha}\cdot\boldsymbol{s}| - \frac{1}{2}\int_{\Gamma_{in}(R_{ij})} (u_\tau)^2|\boldsymbol{\alpha}\cdot\boldsymbol{s}|,\end{aligned}$$

where $\boldsymbol{s}$ denotes the unit tangent vector to ∂R_{ij} and $u_\tau = \nabla u \cdot \boldsymbol{s}$ denotes the tangential derivative along ∂R_{ij}.

Since u_h is continuous across rectangle edges, so is $(u_h)_\tau$, so summing over all rectangles leads to cancellations and the following result.

$$\frac{1}{2}\int_{\Gamma_{out}(\Omega)} [(u_h)_\tau]^2|\boldsymbol{\alpha}\cdot\boldsymbol{s}| = \frac{1}{2}\int_{\Gamma_{in}(\Omega)} [(u_h)_\tau]^2|\boldsymbol{\alpha}\cdot\boldsymbol{s}|.$$

Since this is only a seminorm for functions defined on these boundaries, Winther considered a second test function, which is closer to the spirit of the original analysis. Letting P_x and P_y denote L^2 projections into polynomials of degree $n-1$ in x and y, respectively, Winther used the test function

$v_h = P_{x,y} u_h$, where $P_{x,y} \equiv P_x P_y$. Again dropping the subscript h, we have

$$\begin{aligned}
0 &= (\alpha_1 u_x + \alpha_2 u_y, P_{x,y} u) = (\alpha_1 u_x, P_y u) + (\alpha_2 u_y, P_x u) \\
&= (\alpha_1 [P_y u]_x, P_y u) + (\alpha_2 [P_x u]_y, P_x u) \\
&= \frac{1}{2} \int_{ih}^{[i+1]h} \int_{jh}^{[j+1]h} (\alpha_1 ([P_y u]^2)_x + \alpha_2 ([P_x u]^2)_y)\, dy\, dx \\
&= \frac{1}{2} \int_{jh}^{[j+1]h} \alpha_1 \{[P_y u([i+1]h, y)]^2 - [P_y u(ih, y)]^2\} dy \\
&\quad + \frac{1}{2} \int_{ih}^{[i+1]h} \alpha_2 \{[P_x u(x, [j+1]h)]^2 - [P_x u(x, jh)]^2\} dx \\
&= \frac{1}{2} \int_{\Gamma_{out}(R_{ij})} (P_0 u)^2 |\boldsymbol{\alpha} \cdot \boldsymbol{n}| - \frac{1}{2} \int_{\Gamma_{in}(R_{ij})} (P_0 u)^2 |\boldsymbol{\alpha} \cdot \boldsymbol{n}|,
\end{aligned}$$

where now P_0 represents L^2 projection into polynomials of degree $n-1$ along an edge. Adding h^2 times the first identity to the second identity gives conservation of a norm that is equivalent on piecewise polynomials to the L^2 norm.

4 The Continuous Galerkin method on triangles

This method was originally proposed by Reed and Hill [6] and was analyzed in [2]. Here we note that are two types of triangles: type I triangles with one inflow side and type II triangles with two inflow sides. We then seek an approximate solution in the space of continuous piecewise polynomials of degree $\leq n$ determined on each triangle T by the variational equations

$$(\boldsymbol{\alpha} \cdot \nabla u_h, v_h)_T = (f, v_h)_T$$

where $v_h \in P_{n-1}(T)$ on a type I triangle and $v_h \in P_{n-2}(T)$ on a type II triangle. By continuity, u_h will already be known at $n+1$ degrees of freedom on a type I triangle and at $2n+1$ degrees of freedom on a type II triangle. Since the total number of degrees of freedom for polynomials of degree $\leq n$ is equal to $(n+1)(n+2)/2$, and

$$(n+1)(n+2)/2 = n(n+1)/2 + (n+1) = (n-1)n/2 + (2n+1),$$

we have the same number of equations as unknows on both types of triangles. Once again, the analysis of this method is not so obvious, since the simple test function $v_h = u_h$ is not allowed. However, Winther's analysis gives a clue and Richter noticed that the analogue of u_{xy} for triangles is the choice $v = u_{\tau_1 \tau_2}$, with τ_1 and τ_2 chosen to be the variables along the two inflow sides of a type II triangle or two outflow sides of a type I triangle. This leads

to the following useful identity.

$$-\int_T (\boldsymbol{\alpha}\cdot\nabla u)(u_{\tau_1\tau_2}) = \frac{1}{2}\int_{\partial T}\frac{(\boldsymbol{\alpha}\cdot\boldsymbol{n}_1)(\boldsymbol{\alpha}\cdot\boldsymbol{n}_2)}{\boldsymbol{\alpha}\cdot\boldsymbol{n}}u_\tau^2 - \frac{1}{2}\int_{\Gamma_3}\frac{(\tau_1\cdot\boldsymbol{n}_3)(\tau_2\cdot\boldsymbol{n}_3)}{\boldsymbol{\alpha}\cdot\boldsymbol{n}_3}(\boldsymbol{\alpha}\cdot\nabla u)^2,$$

where Γ_3 is the inflow side of a type I triangle and the outflow side of a type II triangle, and $\boldsymbol{n}_i$ is the unit outward normal to Γ_i. Once again, the continuity of u_h ensures the continuity of $(u_h)_\tau$ across triangle edges and when this identity is summed over the triangles, cancellations occur on the triangle boundaries which lead to a stability result. For the homogeneous problem, $\boldsymbol{\alpha}\cdot\nabla u_h = 0$ on a type I triangle, and although this is not true on a type II triangle, this term appears with the right sign so its presence does not interfere with the basic stability result. As in the approach of Winther, another test function must be used to control the full norm on triangle boundaries.

5 Analysis using characteristic coordinates

There is another approach to error analysis of these methods which can be applied to other methods as well. In this approach, described more fully in [3], we use a coordinate system with one coordinate in the characteristic direction and a second coordinate either orthogonal to the characteristic direction or else lying along the inflow side of a type I triangle or the outflow side of a type II triangle.

In the notation of Fig. 1, a triangle T may be described by

$$T = \{(s,t) : s \in [s_{in}(t), s_{out}(t)], t \in [t_0, t_1]\}.$$

Using this coordinate system, one can integrate along the characteristics to write the exact solution in the form

$$u(s,t) = u_{in}(t) + \int_{s_{in}(t)}^{s} f\,ds.$$

We now show how to get a similar formula for the discontinuous Galerkin solution on a type I triangle. Since the function $s - s_{in}(t)$ is linear and vanishes on $\Gamma_{in}(T)$, $v_h = [s - s_{in}(t)]q$ is a polynomial of degree $\leq n$ when q is a polynomial of degree $\leq n-1$. Choosing this test function in (1), we get

$$\big([s - s_{in}(t)](u_h)_s, q\big)_T = \big([s - s_{in}(t)]f, q\big)_T.$$

Since $(u_h)_s \in P_{n-1}(T)$ and $s - s_{in}(t) \geq 0$ in T, this implies $(u_h)_s = R_{n-1}f$, where R_{n-1} denotes the projection of f into $P_{n-1}(T)$ with respect to the

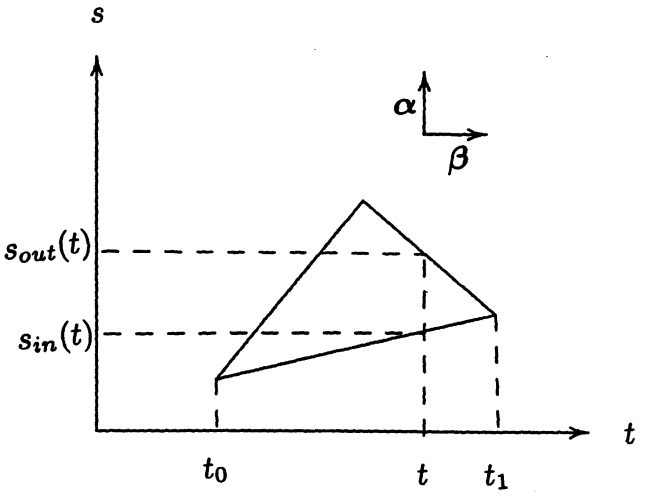


Fig. 1. Characteristic coordinates

weighted L^2 inner product $[f,q] = ([s-s_{in}(t)]f, q)$. Using this result, choosing $v_h = w(t)$ in (1), and rearranging terms, we also get

$$\int_{t_0}^{t_1} (u^+_{h,in} - u^-_{h,in}) w(t)\, dt = (f - R_{n-1}f, w)_T$$
$$= \int_{t_0}^{t_1} \left[\int_{s_{in}(t)}^{s_{out}(t)} (f - R_{n-1}f) \right] w(t)\, dt.$$

Since $u^+_{h,in} - u^-_{h,in} \in P_n[t_0, t_1]$, (polynomials of degree $\leq n$ in t),

$$u^+_{h,in} - u^-_{h,in} = Q_n \int_{s_{in}(t)}^{s_{out}(t)} (f - R_{n-1}f)\, ds,$$

where Q_n denotes L^2 projection into $P_n[t_0, t_1]$. Hence

$$u_h(s,t) = u^-_{h,in} + [u^+_{h,in} - u^-_{h,in}] + \int_{s_{in}(t)}^{s} (u_h)_s\, ds$$
$$= u^-_{h,in} + Q_n \int_{s_{in}(t)}^{s_{out}(t)} (f - R_{n-1}f)\, ds + \int_{s_{in}(t)}^{s} R_{n-1}f\, ds.$$

It follows immediately that

$$u(s_{out}(t),t) - u_h(s_{out}(t),t) = [u_{in} - u_{h,in}^-] + (I - Q_n)\int_{s_{in}(t)}^{s_{out}(t)} (f - R_{n-1}f)\,ds. \tag{2}$$

On a type II triangle, using the test function $v_h = w(t) \in P_n[t_0,t_1]$, we get by integration by parts that

$$\begin{aligned}(f,w)_T &= (\boldsymbol{\alpha}\cdot\nabla u_h, w)_T - \int_{\Gamma_{in}(T)} J[u_h]\, w\,\boldsymbol{\alpha}\cdot\boldsymbol{n} \\ &= \int_{\partial T} u_h\, w\,\boldsymbol{\alpha}\cdot\boldsymbol{n} - \int_{\Gamma_{in}(T)} J[u_h]\, w\,\boldsymbol{\alpha}\cdot\boldsymbol{n} \\ &= \int_{\Gamma_{out}(T)} u_h^-\, w\,\boldsymbol{\alpha}\cdot\boldsymbol{n} + \int_{\Gamma_{in}(T)} u_h^-\, w\,\boldsymbol{\alpha}\cdot\boldsymbol{n} \\ &= \int_{t_0}^{t_1} [u_{h,out}^- - u_{h,in}^-] w(t)\,dt.\end{aligned}$$

When $f = 0$, since $u_{h,out}^-$ and $w(t) \in P_n[t_0,t_1]$, we get that $u_{h,out}^- = Q_n u_{h,in}^-$.

One can then obtain an error analysis for this method by using t-dependent test functions. Setting $e = u - u_h$ and rewriting the discontinuous Galerkin error equation in the form

$$0 = (\boldsymbol{\alpha}\cdot\nabla e, v_h)_T - \int_{\Gamma_{in}(T)} [e^+ - e^-]\boldsymbol{\alpha}\cdot\boldsymbol{n} = -(e, \boldsymbol{\alpha}\cdot\nabla v_h)_T + \int_{t_0}^{t_1} (e_{out}^- - e_{in}^-) v_h\,dt,$$

we get for test functions $v_h = v_h(t)$, that

$$\int_{t_0}^{t_1} (e_{out}^- - e_{in}^-)\, v_h\,dt = 0.$$

Choosing $v_h = Q_n(e_{out}^- + e_{in}^-)$, we get

$$|Q_n e_{out}^-|^2 = |Q_n e_{in}^-|^2.$$

Equivalently,

$$|e_{out}^-|^2 + |(I - Q_n)e_{in}^-|^2 = |e_{in}^-|^2 + |(I - Q_n)e_{out}^-|^2.$$

On a type II triangle, $u_{h,out}^-$ is a polynomial of degree $\le n$ in t, so

$$|(I - Q_n)e_{out}^-| = |(I - Q_n)u_{out}^-| \le Ch^{n+1/2}\|u\|_{n+1,T}.$$

On a type I triangle, we get using (2) and the fact that $u_{h,in}^- \in P_n[t_0,t_1]$ that

$$(I - Q_n)e_{out}^- = (I - Q_n)u_{in} + (I - Q_n)\int_{s_{in}(t)}^{s_{out}(t)} (f - R_{n-1}f)\,ds.$$

Using this formula, it is not difficult to show that

$$|(I - Q_n)e^-_{out}| \leq Ch^{n+1/2}\|u\|_{n+1,T}.$$

We can now sum these identities in the usual way to produce the standard error estimate

$$|e^-_{out}|^2_{\Gamma_{out}(\Omega)} \leq |e^-_{in}|^2_{\Gamma_{out}(\Omega)} + Ch^{2n+1}\|u\|^2_{\Omega,n+1} \leq Ch^{2n+1}\|u\|^2_{\Omega,n+1}.$$

A similar approach can be used to analyze the continuous Galerkin method. The test function $w(t) = Q_{n-1}(e'_{out} + e'_{in})$ produces the basic estimate

$$|Q_{n-1}e'_{out}|^2 = |Q_{n-1}e'_{in}|^2,$$

from which an error estimate can be obtained using a similar technique. The analogue for the continuous Galerkin method of the condition $u^-_{h,out} = Q_n u^-_{h,in}$ holding for the discontinuous Galerkin method on a type II triangle when $f = 0$ is that $[u^-_{h,out}]' = Q_{n-1}[u^-_{h,in}]'$.

6 A nonconforming piecewise linear approximation

A natural question is whether this technique can be used to analyze methods that don't already have another method of error analysis. One class of such methods are those that lie somewhere between the continuous and discontinuous Galerkin methods in the sense that only certain moments are required to be continuous across element edges. Perhaps the simplest example is a method that seeks an approximate solution in the space of nonconforming piecewise linear elements, i.e., piecewise linear functions continuous at the midpoints of triangle edges. On a type II triangle, the solution would already be known at the midpoints of the two inflow edges (i.e., average values are continuous across edges). The remaining degree of freedom could be determined by requiring the finite volume condition that

$$\int_T \boldsymbol{\alpha} \cdot \nabla u_h \, v = \int_T f v$$

for all constant functions v. On a type I triangle, the solution would be known at the midpoint of the inflow edge and several possibilities exist for the remaining two equations. Perhaps the simplest is to require the finite volume condition above and also to require that the first moment be continuous across the inflow edge, i.e., continuity of u_h across the inflow edge of a type I triangle. Again, it is not obvious how to analyze such a method since the test function $v = u_h$ is not allowed. An error analysis, using the t-dependent test functions described above, can be found in [1] in the case when each type II triangle, together with a type I triangle whose inflow side is the outflow side of the type II triangle, forms a parallelogram.

The identity for the homogeneous problem, which is the key to the analysis, is obtained by first integrating by parts to write the homogeneous variational equation in the form

$$\int_{t_0}^{t_1} (u_{h,out}^- - u_{h,in}^+) v_h \, dt,$$

and choose the constant test function

$$v_h = P_{in}(u_{h,out} + u_{h,in})$$

on a type I triangle and

$$v = P_{out}(u_{h,out} + u_{h,in})$$

on a type II triangle, where P_{in} is the L^2 projection into constant functions on $\Gamma_{in}(T)$ on a type I triangle and P_{out} is the L^2 projection into constant functions on $\Gamma_{out}(T)$ on a type II triangle. On a type I triangle, this leads to the identity

$$|P_{in}u_{h,out}|^2 = |P_{in}u_{h,in}|^2$$

Since $P_{in}P_{out} = P_{in}$ on a type I triangle,

$$\begin{aligned} |P_{out}u_{h,out}|^2 &= |P_{in}P_{out}u_{h,out}|^2 + |(I - P_{in})P_{out}u_{h,out}|^2 \\ &= |P_{in}u_{h,in}|^2 + |(I - P_{in})P_{out}u_{h,out}|^2. \end{aligned}$$

On a type I triangle, it is easy to see that $P_{out}P_{in} = P_{in}P_{out}$ and since $\boldsymbol{\alpha} \cdot \nabla u_h = 0$, $u_{h,out}(t) = u_{h,in}(t)$. Hence

$$(I - P_{in})P_{out}u_{h,out} = P_{out}(I - P_{in})u_{h,in} = [du_h/dt]P_{out}Q(t),$$

where $Q(t) = t - t_a$, with t_a being the average value of t on $[t_0, t_1]$. Combining these results, we have on a type I triangle that

$$|P_{out}u_{h,out}|^2 = |P_{in}u_{h,in}|^2 + [du_h/dt]^2|P_{out}Q|^2.$$

In a similar fashion, we can show on a type II triangle that

$$|P_{out}u_{h,out}|^2 = |P_{out}u_{h,in}|^2$$

and then using the fact that $P_{out}P_{in} = P_{out}$, it follows that

$$\begin{aligned} |P_{out}u_{h,out}|^2 &= |P_{out}u_{h,in}|^2 = |P_{out}P_{in}u_{h,in}|^2 \\ &= |P_{in}u_{h,in}|^2 - |(I - P_{out})P_{in}u_{h,in}|^2. \end{aligned}$$

Again, $P_{out}P_{in} = P_{in}P_{out}$, and since $\boldsymbol{\alpha} \cdot \nabla u_h = 0$, $u_{h,in}(t) = u_{h,out}(t)$. Hence

$$(I - P_{out})P_{in}u_{h,in} = P_{in}(I - P_{out})u_{h,out} = [du_h/dt]P_{in}Q.$$

Combining these results, we obtain

$$|P_{out}u_{h,out}|^2 + [du_h/dt]^2|P_{in}Q|^2 = |P_{in}u_{h,in}|^2.$$

In the case when a type I triangle T_1 and a type II triangle T_2 form a parallelogram P, one can show that $|P_{out}Q|^2$ on T_1 is equal to $|P_{in}Q|^2$ on T_2. Since $[du_h/dt]^2$ is continuous across the common boundary of T_1 and T_2, adding the above identities leads to cancellation of the $[du_h/dt]^2$ terms and results in the identity

$$|P_{out}u_{h,out}|^2_{\Gamma_{out}(T_1)} = |P_{in}u_{h,in}|^2_{\Gamma_{in}(T_2)}.$$

Since the above quantities are continuous across triangle edges, we may now sum over all triangles to produce a global stability result

$$|P_{out}u_{h,out}|^2_{\Gamma_{out}(\Omega)} = |P_{in}u_{h,in}|^2_{\Gamma_{in}(\Omega)},$$

which is analogous to the stability result for the continuous problem. An error analysis based on this approach leads to the estimate

$$|P_{out}(u - u_h)|_{\Gamma_{out}(\Omega)} + \|u - u_h\|_\Omega \leq Ch^2\|u\|_{3,\Omega}.$$

Note that the estimate is of optimal order in L^2, but requires additional regularity of the solution. The key idea here is the fact that if the adjacent triangles T_1 and T_2 form a parallelogram and $u \in H^3(T_1 \cup T_2)$, then for all $w \in P_0(T_1 \cup T_2)$,

$$\left| \int_{T_1 \cup T_2} \boldsymbol{\alpha} \cdot \nabla(u - u_I) w \, dx \, dy \right| \leq Ch^2\|u\|_{3,T_1 \cup T_2}\|w\|_{0,T_1 \cup T_2},$$

where u_I is the standard continuous piecewise linear interpolant of u.

References

1. Cai, D-M. : Reduced continuity finite element methods for hyperbolic equations. Ph.D. Dissertation, Rutgers University, (1991)
2. Falk, R.S., Richter, G.R.: Analysis of a continuous finite element method for hyperbolic equations. SIAM J. Numer. Anal. **24** (1987) 257–278
3. Falk, R.S., Richter, G.R.: Local estimates for a finite element method for hyperbolic and convection-diffusion equations. SIAM J. Numer. Anal. **29** (1992) 730–754
4. Johnson, C., Pitkäranta, J.: An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. Math. Comp. **46** (1986) 1–26
5. Lesaint, P., Raviart, P-A.: On a finite element method for solving the neutron transport equation. Mathematical Aspects of Finite Elements in Partial Differential Equations, (C. de Boor, ed.), Academic Press, New York, (1974) 89–123
6. Reed, W. H., Hill, T. R.: Triangular mesh methods for the neutron transport equation. Los Alamos Scientific Laboratory Technical Report LA-UR-73-479 (1973)
7. Winther, R.: A stable finite element method for first-order hyperbolic systems. Math. Comp. **36** (1981) 65–86

Software for the Parallel Adaptive Solution of Conservation Laws by Discontinuous Galerkin Methods

J. E. Flaherty[1], R. M. Loy[2], M. S. Shephard[1], and J. D. Teresco[1]

[1] Scientific Computation Research Center (SCOREC)
and Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180
[2] Mathematics and Computer Science Division
Argonne National Laboratory
Argonne, IL 60439

Abstract. We develop software tools for the solution of conservation laws using parallel adaptive discontinuous Galerkin methods. In particular, the Rensselaer Partition Model (RPM) provides parallel mesh structures within an adaptive framework to solve the Euler equations of compressible flow by a discontinuous Galerkin method (LOCO). Results are presented for a Rayleigh-Taylor flow instability for computations performed on 128 processors of an IBM SP computer. In addition to managing the distributed data and maintaining a load balance, RPM provides information about the parallel environment that can be used to tailor partitions to a specific computational environment.

1 Introduction

By concentrating the computational effort in regions where solution resolution would otherwise be inadequate, adaptive finite element methods (FEMs) provide a reliable, robust, and time- and space-efficient means of solving problems involving partial differential equations [4]. Portions of the finite element mesh may be refined or coarsened (h-refinement), be moved to follow evolving phenomena (r-refinement), or use methods of different order (p-refinement) to enhance resolution and efficiency. In addition to adaptivity, parallel computation is essential for solving large three-dimensional problems in reasonable times. The discontinuous Galerkin (DG) method [5,6] provides an effective means of solving conservation laws on unstructured meshes in a parallel computing environment (§2). The discontinuous basis can capture shock waves and other discontinuities with accuracy, and the compact (nearest neighbor) stencil minimizes interelement communication. This stencil, furthermore, remains compact with high-order polynomial bases, which is (virtually) essential for unstructured mesh computation.

Reusable software libraries allow finite element problems to be solved without concern for the details of the underlying mesh structures, adaptive

procedures, or parallelization. We are developing such libraries to support parallel adaptive finite element computation [10,11,26]. The conventional array-based data representations used for fixed-mesh computation are not well suited for adaptivity by h- or p-refinement [1]. However, alternative structures complicate the automatic (compiler) detection of parallelism. We describe (§3) software to manage distributed mesh data and to provide information about the computational environment by explicit parallelism achieved by message passing using the Message Passing Interface (MPI) [19]. Partitioning and dynamic load balancing algorithms distribute the computation across the processors by a domain decomposition of the spatial (or space-time) mesh.

The DG software is applied to a Rayleigh-Taylor flow instability (§5). This computation represents a preliminary step in the study of thermonuclear flashes on astrophysical bodies, such as neutron stars and white dwarves [18]. Two-dimensional studies [13,14] have shed light on the instability, but the phenomenon is three-dimensional, and such computations are essential for understanding. The problem is, however, quite complex, and the results presented here are only a first step in this direction.

2 The Discontinuous Galerkin Method

We consider three-dimensional conservation laws of the form

$$\mathbf{u}_t(\mathbf{x},t) + \sum_{i=1}^{3} \mathbf{f}_i(\mathbf{x},t,\mathbf{u})_{x_i} = 0, \quad \mathbf{x} \in \Omega, \quad t > 0, \tag{1a}$$

with initial conditions

$$\mathbf{u}(\mathbf{x},0) = \mathbf{u}^0(\mathbf{x}), \quad \mathbf{x} \in \Omega \cup \partial\Omega, \tag{1b}$$

and appropriate well-posed boundary conditions. For the Euler equations, the vector $\mathbf{u}$ specifies the fluid's density, momentum components, and energy. The subscripts t and x_i, $i = 1,2,3$, denote partial differentiation with respect to time and the spatial coordinates. Finite difference schemes for (1), such as the Total Variation Diminishing (TVD) [25,27] and Essentially Non-Oscillatory (ENO) [24] methods, usually achieve high-order accuracy by using a computational stencil that enlarges with order. However, a wide stencil makes the methods difficult to implement on unstructured meshes and limits efficient implementation on parallel computers. Finite element methods have stencils that involve only their neighboring elements regardless of the method order. This allows them to model problems with complex geometries and leads to efficient parallelization. We discretize (1) using a DG finite element method [3,5,6] with a piecewise-continuous spatial basis of polynomials relative to a tetrahedral element Ω_j, $j = 1,2,\ldots,J$, of the mesh on Ω. This basis has a more compact stencil than customary finite element approximations and involves communication only across element faces.

The numerical approximation $\mathbf{U}$ of $\mathbf{u}$ is discontinuous on $\partial\Omega_j$; thus, the flux $\mathbf{f}(\mathbf{u})$ required by the DG method is ambiguous there. It is customarily specified by a "numerical flux" function $\mathbf{h}(\mathbf{U}_j^+, \mathbf{U}_j^-)$ that depends on the solution states $\mathbf{U}_j^+$ and $\mathbf{U}_j^-$ on the interior and exterior, respectively, of $\partial\Omega_j$. Several numerical flux functions are possible [6,24]. Here, we use the method of Colella and Woodward [7,8] to compute an approximate solution to the Riemann problem at $\partial\Omega_j$. This method is based on a Newton's method algorithm of Van Leer [28] but makes the simplifying assumption that wave speeds are the same for both shocks and rarefactions. For efficiency, conditions within rarefactions are computed by linear interpolation to avoid the evaluation of a rational power. Once the ambiguity on $\partial\Omega_j$ has been resolved, the flux may easily be computed. Because of a required iteration, the Colella and Woodward flux is 2–3 times more expensive to evaluate than Van Leer's flux vector splitting [9,17,29], but it offers much greater resolution at contact discontinuities.

Computations with polynomial degrees $p > 0$ require flux or solution limiting to preserve a monotonic behavior near discontinuities. Biswas et al. [3] describe an adaptive solution limiting that avoids "flattening" the solution near smooth extrema and maintains the expected $O(h^{p+1})$, $h = \max_{1 \le j \le J} \operatorname{diam}(\Omega_j)$, L^1 convergence rate when solutions are smooth [3,6]. The results in §6 use a piecewise-constant ($p = 0$) basis with explicit Euler integration in time; hence, limiting is unnecessary.

3 Rensselaer Partition Model

The *Rensselaer Partition Model* (RPM) [26] provides distributed mesh data structures and information about the parallel computational environment in which a program is executing. The basic mesh data structures in RPM are provided by the *SCOREC Mesh Database* (MDB) [1]; however, many of the ideas may be applied to other systems. MDB includes operators to query and update a mesh data structure consisting of a full mesh entity hierarchy: three-dimensional *regions*, and their bounding *faces*, *edges*, and *vertices*, with bidirectional links between mesh entities of consecutive order. Regions serve as finite elements in three dimensions, while faces are finite elements in two dimensions or interface elements in three dimensions. The full entity hierarchy allows efficient mesh modification during h-refinement [22] and facilitates p-refinement [21] by allowing attachment of degrees of freedom to the mesh entities and by providing necessary geometric information. Mesh entities have an explicit *geometric classification* relative to a geometric (CAD) model of the problem domain. This allows the mesh to remain correct with respect to the geometry during h- or p-refinement. Mesh entities are stored with the geometry, so inverse classification information (retrieval of all mesh entities classified on a given model entity) is readily available. This is useful, for example, when applying a boundary condition on a model face. Rather than

visiting all faces in the mesh and querying each to check whether it is on the desired boundary, a list of the needed entities is traversed directly.

Each entity in a distributed finite element mesh is uniquely assigned to a *partition*. Each partition is assigned to a specific *process*, with the possibility that multiple partitions may be assigned to a single process. "Process" in this context refers to an address space. The model is hierarchical, with partitions assigned to a *process model* and processes assigned to a *machine model*.

The machine model represents the computational environment: the processing nodes and their network interconnections. The process model maps processes to the computer and maps interprocess communication to intercomputer networks or, perhaps, to a shared-memory interface. Partitions know the mesh entities that they contain, and mesh entities know their partition assignments (*partition model classifications*).

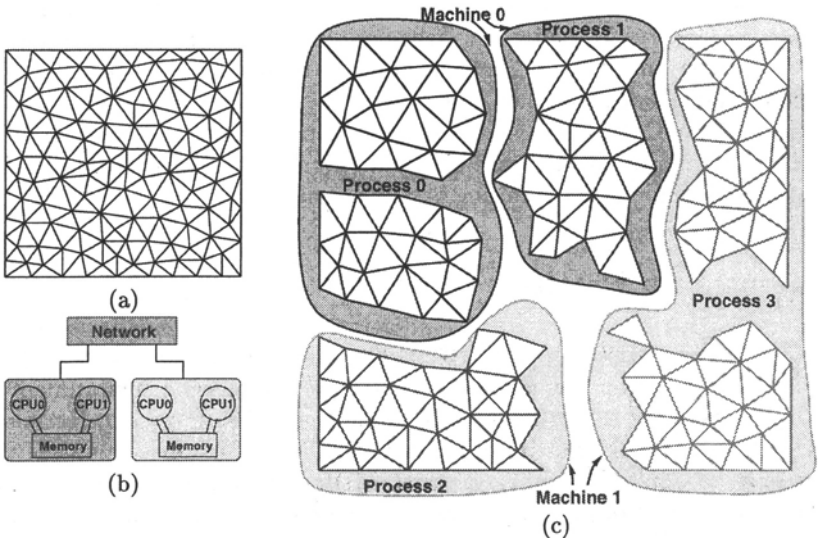


Fig. 1. A sample two-dimensional mesh (a), target parallel environment in which the mesh is to be partitioned (b), and partitioning of the mesh and assignment to processes and machines for the parallel environment (c).

In Figure 1, we see a sample two-dimensional mesh (a) and a target parallel environment consisting of two 2-way SMP workstations connected by a network (b). Figure 1(c) shows a partitioning of this mesh and the assignment of those partitions to the processes and machines of the target environment. Six partitions are created and assigned to four processes, since four processors are available. Two processes are assigned two partitions, while the other two

are assigned only a single partition. The four processes are further assigned to the available machines, two to each.

Mesh entities are replicated only when on a partition boundary that is also a process boundary. Figure 2 shows two-dimensional examples of mesh faces that share a common edge across a partition boundary. The shared mesh edge is classified on the partition boundary in each case. On the left, the partition boundary is local to the process, so the mesh entity need not be replicated and is stored only with the partition boundary mesh. On the right, the partition boundary is also on the process boundary. This partition boundary is replicated in each process, so any mesh entity classified on this boundary must also be replicated.

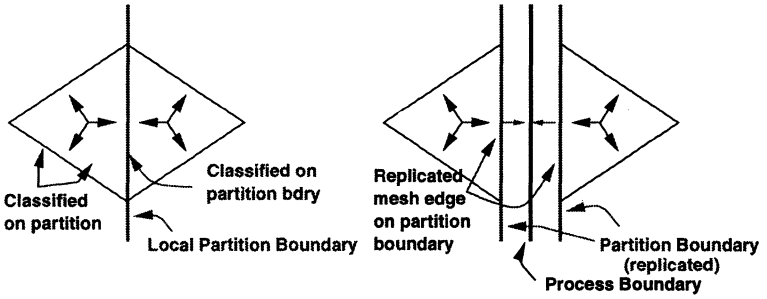


Fig. 2. Partition classification of mesh entities at a same-process partition (left) and at a process-boundary partition (right).

4 Adaptive Methods

Adaptive spatial h-refinement [20,23] is edge based, using error indicators to guide enrichment. An element may be subdivided isotropically or anisotropically, according to predefined templates, depending on the number of its edges selected for refinement. Coarsening is performed when a convex polyhedron of elements request it. A central vertex is identified, the interior edges of the polyhedron are removed, and the polyhedron is remeshed to form fewer elements. Both refinement and coarsening are performed on distributed meshes. During refinement, interprocessor communication is required to update shared vertices, edges, and faces; however, element migration is not necessary. Coarsening requires that the entire polyhedron of elements lie on the same processor, so element migration may be required if the mesh near an interprocessor boundary is marked for coarsening.

With a wide range of element sizes, it is advantageous to use a local refinement method (LRM) [12,16] where spatially dependent time steps are based upon the Courant stability condition. In a given time period, a small

number of large time steps will be taken on large elements, while the opposite will occur on small elements.

The time step for Ω_j is determined from the Courant condition as

$$\Delta t_j = \alpha \frac{r_j}{v_j}, \quad \alpha \leq 1, \tag{2}$$

where r_j is the radius of Ω_j's inscribed sphere and v_j is the maximum signal speed on Ω_j. For the Euler equations, v_j is the sum of the fluid speed and the sound speed. The parameter α is introduced to maintain stability in areas of mesh gradation. We empirically chose $\alpha = 0.65$, but a more thorough analysis is necessary.

Elements may take any stable time step; however, small differences in element sizes and shapes lead to minor differences in time steps. These differences, in turn, lead to time stepping many isolated elements, which causes additional flux evaluations and interpolations. Efficiency can, thus, be improved by rounding time steps down to the next lower (fractional) power of two. This time stepping also helps to organize the computation [15].

Temporal interpolation requires storage for solution data at the previous and current times. Additional space may be required so that the solution may be synchronized and interpolated to a common time for checkpointing or outputting. The interval between synchronization times is referred to as a *major* step. Each major step is composed of several smaller steps, each of which performs a single time step on elements that have the necessary data from their neighbors.

5 Rayleigh-Taylor Flow

The resulting software package implementing the parallel adaptive DG solution of the compressible Euler Equations is called *LOCO*. It has been built using the parallel structures and dynamic load balancing algorithms within RPM.

In collaboration with scientists at the University of Chicago and Argonne National Laboratory, we are working toward complete simulations of thermonuclear flashes on astrophysical bodies such as neutron stars and white dwarves. One crucial aspect of these simulations is the correct modeling of the flame front as it leaves the surface of a compact star in a deflagration stage. Because the relatively dense nuclear fuel lies above the less-dense nuclear ash, the front is subject to Rayleigh-Taylor instabilities. These dramatically alter the shape and area of the burn region and, consequently, the duration and strength of the nuclear flash. Large problems sizes are necessary to accurately model this phenomenon because fine-scale features can dramatically affect the large-scale features.

As a preliminary step, we solve a Rayleigh-Taylor instability problem in a rectangular parallelepiped ($x, y \in [0, 0.25], z \in [0, 1]$) containing an ideal gas

with $\gamma = 5/3$. Initially, the gas in the top half of the domain has density $\rho = 2$ and that in the bottom half has $\rho = 1$; thus, the Atwood Number is $1/3$. The interface between the two regions is sharp. Pressure is unity at the top of the domain and increases toward the bottom with hydrostatic gradient ρg, where $g = 1$ is the acceleration of gravity acting in the z direction. Far-field conditions are applied on the sides, and the pressure is prescribed at the top and bottom to maintain the hydrostatic equilibrium. An initial single-mode sinusoidal velocity perturbation [30] is

$$V_z = -\epsilon_z \cos 8\pi x \cos 8\pi y \sin^\tau \pi z,$$

$$V_x = \epsilon_{xy} \sin 8\pi x \cos 8\pi y \cos \pi z \sin^{\tau-1} \pi z,$$

$$V_y = \epsilon_{xy} \cos 8\pi x \sin 8\pi y \cos \pi z \sin^{\tau-1} \pi z,$$

where

$$\epsilon_z = M_0\sqrt{\gamma/2}, \qquad \epsilon_{xy} = -\epsilon_z \tau/16.$$

The velocity perturbation has magnitude $M_0 = 0.05$ and "tapers off" from the interface with a factor of $\tau = 6$. The planar cross terms V_x and V_y are used for consistency with other software and are of importance with incompressible flows.

6 Computational Results

The Rayleigh-Taylor problem was solved on 128 processors (32 4-way SMP nodes) of an IBM SP computer. Error indicator and refinement tolerances were chosen to detect the interface between the high- and low-density regions, and refined to a given edge length in that region. The initial mesh consisted of 234,421 regions. At $t = 0.28$, the mesh has been adaptively refined to 5,116,334 regions. Octree partitioning with a Morton traversal (OCTPART) and interprocessor boundary smoothing [11] was used to rebalance the computational load after each adaptive enrichment. Details regarding the parallel efficiency of OCTPART and other tools used in this computation are reported elsewhere [12,11].

Figure 3 shows the fluid density at $t = 0.28$ with (left) and without (right) mesh projections on a plane through the center of the domain. The instability is beginning; however, additional computation is necessary to see the complex flow that develops. The adaptive process has clearly concentrated the mesh in the interface zone. The interface is much more sharply defined than previous simulations, which employed a van Leer flux vector splitting rather than the Colella and Woodward fluxes.

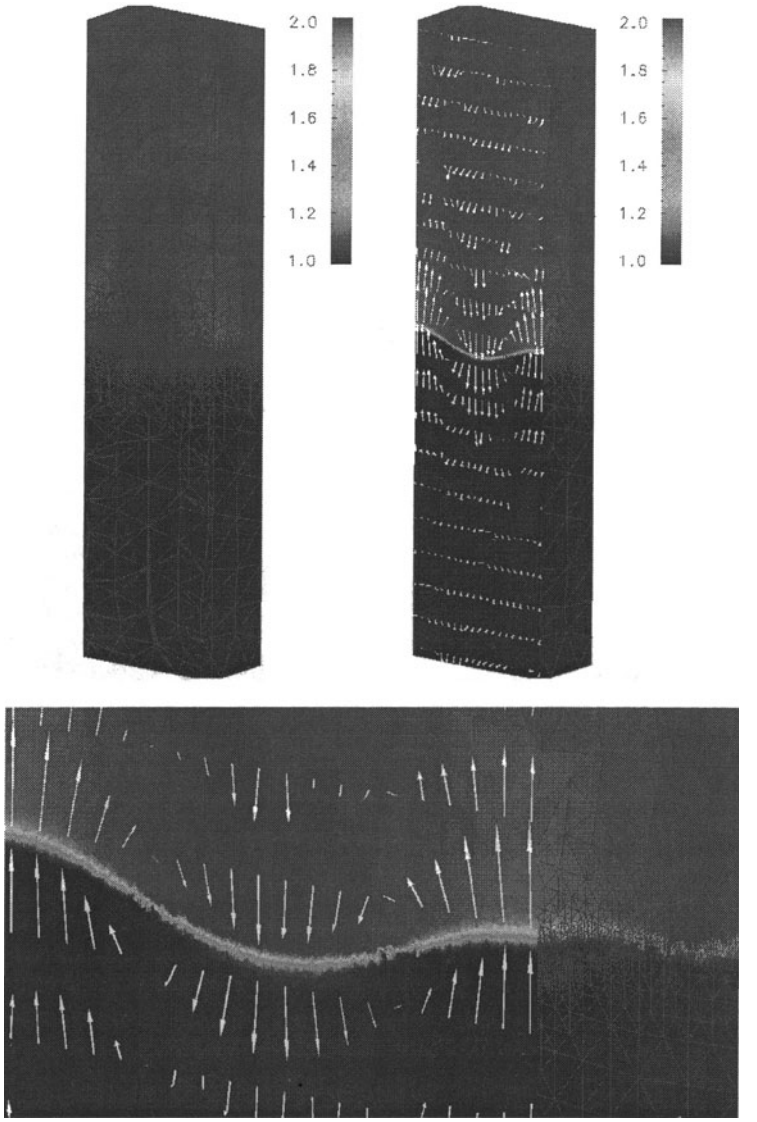


Fig. 3. Densities at $t = 0.28$ for a Rayleigh-Taylor flow projected on a plane through the center of the domain. Densities range from 1 (blue) to 2 (red). The projection on the upper left includes the mesh. Arrows shown on the cut plane at the upper right indicate velocity. The projection at the bottom is a closer view of the interface zone.

7 Discussion

The Rayleigh-Taylor problem is a complex and severe test of an adaptive solution procedure. Additional computations are ongoing. With meshes rang-

ing into millions of regions, we need a hierarchical visualization system to examine the results. Such a system is under development using an octree decomposition of the spatial domain.

A higher-order basis would reduce the spurious diffusion of the piecewise-constant basis used here. This has been developed for two-dimensional flows [3] and is being incorporated into the three-dimensional software. As noted, higher-order requires limiting and the adaptive procedure of Biswas et al. [3] is being extended to unstructured meshes for this purpose. Estimates of discretization errors will be needed both to evaluate accuracy and to guide adaptive enrichment. Possibilities for these include use of superconvergence at Radau points [3] (although extending this idea to unstructured meshes presents a challenge) and the linear-problem estimates of Bey et al. [2]. Adaptive p- and hp-refinement procedures will be possible once these developments have been completed.

RPM is capable of handling the heterogeneities introduced by p-refinement. All of the load balancing procedures include capabilities to weight mesh entities. As noted, weighting due to the LRM was included here. While procedures to handle communications hierarchies are in place (§3), these have to be examined more closely and extended to account for memory hierarchies (cache utilization).

Acknowledgments

We thank Bruce Fryxell and Henry Tufo for many helpful discussions regarding the Rayleigh-Taylor problem and for providing software for the initial perturbation and Riemann flux evaluation.

Flaherty, Shephard, and Teresco were supported by the NSF under Grant ASC-9720227, by the USARO under Contract DAAG55-98-1-0200, by the DOE under Phase 1 ASCI Grant B341495, and by Simmetrix, Inc. Loy was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research Research, U.S. Department of Energy, under Contract W-31-109-Eng-38. The U.S. Government retains for itself, and others acting on its behalf, a paid-up, nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government. Computer systems used include the IBM SP systems at the Maui High Performance Computing Center, the IBM SP at Rensselaer, and the IBM SP Blue-Pacific at Lawrence Livermore National Laboratory.

References

1. M. W. Beall and M. S. Shephard. A general topology-based mesh data structure. *Int. J. Numer. Meth. Engng.*, 40(9):1573–1596, 1997.

2. K. S. Bey, A. Patra, and J. T. Oden. *hp*-version discontinuous Galerkin methods for hyperbolic conservation laws: a parallel adaptive strategy. *Int. J. Numer. Meth. Engng.*, 38(22):3889–3907, 1995.
3. R. Biswas, K. D. Devine, and J. E. Flaherty. Parallel, adaptive finite element methods for conservation laws. *Appl. Numer. Math.*, 14:255–283, 1994.
4. K. Clark, J. E. Flaherty, and M. S. Shephard. *Appl. Numer. Math., special ed. on Adaptive Methods for Partial Differential Equations*, 14, 1994.
5. B. Cockburn, S.-Y. Lin, and C.-W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: One-Dimensional systems. *J. Comput. Phys.*, 84:90–113, 1989.
6. B. Cockburn and C.-W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws II: General framework. *Math. Comp.*, 52:411–435, 1989.
7. P. Colella. Glimm's method for gas dynamics. *SIAM J. Scien. Stat. Comput.*, 3(1):76–110, 1982.
8. P. Colella and P. R. Woodward. The piecewise parabolic (PPM) method for gas-dynamical simulations. *J. Comput. Phys.*, 54:174–201, 1984.
9. K. D. Devine, J. E. Flaherty, R. Loy, and S. Wheat. Parallel partitioning strategies for the adaptive solution of conservation laws. In I. Babuška, J. E. Flaherty, W. D. Henshaw, J. E. Hopcroft, J. E. Oliger, and T. Tezduyar, editors, *Modeling, Mesh Generation, and Adaptive Numerical Methods for Partial Differential Equations*, volume 75, pages 215–242, Berlin-Heidelberg, 1995. Springer-Verlag.
10. J. E. Flaherty, M. Dindar, R. M. Loy, M. S. Shephard, B. K. Szymanski, J. D. Teresco, and L. H. Ziantz. An adaptive and parallel framework for partial differential equations. In D. F. Griffiths, D. J. Higham, and G. A. Watson, editors, *Numerical Analysis 1997 (Proc. 17th Dundee Biennial Conf.)*, number 380 in Pitman Research Notes in Mathematics Series, pages 74–90. Addison Wesley Longman, 1998.
11. J. E. Flaherty, R. M. Loy, C. Özturan, M. S. Shephard, B. K. Szymanski, J. D. Teresco, and L. H. Ziantz. Parallel structures and dynamic load balancing for adaptive finite element computation. *Appl. Numer. Math.*, 26:241–263, 1998.
12. J. E. Flaherty, R. M. Loy, M. S. Shephard, B. K. Szymanski, J. D. Teresco, and L. H. Ziantz. Adaptive local refinement with octree load-balancing for the parallel solution of three-dimensional conservation laws. IMA Preprint Series 1483, Institute for Mathematics and its Applications, University of Minnesota, 1997. To appear, *J. Parallel and Dist. Comput.*
13. B. Fryxell. *Personal communication.*
14. B. Fryxell, E. Müller, and D. Arnett. Hydrodynamics and nuclear burning. *Max-Planck-Institut für Astrophysik Report 449*, 1989.
15. W. L. Kleb and J. T. Batina. Temporal adaptive Euler/Navier-Stokes algorithm involving unstructured dynamic meshes. *AIAA J.*, 30(8):1980–1985, 1992.
16. R. M. Loy. *Adaptive Local Refinement with Octree Load-Balancing for the Parallel Solution of Three-Dimensional Conservation Laws.* PhD thesis, Computer Science Dept., Rensselaer Polytechnic Institute, Troy, 1998.
17. R. A. Ludwig, J. E. Flaherty, F. Guerinoni, P. L. Baehmann, and M. S. Shephard. Adaptive solutions of the Euler equations using finite quadtree and octree grids. *Computers and Structures*, 30:327–336, 1988.
18. A. Malagoli. http://www.flash.uchicago.edu/scientific.htm. URL, 1997.
19. Message Passing Interface Forum, University of Tennessee, Knoxville, Tennessee. *MPI: A Message Passing Interface Standard*, first edition, 1994.

20. L. Oliker, R. Biswas, and R. C. Strawn. Parallel implementaion of an adaptive scheme for 3D unstructured grids on the SP2. In *Proc. 3rd International Workshop on Parallel Algorithms for Irregularly Structured Problems*, Santa Barbara, 1996.
21. M. S. Shephard, S. Dey, and J. E. Flaherty. A straightforward structure to construct shape functions for variable p-order meshes. *Comp. Meth. in Appl. Mech. and Engng.*, 147:209–233, 1997.
22. M. S. Shephard, J. E. Flaherty, C. L. Bottasso, H. L. de Cougny, C. Özturan, and M. L. Simone. Parallel automatic adaptive analysis. *Parallel Comput.*, 23:1327–1347, 1997.
23. M. S. Shephard, J. E. Flaherty, H. L. de Cougny, C. Özturan, C. L. Bottasso, and M. W. Beall. Parallel automated adaptive procedures for unstructured meshes. In *Parallel Comput. in CFD*, number R-807, pages 6.1–6.49. Agard, Neuilly-Sur-Seine, 1995.
24. C.-W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes, II. *J. Comput. Phys.*, 27:1–31, 1978.
25. P. K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM J. Numer. Anal*, 21:995–1011, 1984.
26. J. D. Teresco, M. W. Beall, J. E. Flaherty, and M. S. Shephard. A hierarchical partition model for adaptive finite element computation. To appear, *Comp. Meth. in Appl. Mech. and Engng.*, 1998.
27. B. Van Leer. Towards the ultimate conservative difference scheme. IV. A new approach to numerical convection. *J. Comput. Phys.*, 23:276–299, 1977.
28. B. Van Leer. Towards the ultimate conservative difference scheme. V. A second order sequel to Godunov's methods. *J. Comput. Phys.*, 32:101–136, 1979.
29. B. Van Leer. Flux vector splitting for the Euler equations. ICASE Report 82-30, ICASE, NASA Langley Research Center, Hampton, 1982.
30. Y.-N. Young, H. Tufo, and R. Rosner. On the miscible Rayleigh-Taylor instability: Mixing layer width scaling in 2 and 3-d. In progress, 1999.

Simulation of Gravity Flow of Granular Materials in Silos*

Pierre A. Gremaud[1] and John V. Matthews[1]

Department of Mathematics and Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC 27695-8205, USA

Abstract. The problem of determining the steady state flow of granular materials in silos under the action of gravity is considered. In the case of a Mohr-Coulomb material, the stress equations correspond to a system of hyperbolic conservation laws with source terms and nonlinear boundary conditions. A higher order Discontinuous Galerkin method is proposed and implemented for the numerical resolution of those equations. The efficiency of the approach is illustrated by the computation of the stress fields induced in silos with sharp changes of the wall angle.

1 Introduction

In this paper, the steady state flow of incompressible granular materials under the action of gravity is investigated. Of special interest is the case of flows in silos and bins. Indeed, manufacturing industries routinely store and handle vast quantities of raw materials in granular form. The material is usually retrieved through outlets at the bottom of the containers. Serious difficulties during the withdrawal process are often observed. Those range from dead zones of materials sticking the container's walls to violent vibrations that can cause the complete collapse of the structure. In spite of their commonness, those problems are poorly understood. It is proposed here to analyze numerically the structure and properties of the corresponding flows. This work is, to the authors' knowledge, the first application of a higher order numerical method –a third order Discontinuous Galerkin method– to this type of applications.

The two main physical assumptions, which are discussed in Section 2, are that only established, steady state, flows are considered, and that the material is assumed to be everywhere at yield. Most of the existing work in this field deals with steady state flows in conical hoppers, i.e., when using spherical coordinates, in domains such as

$$\{(r, \theta, \varphi); r > 0, 0 \leq \theta < \theta_w, 0 \leq \varphi \leq 2\pi\},$$

* This research was supported by the Army Research Office (ARO) through grants DAAH04-95-1-0419 and DAAH04-96-1-0097, by the National Science Foundation (NSF) through grant DMS-9818900, and by a grant from the North Carolina Supercomputing Center.

which corresponds to an infinite, converging hopper of half opening angle θ_w. The attention devoted to this case stems from two reasons. First, in a great number of applications, the containers are indeed axisymmetric, if not downright piecewise conical. Second, as a consequence of the invariance of the domain under the scaling transformation $(r, \theta, \varphi) \mapsto (\lambda r, \theta, \varphi)$ where $\lambda > 0$, similarity solutions, the so-called *radial solutions*, can be constructed. This was first observed by Jenike [6], and has played a fundamental role in the design of industrial hoppers ever since [7], [11]. The radial solutions can be found numerically by solving systems of ordinary differential equations, more precisely, boundary value problems. Their behavior is well documented, see e.g. [4], [8]. One should note, however, that the domain of applicability of such an approach is quite limited [5]. Indeed, the very important case of the junction between conical hoppers of different wall angles for instance, see Figure 1 p.7, clearly cannot be treated under the assumption of radial symmetry. Further, even if the considered domains have the necessary symmetry properties, if and when the radial solutions effectively correspond to approximations of what is observed in practice is not clear. The full system of partial differential equations describing the problems at hand has thus to be solved. This is where our contribution lies.

The model is presented in Section 2. Section 3 is devoted to the description of the numerical method. Computational results are discussed and analyzed in Section 4. Finally, some conclusions and remarks on future work are offered in Section 5.

2 The model

The equations governing the *time dependent* flow of granular material under gravity are derived and analyzed in [12]. Those are found to be linearly ill-posed in most cases of practical interest. To the authors' knowledge, the situation is not fully understood, mathematically or otherwise. In practice, strongly time-dependent problems are usually observed in conjunction with *funnel flows*, i.e., flows for which the motion essentially takes place in the central part of the silo. This paper deals exclusively with *mass flows*, i.e., flows for which all the material is mobilized. In this context, established, steady state flows can be observed.

The spatial domain is assumed to be axisymmetric, but not necessarily right conical. The particles are assumed to have no motion in the axial (or φ) direction. Even though this assumption may be counterintuitive to fluid dynamicists, it is an established experimental fact for granular materials. The stress tensor takes the form

$$T = \begin{bmatrix} T_{rr} & T_{r\theta} & 0 \\ T_{r\theta} & T_{\theta\theta} & 0 \\ 0 & 0 & T_{\varphi\varphi} \end{bmatrix}. \tag{1}$$

Neglecting the inertial terms, conservation of momentum yields

$$\nabla \cdot T = \rho g, \tag{2}$$

in which ρ is the density, taken to be constant, and the vector g is the acceleration due to gravity.

Plastic deformation is assumed everywhere. Constitutive models based on plasticity are conveniently expressed in terms of the principal stresses $\sigma_i, i = 1,2,3$, i.e., the eigenvalues of the stress tensor T. If the principal stresses are ordered $\sigma_1 \geq \sigma_2 \geq \sigma_3$, then the *Mohr-Coulomb* yield condition reads

$$\frac{\sigma_1}{\sigma_3} = \frac{1+\sin\delta}{1-\sin\delta},$$

where δ is the angle of internal friction. This relation can be derived from the law of sliding friction [7], Chap. 3. This condition can be expressed in the original stress variables

$$(T_{rr} - T_{\theta\theta})^2 + 4T_{r\theta}^2 = \sin^2\delta\,(T_{rr} + T_{\theta\theta})^2. \tag{3}$$

The *Haar-von Karman* assumption can be invoked to evaluate the circumferential stress $T_{\varphi\varphi}$. Indeed, the Mohr-Coulomb analysis merely states $\sigma_1 \geq T_{\varphi\varphi} \geq \sigma_3$. For axisymmetric converging hoppers, the Haar-von Karman assumption states that $T_{\varphi\varphi}$ is in fact the major principal stress. One can write the corresponding equations in terms of two unknowns T_{rr} and $T_{r\theta}$

$$\begin{aligned} \partial_\tau T_{rr} - \partial_\theta T_{r\theta} &= f(\tau,\theta,T_{rr},T_{r\theta}) \\ \partial_\tau T_{r\theta} - \partial_\theta T_{\theta\theta} &= g(\tau,\theta,T_{rr},T_{r\theta}) \end{aligned} \tag{4}$$

where some simplifications result from the use of the new variable $\tau = -\ln r$; we do not bother to rename the stress variables. The right hand side terms are given by

$$\begin{aligned} f(\tau,\theta,T_{rr},T_{r\theta}) &= \tfrac{3-s}{2}\,T_{rr} + \cot\theta\,T_{r\theta} - \tfrac{3+s}{2}\,T_{\theta\theta} + \rho g\,e^{-\tau}\cos\theta \\ g(\tau,\theta,T_{rr},T_{r\theta}) &= -\tfrac{1+s}{2}\cot\theta\,T_{rr} + 3\,T_{r\theta} + \tfrac{1-s}{2}\cot\theta\,T_{\theta\theta} - \rho g\,e^{-\tau}\sin\theta \end{aligned} \tag{5}$$

where we have set $s = \sin\delta$, and, for future reference, $c = \sqrt{1-s^2} = \cos\delta$. The equation of state, which relates $T_{\theta\theta}$ to the unknowns T_{rr} and $T_{r\theta}$ is the yield condition (3). It should be noticed that (3) is the equation of a cone in the space $(T_{rr}, T_{r\theta}, T_{\theta\theta})$, whose central line bisects the $(T_{rr}, T_{\theta\theta})$-plane. The corresponding relation between $T_{\theta\theta}$ and the unknowns is therefore not a proper functional relation, but rather assigns the dependent variables $(T_{rr}, T_{r\theta})$ to lie on a manifold: the yield surface. The situation greatly simplifies, however, in the case of converging hoppers which we consider in this paper. Indeed, because the large lateral compression taking place, it can then

be argued that $T_{\theta\theta}$ lies on the "top" of the yield surface. Accordingly, by solving (3) for $T_{\theta\theta}$, the equation of state completing (4, 5) is

$$T_{\theta\theta} = h(T_{rr}, T_{r\theta}) \equiv \frac{1+s^2}{1-s^2} T_{rr} + 2\sqrt{\frac{s^2}{(1-s^2)^2} T_{rr}^2 - \frac{1}{1-s^2} T_{r\theta}^2}. \qquad (6)$$

This corresponds to the so-called passive state [7], [9], as opposed to the active state which is the other solution from (3), obtained by changing the sign in front of the radical in (6); see §5.

Rewriting (4) as

$$\partial_r U + \partial_\theta F(U) = G(r, \theta, U), \qquad (7)$$

with the obvious notation, one can then analyze the eigenvalues $\lambda_{1,2}$ of the Jacobian F'. A few calculations lead to

$$\lambda_{1,2} = \mp \tan\delta \mp \frac{1}{c}\sqrt{\frac{s T_{rr} \mp c T_{r\theta}}{s T_{rr} \pm c T_{r\theta}}}. \qquad (8)$$

A quick analysis [3] reveals the eigenvalues to be real and distinct, provided that one stays "in the cone", i.e., $|\frac{T_{r\theta}}{T_{rr}}| < \tan\delta$. In other words, the steady state stress equations (4, 5, 6) form a strictly hyperbolic system of nonlinear conservation laws with source terms. The radial and angular variables r and θ can be thought of as time-like and space-like variables, respectively.

The system has to be completed with "initial" and boundary conditions. The "initial" condition used here corresponds to prescribing the stress high up in the hopper, say, on an "r = constant" surface, and solve down from there. Several arguments can be considered to justify the fact that the stress information travels downward, see [3], [7], or [11] for more details. In the calculations presented in this paper, the "initial" condition is computed from the radial stress field; see §4 for details. Finally, the boundary conditions are given by the law of sliding friction. At any point on the wall, the magnitude of the tangential stress $|T_T|$ is proportional to the magnitude of the normal stress $|T_N|$, i.e.,

$$|T_T| = \mu |T_N|,$$

where $\mu > 0$ is the coefficient of wall friction. In a purely radial geometry, the above boundary condition becomes

$$T_{r\theta} = \pm \mu T_{\theta\theta} \qquad \text{on the walls} \qquad (9)$$

with a "+" sign on one side of the hopper and a "-" sign on the other. Observe that those conditions are nonlinear in the unknowns $T_{rr}, T_{r\theta}$.

At this point, it is worth mentioning that the equations can in fact be somewhat "simplified" by the use of the so-called Sokolovskii variables. Those correspond to the natural parameterization of the conical yield surface. This

approach was for instance taken in [8] and [9]. However, the use of this nonlinear change of variables has the unfortunate side effect of destroying the conservation form of the equations, losing in this way the ability to compute shocks in any reliable way. This prevents, for instance, the computation of stresses occurring at the junction between conical hoppers of different wall angles, one of the main goals of this work. Further, as is well known, many purely numerical problems also appear when solving systems in nonconservation form.

3 The algorithm

For the sake of simplicity, we only describe the algorithm in the case of a conical hopper. The method used is a formally third order Discontinuous Galerkin scheme, see [2] and the references listed therein.

Let $\tau_0 = -\ln r_0$, where r_0 is the value of the radial variable we start from. Let $\theta_w > 0$ be the half opening angle, and let $\Delta\theta = \theta_w/N$ be the mesh size, N being the number of cells. In this axisymmetric setting, the problem is spatially essentially one-dimensional, and thus no efforts have been made to adapt the mesh. We define

$$V = \{v \in L^\infty(0,\theta_w)^2 : v|_{K_j} \in P^k(K_j)^2, j = 1,\dots,N\},$$

where $K_j = [\theta_{j-1}, \theta_j]$ is the j-th cell, with $\theta_j = j\,\Delta\theta$, and $P^k(K_j)$ stands for the space of the polynomials of degree at most k in K_j. We use $k = 2$ or 3 below; see §4. A semidiscretization consists of looking for $U_h(\tau,\cdot) \in V$, $\tau > \tau_0$, such that $U_h(\tau_0,\cdot) = \Pi_V(U(\tau_0,\cdot))$, where $U(\tau_0,\cdot)$ is an "initial condition", see §4, Π_V is a projection operator into V, and

$$\begin{aligned}\frac{d}{d\tau}\int_{K_j} U_h(\tau,\theta)\,v(\theta)\,d\theta + \Delta_+\left(v(\theta_{j-1/2})\,H_{j-1/2}\right)\\ - \sum_{l=1}^{5}\omega_l F(U_h(\tau,\theta_{j,l}))\,\frac{d}{d\theta}v(\theta_{j,l})\Delta\theta\\ = \sum_{l=1}^{5}\omega_l G(\tau,\theta_{j,l},U_h(\tau,\theta))\,v(\theta_{j,l})\Delta\theta, \qquad \forall v \in V, j = 1,\dots,N.\end{aligned}$$

In the previous expression, Δ_+ stands for the usual difference operator, $\Delta_+ U_j = U_{j+1} - U_j$ and the coefficients ω_l and the nodes $\theta_{j,l}$, $l = 1,\dots,5$, $j = 1,\dots,N$ stem from the use the classical 5-point Gaussian quadrature formula. We use the local Lax-Friedrichs flux

$$H_{j+1/2} = \frac{1}{2}\left(F(U^-_{j+1/2}) + F(U^+_{j+1/2}) - \alpha_{j+1/2}(U^+_{j+1/2} - U^-_{j+1/2})\right),$$

where $\alpha_{j+1/2}$ is the magnitude of the largest eigenvalue of a properly chosen Roe average matrix $A_{j+1/2} \approx (\frac{dF}{dU})_{U=U_{j+1/2}}$, see [3] for details and [10] for

background. The mass matrix can be made diagonal by choosing the basis functions as Legendre polynomials over each cell [1].

The coefficients of $U_h(\tau,\cdot)$ can then be grouped in a vector $\mathcal{U}(\tau)$. The unknown vector $\mathcal{U}(\tau)$ satisfies the following system of ODEs

$$\frac{d}{d\tau}\mathcal{U} = \mathcal{F}(\mathcal{U}) + \mathcal{G}(\tau,\mathcal{U}) \tag{10}$$

where $\mathcal{F}(\mathcal{U})$ and $\mathcal{G}(\tau,\mathcal{U})$ come respectively from the discretization of $F(U)$ and $G(\tau,\theta,U)$ in (7). Note that we use below an unsplit approach. This is justified first by the fact that the source term $\mathcal{G}(\tau,\mathcal{U})$ is not stiff and second, by the realization that the delicate interplay between $\mathcal{G}(\tau,\mathcal{U})$ and the nonlinear boundary conditions would render the implementation of a split algorithm a lot more involved than the present approach.

The discretization with respect to τ involves a third order TVD Runge-Kutta [13] combined with a local slope limiting process. Let $\Delta\tau > 0$ be a constant increment in τ and let $\tau^n = \tau_0 + n\Delta\tau$; the algorithm reads then as follows, see e.g. [2]

- Set $U^{(0)} = \Lambda\Pi(U_h(\tau_0,\cdot))$;
- For $n = 0,\ldots,N-1$, compute U_h^{n+1}:
  1. set $U^{(0)} = U_h^n$;
  2. for $i = 1,\ldots,3$, compute the intermediate stages

$$U^{(i)} = \Lambda\Pi\left(\sum_{j=0}^{i-1}\alpha_{ij}U^{(j)} + \Delta\tau\,\beta_{ij}(\mathcal{F}(U^{(j)}) + \mathcal{G}(\tau^n + d_j\Delta\tau, U^{(j)}))\right);$$

  3. set $U_h^{n+1} = U^{(3)}$.

The numerical parameters $\{\alpha_{ij}\}$, $\{\beta_{ij}\}$ and $\{d_j\}$, $i = 1,2,3$, $j = 0,1,2$ are respectively defined as

$$\begin{array}{lll} 1 & 1 & 0 \\ 3/4\ 1/4 & 0\ 1/4 & 1 \\ 1/3\ \ 0\ \ 2/3 & 0\ \ 0\ \ 2/3 & 1/2 \end{array}$$

An experimental stability condition of the type $|\lambda_{max}|\frac{\Delta\tau}{\Delta\theta} \leq \frac{1}{2k+1}$ was used, where λ_{max} is the largest eigenvalue in modulus of F', see (8). No rigorous stability analysis results seem to be available. A thorough description of the local slope limiting operator $\Lambda\Pi$, which is based on the use of a corrected minmod function, can be found, e.g., in [2]; it is not repeated here. Note that both $\Lambda\Pi$ and the proper implementation of the boundary conditions (9) require transformation to the characteristic fields, see [3] for implementation details pertaining to the present application.

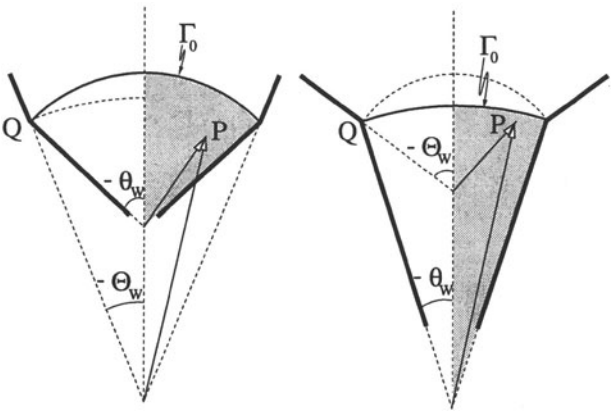


Fig. 1. Geometrical situation; left: transition to a flatter wall angle, right: transition to a steeper wall angle. The radial stress is used to generate an initial condition on the curve Γ_0. The domains of calculation are shaded.

4 Computational results

We analyze the influence of abrupt changes in the wall angle on the stress field. The geometrical situation is illustrated in Figure 1.

Any point P admits two representations, namely, (R, Θ) and (r, θ), corresponding to the natural coordinate systems for the upper and lower hopper respectively. The transition is located through the point Q, see again Figure 1, where $Q = (R_0, -\Theta_w) = (r_0, -\theta_w)$. For given values of the material parameters δ and μ, the numerical approach consists then in

- generating the radial stress field T in the upper hopper $\{(R, \Theta) : R > 0, |\Theta| \leq \Theta_w\}$ [4], [7]; by construction, at a point (R, Θ), the radial stress field is given by $RT(\Theta)$;
- interpolating the radial stress field on the curve $\Gamma_0 = \{(r, \theta) : r = r_0, |\theta| \leq \theta_w\}$, leading to a stress tensor S_0;
- changing to the new coordinate system through $\mathcal{R}^T(\Theta - \theta) S_0 \mathcal{R}(\Theta - \theta)$, where $\mathcal{R}(\alpha)$ is the rotation matrix of angle α;
- solving in the lower hopper $\{(r, \theta) : 0 < r < r_0, 0 < \theta < \theta_w\}$ using the algorithm described in §3.

Notice that by (anti)symmetry, one can solve in one half of the domain only, the law of sliding friction (9) valid on the walls being replaced by the symmetry condition $T_{r\theta} = 0$ on the central line.

Some comments are in order. First, the fact that the initial condition is generated from the radial stress field implicitly assumes that this very solution is sought and realized by the problem in the upper part of the hopper. Second, it also takes as granted that the radial solution reaches the curve Γ_0 unperturbed by the wall corner. This last point is clearly satisfied, assuming

again a downward propagation of the information for the stresses, in case of a transition to a flatter hopper, i.e., $\Theta_w < \theta_w$, see Figure 1, left. If the transition is to a steeper hopper, $\theta_w < \Theta_w$, see Figure 1, right, a quick analysis based on the characteristic curves reveals that the difference in angles should not be too large, namely

$$\Theta_w - \theta_w < \arctan |\frac{1}{\lambda_{max}}|, \qquad \theta_w, \Theta_w > 0,$$

where λ_{max} is again the largest eigenvalue in modulus of F', see (8), evaluated for the radial stress field on the curve $\{(R_0, \Theta) : |\Theta| \leq \Theta_w\}$. Third, the case $\theta_w = \Theta_w$ can be used to check that the algorithm effectively preserves the radial solution. The interaction between the boundary conditions and the forcing terms renders this numerically delicate. Further, the radial solution itself may be unstable [4]. The present approach preserves the radial solution with a great degree of accuracy, see [3] for more details. Finally, it should be noted that the above initial condition is not consistent with the boundary condition (9), unless $\Theta_w = \theta_w$.

In the calculations below, the opening angles Θ_w and θ_w are respectively taken as 15^o and 10^o in a first experiment, and as 10^o and 12^o, in a second. The material parameters correspond to the case of corn in a steel hopper. The angle of internal friction δ is 32.1^o, while the angle of wall friction is 11.7^o, in other words, the wall friction is $\mu = \tan 11.7^o$.

Figure 2 corresponds to the case of a transition to a steeper hopper, whereas in Figure 3, the transition is to a shallower hopper. Some predictions about such transitions can be found in the literature, see e.g. [7], §7.12. Considerations based on analogy with corresponding Fluid Dynamics problems and/or on the use of radial solutions have been advanced, predicting smooth "rarefaction" waves for transitions to steeper hoppers, and shocks in the case of transitions to wider ones. We note here that neither of those types of arguments can be fully justified. The present results shed some light on this problem.

First, in the case of a transition to a steeper hopper, the results reported in Figure 2 clearly show that there is indeed formation of a rarefaction wave. One can observe, however, that after the waves generated on opposite sides start to interact, they sharpen considerably and shocks appear. In Figure 2, both the P^1 and P^2 cases are reported. Although the P^2 results offer a slightly better resolution, they also suffer from small oscillations that can be observed around the center line. The small oscillations stem from numerical difficulties linked to the form of the forcing terms there, see the $\cot\theta$ term in (5), and the delicate interplay with the limiting procedure. Such oscillations were not observed for the P^1 case.

In the case of a transition to a shallower domain, the results of Figure 3 show the immediate formation of shocks in the stress field. In this case, the oscillations of the P^2 calculations are more pronounced; we only display the P^1 results. A more in depth analysis of those issues is presented [3].

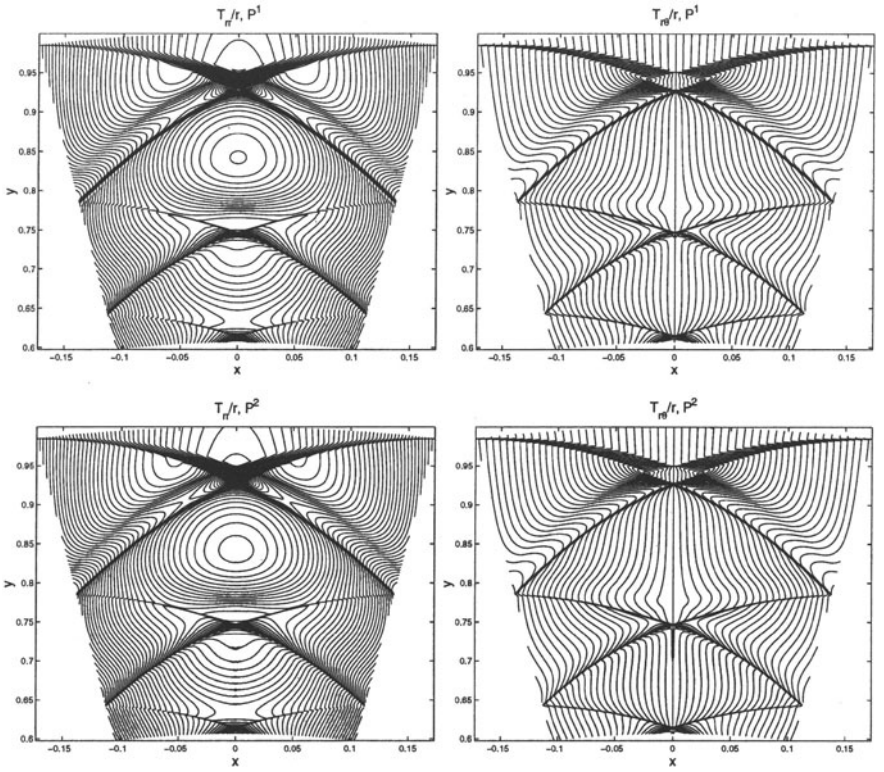


Fig. 2. Structure of the stress field induced by a transition from an opening angle of 15^o to one of 10^o. Values of the parameters: $\delta = 32.1^o$, $\mu = \tan(11.7^o)$ (corn, steel wall). First row, P^1 elements; second row, P^2 elements.

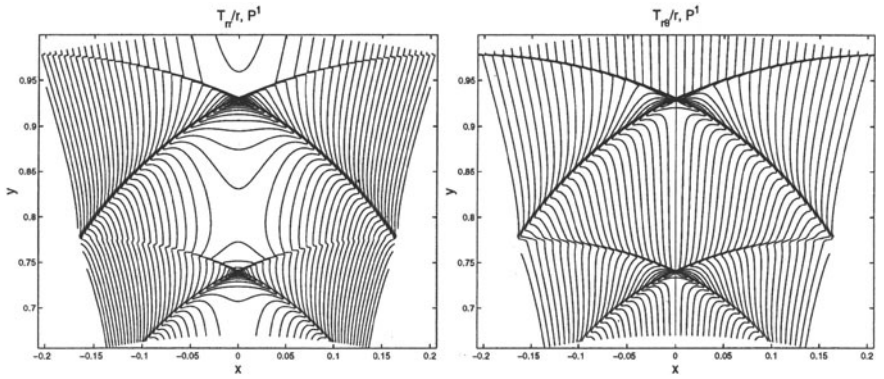


Fig. 3. Structure of the stress field induced by a transition from an opening angle of 10^o to one of 12^o. Values of the parameters: $\delta = 32.1^o$, $\mu = \tan(11.7^o)$ (corn, steel wall). P^1 approximations

5 Concluding remarks

We have presented a numerical study of stress fields induced by the discharge of granular materials in hoppers. The stress equations correspond to a system of hyperbolic conservation laws with several nonstandard features. A higher order Discontinuous Galerkin method has been implemented. The corresponding numerical results partially confirm several "educated guesses" that had been made about the stress field structure induced by changes in the wall angle. A more complete picture of the flow should involve the resolution of the velocity equations, which will be covered in future publications.

Acknowledgments

The authors would like to thank David Schaeffer and Michael Shearer for many helpful discussions.

References

1. Cockburn, B., Lin, S.Y., Shu, C.W.: TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: one dimensional systems. J. Comput. Phys. **84** (1989) 90–113
2. Cockburn, B., Shu, C.W.: The Runge-Kutta discontinuous Galerkin method for conservation laws V: multidimensional systems. J. Comput. Phys. **141** (1998) 199–224
3. Gremaud, P.A., Matthews, J.V.: On the computation of hopper flows. In preparation.
4. Gremaud, P.A., Matthews, J.V., Shearer, M.: Similarity solutions for granular flows in hoppers. Proceedings of the SIAM/AMS Conference on Nonlinear PDEs, Dynamics and Continuum Physics, J. Bona, K. Saxton, R. Saxton, Eds., (1998), AMS Contemporary Mathematics Series, to be published.
5. Gremaud, P.A., Schaeffer, D., Shearer, M.: Numerical determination of flow corrective inserts for granular materials in conical hoppers. To be published in Int. J. Nonlinear Mech.
6. Jenike, A.: Gravity flows of bulk solids. Bulletin No. 108, vol. 52, Utah Eng. Expt. Station, University of Utah, Salt Lake City (1961)
7. Nedderman, R.M.: Static and kinematic of granular materials. Cambridge University Press (1992).
8. Pitman, E.B.: The stability of granular flow in converging hoppers. SIAM J. Appl. Math. **48** (1988) 1033–1052
9. Ravi Prakash, J., Kesava Rao, K.: Steady compressible flow of cohesionless granular materials through a wedge-shaped bunker. J. Fluid Mech. **225** (1991) 21–80
10. Roe, P.L.: Approximate Riemann solvers, parameter vectors, and difference schemes. J. Comp. Phys. **43** (1981) 357–372
11. Royal, A.T.: Private communication. Jenike & Johanson, Inc. (1998).
12. Schaeffer, D.G.: Instability in the evolution equations describing incompressible granular flow. J. Diff. Eq. **66** (1987) 19–50.
13. Shu, C.W., Osher, S.: Efficient implementation of essentially non-oscillatory shock-capturing schemes. J. Comp. Phys. **77** (1988) 439–471.

A Comparison of Discontinuous and Continuous Galerkin Methods Based on Error Estimates, Conservation, Robustness and Efficiency

Thomas J. R. Hughes, Gerald Engel, Luca Mazzei, and Mats G. Larson

Stanford University, Division of Mechanics and Computation,
Durand Building, Stanford, California 94305-4040, USA

Abstract. The Discontinuous Galerkin Method (DGM) and Continuous Galerkin Method (CGM) are investigated and compared for the advection problem and the diffusion problem. First, error estimates for Stabilized Discontinuous Galerkin Methods (SDGMs) are presented. Then, conservation laws are discussed for the DGM and CGM. An advantage ascribed to the DGM is the local flux conservation property. It is remarked that the CGM is not only globally conservative, but locally conservative too when a simple post-processing procedure is used. Next, the robustness of different DGMs is investigated numerically. Lastly, the efficiency of the DGM and CGM is compared.

1 Introduction

The DGM has established itself as an important alternative for solving advection problems, for which the CGM lacks robustness. A feature of the DGM considered important by the CFD community is that it conserves flux locally. These facts have recently led to increased efforts to design a DGM for diffusion problems with the ultimate goal to solve advective-diffusive problems, such as the Navier-Stokes Equations, taking advantage of its good performance in the advection-dominated limit (see, e.g., Oden et al. [7]). In this paper we compare the CGM and DGM in the framework of stabilized methods.

First we state the advection and diffusion model problems considered. Then, a parameterized SDGM finite element formulation is introduced. We continue by stating, without proof, stability and convergence results for the methods proposed. Next, we discuss conservation of DGMs and CGMs. While the conservation properties of DGMs are transparent, those of CGMs require elucidation. We describe new local and global conservation laws for CGMs which have recently been derived elsewhere [4]. Then we consider the robustness of the DGM. Our analysis is concluded by a comparison of the efficiency between the CGM and DGM. The main results are summarized in the conclusions.

2 Model Problems

2.1 Advection

Our first model problem is the hyperbolic advection equation

$$\boldsymbol{u} \cdot \boldsymbol{\nabla} \phi = \mathrm{f} \qquad \text{in } \Omega \ , \tag{1}$$

$$\phi = \mathrm{g} \qquad \text{on } \Gamma^- \ , \tag{2}$$

where $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$ is the velocity field, assumed smooth and solenoidal, i.e. $\boldsymbol{\nabla} \cdot \boldsymbol{u} = 0$; Ω is a bounded domain in $\mathbb{R}^d$, d =1, 2 or 3, with Lipschitz boundary Γ; and $\Gamma^- = \{\boldsymbol{x} \in \Gamma : \boldsymbol{u} \cdot \boldsymbol{n} < 0\}$, $\boldsymbol{n} = \boldsymbol{n}(\boldsymbol{x})$ being the outward normal to Γ, denotes the inflow boundary. The outflow boundary is denoted by $\Gamma^+ = \Gamma \backslash \Gamma^-$. We also define the advective flux, $\boldsymbol{\sigma}^{\mathrm{a}}(\phi) = -\boldsymbol{u}\phi$, its component along $\boldsymbol{n}$, $\sigma_n^{\mathrm{a}}(\phi) = -(\boldsymbol{u}\phi) \cdot \boldsymbol{n} = -(\boldsymbol{u} \cdot \boldsymbol{n})\phi = -u_n \phi$, and an abstract notation for the advective operator, $\mathcal{L}_{\mathrm{a}}(\,\cdot\,) = \boldsymbol{u} \cdot \boldsymbol{\nabla}(\,\cdot\,)$.

2.2 Diffusion

The second model problem we consider is the elliptic diffusion equation with Dirichlet and Neumann boundary conditions:

$$-\boldsymbol{\nabla} \cdot \boldsymbol{A} \boldsymbol{\nabla} \phi = \mathrm{f} \qquad \text{in } \Omega \ , \tag{3}$$

$$\phi = \mathrm{g} \qquad \text{on } \Gamma_{\mathrm{g}} \ , \tag{4}$$

$$\sigma_n^{\mathrm{d}}(\phi) = \mathrm{h} \qquad \text{on } \Gamma_{\mathrm{h}} \ . \tag{5}$$

Γ_{g} and Γ_{h} denote, respectively, the Dirichlet and Neumann parts of the boundary Γ and are such that $\Gamma_{\mathrm{g}} \cap \Gamma_{\mathrm{h}} = \emptyset$ and $\overline{\Gamma_{\mathrm{g}} \cup \Gamma_{\mathrm{h}}} = \Gamma$. Moreover, we assume that the diffusivity matrix $\boldsymbol{A}$ in (3) is symmetric, bounded and such that there exists a constant diffusivity $A > 0$ satisfying $A\,\mathbf{v}^T\mathbf{v} \leq \mathbf{v}^T \boldsymbol{A}(\mathbf{x})\,\mathbf{v}$ for all $\mathbf{v} \in \mathbb{R}^d$. In analogy to the advection case we also define the diffusive flux, $\boldsymbol{\sigma}^{\mathrm{d}}(\phi) = \boldsymbol{A}\boldsymbol{\nabla}\phi$, and its component along $\boldsymbol{n}$, $\sigma_n^{\mathrm{d}}(\phi) = (\boldsymbol{A}\boldsymbol{\nabla}\phi) \cdot \boldsymbol{n}$.

3 Discontinuous Finite Element Formulation

Consider a partition $\mathcal{P}(\Omega) = \{\Omega_e\}$, $e = 1, 2, ..., N_{\mathrm{el}}$, where N_{el} is the number of elements. We assume that the elements are shape-regular (see [2], Thm. 4.4.4) and we use $h_e = \mathrm{diam}(\Omega_e)$ as the element characteristic dimension. For any element $\Omega_e \in \mathcal{P}(\Omega)$, $P_k(\Omega_e)$ is the finite dimensional space of all polynomials of degree less than, or equal to, k defined on Ω_e. With these, we define the space in which we are going to approximate the solution of our problems as

$$\mathcal{V}^h = \left\{ v^h \in L_2(\Omega) \mid v^h|_{\Omega_e} \in P_k(\Omega_e) \ \ \forall \Omega_e \in \mathcal{P}(\Omega) \right\} \ . \tag{6}$$

In the following we will use the standard notation $(f_1, f_2)_\Theta$ to indicate the L_2 inner product over a domain Θ. Also, we denote by $\widetilde{\Omega} = \bigcup_e \Omega_e$ the union of element interiors and by $\Gamma_{\mathrm{int}} = \bigcup_r \bigcup_{(s \geq r)} \overline{\Omega_r} \bigcap \overline{\Omega_s}$ the union of interior boundaries of the elements.

3.1 Advection

For the hyperbolic case, consider the following discontinuous Galerkin formulation for (1)–(2): Find $\phi^h \in \mathcal{V}^h$ such that

$$B_{\mathrm{a}}(w^h, \phi^h) = L_{\mathrm{a}}(w^h) \quad \forall w^h \in \mathcal{V}^h \ , \tag{7}$$

where

$$\begin{aligned} B_{\mathrm{a}}(w^h, \phi^h) &= \left(\nabla w^h, \boldsymbol{\sigma}^{\mathrm{a}}\left(\phi^h\right)\right)_{\widetilde{\Omega}} - \left(w^h, \sigma_n^{\mathrm{a}}\left(\phi^{h-}\right)\right)_{\Gamma+} \\ &\quad + \left(\llbracket w^h \rrbracket, \sigma_n^{\mathrm{a}}\left(\phi^{h-}\right)\right)_{\Gamma_{\mathrm{int}}} + \left(\tau \mathcal{L}_{\mathrm{a}} w^h, \mathcal{L}_{\mathrm{a}} \phi^h\right)_{\widetilde{\Omega}} \ , \end{aligned} \tag{8}$$

$$L_{\mathrm{a}}(w^h) = \left(w^h, \mathrm{f}\right)_{\widetilde{\Omega}} + \left(w^h, \sigma_n^{\mathrm{a}}(\mathrm{g})\right)_{\Gamma^-} + \left(\tau \mathcal{L}_{\mathrm{a}} w^h, \mathrm{f}\right)_{\widetilde{\Omega}} \ . \tag{9}$$

The jump operator on $\gamma \subset \Gamma_{\mathrm{int}}$ is defined by $\llbracket f \rrbracket = f^+ - f^-$, where $f^{\pm}|_{\gamma} = \lim_{\epsilon \to 0} f(\boldsymbol{x} \pm \epsilon \boldsymbol{n})$, $\boldsymbol{x} \in \gamma$, and $\boldsymbol{n} \perp \gamma$ is such that $\boldsymbol{u} \cdot \boldsymbol{n} > 0$.

This is a stabilized formulation based on the DGM (cf. Johnson and Pitkäranta [5]), in which $\tau = O(h/|\boldsymbol{u}|) > 0$, is the stabilization parameter.

3.2 Diffusion

For the elliptic case, consider the following family of discontinuous Galerkin formulations for (3)–(5): Find $\phi^h \in \mathcal{V}^h$ such that

$$B_{\mathrm{d}}(w^h, \phi^h) = L_{\mathrm{d}}(w^h) \quad \forall w^h \in \mathcal{V}^h \ , \tag{10}$$

where

$$\begin{aligned} B_{\mathrm{d}}(w^h, \phi^h) &= \left(\nabla w^h, \boldsymbol{\sigma}^{\mathrm{d}}\left(\phi^h\right)\right)_{\widetilde{\Omega}} - \left(w^h, \sigma_n^{\mathrm{d}}\left(\phi^h\right)\right)_{\Gamma_{\mathrm{g}}} + \left(\alpha \sigma_n^{\mathrm{d}}\left(w^h\right), \phi^h\right)_{\Gamma_{\mathrm{g}}} \\ &\quad - \left(\llbracket w^h \rrbracket, \langle \sigma_n^{\mathrm{d}}\left(\phi^h\right)\rangle\right)_{\Gamma_{\mathrm{int}}} + \left(\alpha \langle \sigma_n^{\mathrm{d}}\left(w^h\right)\rangle, \llbracket \phi^h \rrbracket\right)_{\Gamma_{\mathrm{int}}} \\ &\quad + \left(\tau w^h, \phi^h\right)_{\Gamma_{\mathrm{g}}} + \left(\tau \llbracket w^h \rrbracket, \llbracket \phi^h \rrbracket\right)_{\Gamma_{\mathrm{int}}} \ , \end{aligned} \tag{11}$$

$$L_{\mathrm{d}}(w^h) = \left(w^h, \mathrm{f}\right)_{\widetilde{\Omega}} + \left(\alpha \sigma_n^{\mathrm{d}}\left(w^h\right), \mathrm{g}\right)_{\Gamma_{\mathrm{g}}} + \left(w^h, \mathrm{h}\right)_{\Gamma_{\mathrm{h}}} + \left(\tau w^h, \mathrm{g}\right)_{\Gamma_{\mathrm{g}}} \ , \tag{12}$$

and $\langle f \rangle = \frac{1}{2}(f^+ + f^-)$. Here, $\alpha \in \mathbb{R}$ is the parameter differentiating the various Galerkin methods, and $\tau = O(A/h) > 0$ is again the stabilization parameter. This setup allows us to study and to generalize within a unified framework several methods previously introduced, as well as introduce methods still unexplored. We note, in particular, that for $\tau = 0$ and $\alpha = -1$, (11)–(12) yield a symmetric method known as the Global Element Method [3]. Setting $\tau = 0$ and $\alpha = 1$ yields the method of Oden et al. [7] in which the interface terms constitute a skew-symmetric operator. If we choose $\alpha = -1$ and $\tau > 0$, we obtain Nitsche's method [6,8,9], generalized to the discontinuous case.

4 Error estimates

4.1 Advection

Consistency. For sufficiently smooth ϕ, we have

$$B_{\mathrm{a}}(w^h, e) \;=\; 0 \quad \forall w^h \in \mathcal{V}^h \;, \tag{13}$$

where $e = \phi^h - \phi$ is the error.

Stability. The relevant norm for (7) is

$$\begin{aligned} |||w^h|||_{\mathrm{a}}^2 \;=\; & \tfrac{1}{2}|| \, |u_n|^{1/2} w^{h-} \, ||_{\Gamma^+}^2 + \tfrac{1}{2}|| \, |u_n|^{1/2} w^{h+} \, ||_{\Gamma^-}^2 \\ & + \tfrac{1}{2}|| \, |u_n|^{1/2} [\![w^h]\!] \, ||_{\Gamma_{\mathrm{int}}}^2 + ||\tau^{1/2} \boldsymbol{u} \cdot \boldsymbol{\nabla} w^h||_{\Omega}^2 \;, \end{aligned} \tag{14}$$

where $||\cdot||_{\Theta}$ denotes the L_2-norm over the domain Θ.

Lemma 1. *We have*

$$B_{\mathrm{a}}(w^h, w^h) = |||w^h|||_{\mathrm{a}}^2 \quad \forall w^h \in \mathcal{V}^h \;. \tag{15}$$

Convergence. Let us introduce the interpolant $\widetilde{\phi}^h$ and decompose the error as follows: $e = (\phi^h - \widetilde{\phi}^h) + (\widetilde{\phi}^h - \phi) = e^h + \eta$, where e^h is the part of the error in the finite element space, and η is the interpolation error. With the previous results we obtain the following theorem:

Theorem 2. *Given the consistency condition (13) and the stability condition (15), and assuming that interpolation estimates of the form*

$$|\eta|_{H^l(\Omega_e)} \;\leq\; C_{\mathrm{i}} h_e^{k+1-l} |\phi|_{H^{k+1}(\Omega_e)} \tag{16}$$

hold, then the following error estimate holds for (7):

$$|||e|||_{\mathrm{a}}^2 \;\leq\; C_{\mathrm{a}} \sum_{e=1}^{N_{\mathrm{el}}} h_e^{2k+1} \; |\phi|_{H^{k+1}(\Omega_e)}^2 \;. \tag{17}$$

4.2 Diffusion

Consistency. For sufficiently smooth ϕ, we have

$$B_{\mathrm{d}}(w^h, e) \;=\; 0 \quad \forall w^h \in \mathcal{V}^h \;. \tag{18}$$

Stability. By appropriately generalizing previous results [1,7], it can be shown that the pertinent norm for (10) is

$$|||w|||_{\mathrm{d}}^2 = ||\boldsymbol{A}^{1/2}\nabla w^h||_{\tilde{\Omega}}^2 + ||\tau^{1/2}w^h||_{\Gamma_{\mathrm{g}}}^2 + ||\tau^{1/2}[\![w^h]\!]||_{\Gamma_{\mathrm{int}}}^2 \ . \tag{19}$$

The following results hold:

Lemma 3. *The bilinear form (11) is stable, that is, for any $\alpha \in \mathbb{R}$ there exists a positive constant m such that*

$$B_{\mathrm{d}}(w^h, w^h) \geq m|||w^h|||_{\mathrm{d}}^2 \quad \forall w^h \in \mathcal{V}^h \ . \tag{20}$$

Remarks

1. (20) is not uniform in α. In particular, it can be shown that we need $\tau = O((A/h)(1 + C|\alpha - 1|^2))$, where C is the product of constants from an inverse estimate and a trace inequality. Note that C increases with k.
2. The presence of the stabilization term makes convergence possible for a wider variety of methods. The parameter α allows "tuning" of the different methods in order to obtain improved convergence properties.

Convergence. As for the advection case, we can prove the following theorem:

Theorem 4. *Given the consistency condition (18) and the stability condition (20), and assuming that the interpolation estimate (16) holds, then the following error estimate holds for (10):*

$$|||e|||_{\mathrm{d}}^2 \ \leq \ C_{\mathrm{d}} \sum_{e=1}^{N_{\mathrm{el}}} h_e^{2k} \ |\phi|_{H^{k+1}(\Omega_e)}^2 \ . \tag{21}$$

5 Conservation

In comparing discontinuous and continuous Galerkin methods, the local conservation property of the former is often identified as an advantageous property, although precisely what is the advantage is often not explained. Let us take the point of view here that local conservation is at least desirable, possibly helpful, and certainly not harmful. Local conservation, and in particular element conservation, emanates from the property that the weighting function can be set exactly to value 1 on the subdomain or element of interest and zero elsewhere. Due to the discontinuous nature of the weighting function space, this is possible in the discontinuous Galerkin method on an element-by-element basis. (In the finite volume method, a similar property holds for the volumes, or covolumes, depending on whether the method is cell, or node, centered, respectively.)

In contrast, it is usually said that the continuous Galerkin method is globally conservative, but not locally conservative. We have trouble with this statement on both counts and are of the opinion that the conservation law structure of the continuous Galerkin method is not very well understood. In a recent study we endeavored to shed some light on this subject [4].

Global conservation requires that the weighting function whose value is precisely 1 throughout the domain of the boundary value problem be present in the weighting function space. This is only the case for no Dirichlet boundary conditions because strong enforcement of the Dirichlet condition necessitates that weighting functions take value zero on the Dirichlet portion of the boundary. Consequently, global conservation only occurs when we have all Neumann boundary conditions. In cases where there are Dirichlet conditions, we can say nothing about global conservation.

However, there is a well-known remedy to the problem of global conservation. Introduce a modified (i.e., "mixed") formulation with an auxiliary field which amounts to the flux on the Dirichlet portion of the boundary. The modified formulation reduces to the usual continuous Galerkin method plus a "post-processing" calculation to determine the flux. This field is expanded in terms of the basis functions omitted to satisfy the homogeneous Dirichlet boundary condition. The resulting flux possesses remarkable properties: (i) It is the missing link in the global conservation structure of the method, and (ii) it achieves superior convergence characteristics. The global conservation law of the governing theory is then obtained for the (modified) continuous Galerkin method. This result then confirms the usual assertion that the continuous Galerkin method is globally conservative.

In [4] we endeavored to obtain a conservation law for a subdomain consisting of a union of connected element domains. It is usually thought that this is not possible because the weighting function taking on value 1 on the subdomain, and identically zero elsewhere, is not available in the continuous Galerkin method. However, we pointed out that the method of establishing global conservation is a paradigm capable of exposing the local conservation structure of the continuous Galerkin method as well. For the subdomain under consideration, we introduced an auxiliary boundary flux field and developed a modified formulation which reduces to the usual continuous Galerkin method plus the previous modification to attain global conservation. With the usual solution of the global auxiliary boundary flux in hand, the new modification entails a subsequent "post-processing" calculation for the auxiliary boundary flux on the subdomain. We showed that this flux is the missing link to conservation on the subdomain and that the formulation thereby attains the exact conservation law on the subdomain. With respect to the subdomain, the auxiliary boundary flux possesses the same remarkable properties as the one obtained to achieve global conservation.

We specialized the results to an individual element subdomain and determined the element conservation law. This result seems to us to refute the notion that the continuous Galerkin method is not locally conservative.

We also argued that the auxiliary flux is a continuous redistribution of the element nodal fluxes which likewise are a conserved quantity. In fact, all conservation properties of the auxiliary fields emanate from the conservation of nodal fluxes. This is where the fundamental conservation structure of the continuous Galerkin method resides and this is why one is able to redistribute the fluxes continuously in a conservative way. It seems that this observation had been missed heretofore.

6 Robustness

6.1 Advection

The advection problem (1) with boundary condition (2) was solved by the DGM and SDGM (7) in one dimension in the interval $]0,1[$ for different orders of interpolation k. The domain was discretized into ten elements of equal length. The problem considered can be written as

$$\frac{d}{dx}\phi = \delta(x - x_0) \qquad \text{in }]0,1[\ , \tag{22}$$

$$\phi = 1 \qquad \text{at } x = 0\ . \tag{23}$$

The location x_0 of the Dirac delta function was varied within the element $0.3 < x < 0.4$.

Monotonicity and Continuity. The DGM with constant interpolation is the only method which retains the monotonicity of the exact solution (Fig. 1). For $k \geq 1$, monotonicity is lost (Figs. 2–4). Stabilization improves the approximate solution (Fig. 2) and yields monotonicity and continuity in the linear case in the limit as $\tau \to \infty$. For higher-order interpolation ($k \geq 2$), however, monotonicity is lost both for the DGM and SDGM (Figs. 3, 4). The SDGM yields continuity in the limit as $\tau \to \infty$, but not monotonicity. In other words, least-squares stabilization cannot ensure an adequate approximation of the solution for higher-order interpolation in the element where the Dirac delta function is located. One also concludes that the p-method is *not* the method of choice in the case of shocks, and that constant interpolation most closely reflects the character of the exact solution.

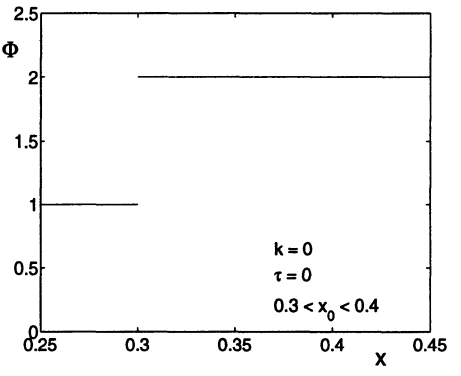


Fig. 1. Discontinuous Galerkin Method for Advection with Constant Interpolation ($k = 0$)

Dependence on Location of Dirac Delta Function. For constant interpolation, the approximation is independent of the location of the Dirac delta function within the element (Fig. 1). For $k \geq 1$, the approximations of the DGM and SDGM strongly depend on the location x_0 of the Dirac delta function. When the Dirac delta function is very close to an upwind node, the DGM yields more accurate results than the SDGM, which can be seen for the case of cubic interpolation in Fig. 4 (i). In the case when the Dirac delta function is close to a downwind node, the results of the SDGM are better than those of the DGM (Fig. 4 (ii)). Again, constant interpolation yields the best results, and the results for linear interpolation with stabilization are acceptable.

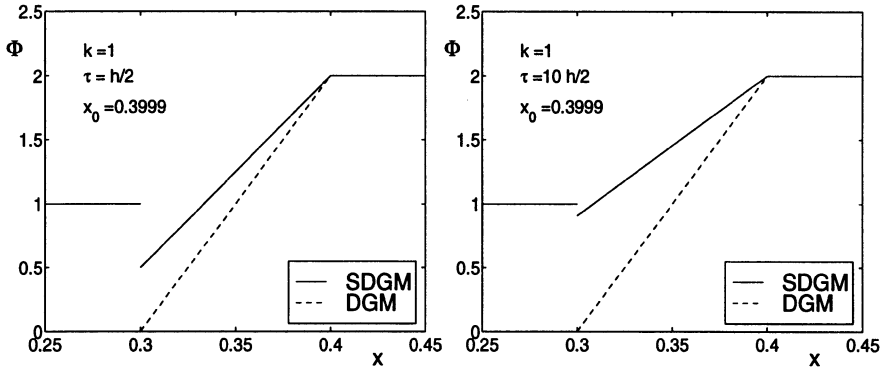


Fig. 2. Discontinuous (DGM) and Stabilized Discontinuous Galerkin Methods (SDGM) for Advection with Linear Interpolation ($k = 1$)

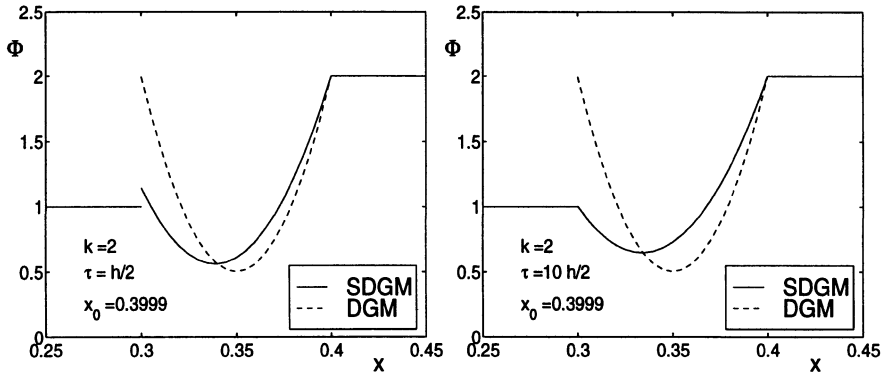


Fig. 3. Discontinuous (DGM) and Stabilized Discontinuous Galerkin Methods (SDGM) for Advection with Quadratic Interpolation ($k = 2$)

Localization Property. An important observation from Figs. 2–4 is that for advection, both DGM and SDGM *localize* the non-monotonicity within one single element. For the diffusion case, however, which will be considered in the following section, this localization property is lost, and the error of one element may pollute the solution in all other elements (see, e.g., Fig. 5). It is easy to show analytically, and may be inferred from Figs. 1–4, that, for the advection problem, at the downwind node in the element containing the delta function, the solution is exact for all k. This follows from the local conservation property.

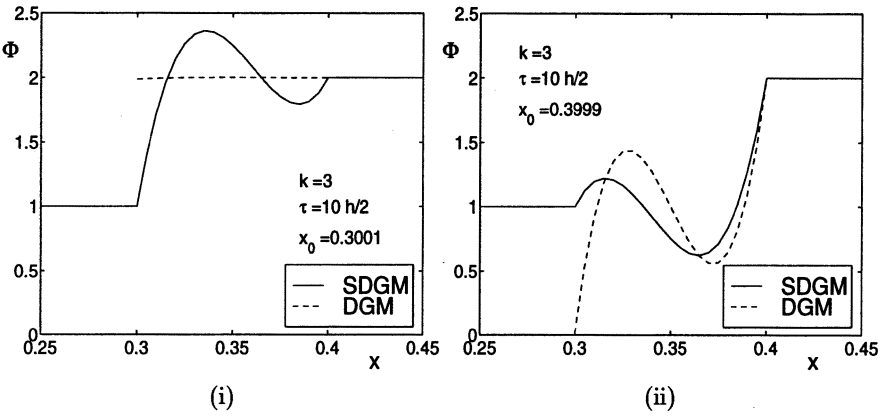


Fig. 4. Discontinuous (DGM) and Stabilized Discontinuous Galerkin Methods (SDGM) for Advection with Cubic Interpolation ($k = 3$)

6.2 Diffusion

We considered the following one-dimensional diffusion problem (cf. (3)–(5)):

$$-\frac{d^2}{dx^2}\phi = \delta(x - x_0) \qquad \text{in }]0,1[\ , \tag{24}$$

$$\phi = 0 \qquad \text{for } x \in \{0,1\}\ . \tag{25}$$

Again, the domain was discretized into ten elements of equal length, and the location x_0 of the Dirac delta function was varied within the element $0.6 < x < 0.7$. We compared stabilized ($\tau > 0$) and non-stabilized ($\tau = 0$) members of the family of DGMs (10) for different orders of interpolation.

Skew-Symmetric and Symmetric Formulations. We refer to the case $\alpha = -1$ and $\alpha = 1$ as the symmetric and skew-symmetric formulations, respectively. For $\alpha = -1$, the entire formulation is symmetric, whereas for $\alpha = 1$ only the Galerkin interface terms are skew-symmetric, the remaining terms

being symmetric. The non-stabilized skew-symmetric method, introduced by Oden et al. [7], is stable for quadratic and higher-order interpolation. However, as can be seen in Fig. 5 (i) for quadratic interpolation, the approximation can be quite inaccurate in some cases. Increasing the stabilization parameter τ leads to better approximations (Figs. 5 (ii), (iii)). Let us mention at this point that in the case of linear interpolation, the skew-symmetric method without stabilization is unstable, while good convergence is achieved for the stabilized version.

The stabilization terms in (10) (i.e., τ terms) were originally introduced by Nitsche [6]. However, Nitsche's method amounts to the symmetric, rather than skew-symmetric, formulation, which is unstable for τ not large enough. Invoking Nitsche's stabilization terms in the skew-symmetric formulation results in a robust method which seems to yield the best discontinuous approximation to the diffusion problem as can be seen from Fig. 5.

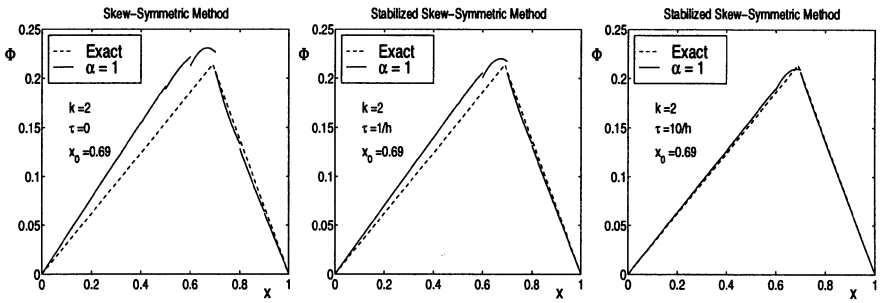


Fig. 5. Quadratic DGM and SDGM for the Diffusion Problem

Global Pollution. For the diffusion problem, the advantageous localization property of the DGM encountered in the advection case is *lost*, and an off-centered Dirac delta function in one single element causes a deteriorated approximation globally. This problem can be fixed by increasing τ (Fig. 5).

Choice of Stabilization Parameter. In order to obtain a stable discrete formulation for the symmetric formulation ($\alpha = -1$), we need to choose τ sufficiently large.

In contrast, the skew-symmetric formulation ($\alpha = 1$), is stable for all $\tau > 0$. This is a desireable property, since choosing a suitable τ for the symmetric formulation may be difficult without knowledge of the smallest eigenvalue. Note that the condition number increases for large values of τ, and therefore it is also important not to choose the stabilization parameter too large.

We investigated the minimum and maximum eigenvalues λ of the discrete problem as a function of τ for $\alpha = -1$ and $\alpha = 1$, and linear, quadratic and

cubic interpolation. In Fig. 6, one can see in each graph the largest and smallest eigenvalues as functions of τ. The circles in the graphs indicate the value of τ when the smallest eigenvalue becomes positive. Note, in particular, that the symmetric formulation is indefinite for τ too small, and that the critical value of τ increases with the order of interpolation. (On the scale of the graphs, the smallest eigenvalue often plots as zero, even though it is positive.)

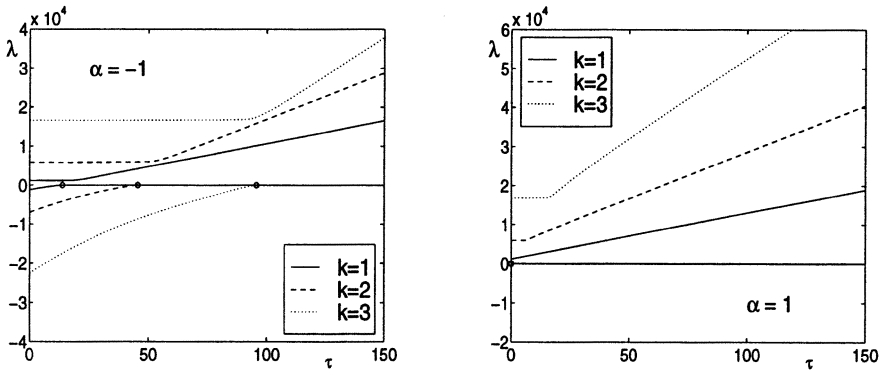


Fig. 6. Largest and Smallest Eigenvalues λ as Functions of τ. The Circles in the Graphs Indicate the Value of τ when the Smallest Eigenvalue becomes Positive

7 Efficiency

The number of unknowns of a discretized problem is a good indicator for the efficiency of a numerical method. The DGM enables constant interpolation on an element, which is impossible with the CGM, and this is an interesting possibility for certain problems. For the orders of interpolation commonly used in the CGM, the DGM seems inefficient. Table 1 gives an overview of the ratio of the number of unknowns in the DGM to the number of unknowns in the CGM for different orders of interpolation k and commonly used two- and three-dimensional elements. In the case of triangles and tetrahedra, the ratio is based on regular meshes obtained from subdivisions of quadrilateral and hexahedral meshes, respectively. Note that, in the limit as $k \to \infty$, this ratio approaches 1. Very high-order DGMs involve a similar number of unknowns as corresponding CGMs.

8 Conclusions

An error analysis showed that, with appropriately chosen stabilization parameters, all DGMs converge with optimal rates of convergence. Local conservation properties of the DGM are shared by the CGM after using a post-

Table 1. Number of Unknowns for the DGM as a Multiple of the Number of Unknowns of the CGM for Different Elements and Orders of Interpolation k

k	Quadrilateral	Triangle	Hexahedron	Tetrahedron
0*	1	2	1	5
1	4	6	8	20
2	2.25	3	3.38	7.14
3	1.78	2.22	2.37	4.35
∞	1	1	1	1

* For $k = 0$, normalization is with respect to the number of unknowns of the CGM for linear interpolation, i.e., $k = 1$.

processing procedure. The numerical results indicate that the DGM for advection with constant interpolation attains monotonicity, but this property is lost for higher-order interpolation. The localization property of the DGM for the advection problem ensures that local errors do not pollute the approximate solution globally. For the diffusion problem, numerical experiments showed the superiority of the skew-symmetric form over the symmetric form of Nitsche, and the need for stabilization. A comparison of the number of unknowns of the DGM and the CGM revealed that the DGM involves significantly more unknowns for the same interpolation order, except for k very large.

References

1. R. Becker and P. Hansbo. A Finite Element Method for Domain Decomposition with Non-Matching Grids. Technical Report 3613, INRIA, Sophia Antipolis, France, Jan. 1999.
2. S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods.* Springer-Verlag, New York, 1994.
3. M. L. Delves and C. A. Hall. An Implicit Matching Principle for Global Element Calculations. *J. Inst. Maths. Appl.*, **23**:223–234, 1979.
4. T. J. R. Hughes, G. Engel, L. Mazzei, and M. G. Larson. The Continuous Galerkin Method is Locally Conservative. Preprint, 1999.
5. C. Johnson and J. Pitkäranta. An Analysis of the Discontinuous Galerkin Method for a Scalar Hyperbolic Equation. *Math. Comp.*, **46**:1–26, 1986.
6. J. Nitsche. Über ein Variationsprinzip zur Lösung von Dirichlet Problem bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind. *Abh. Math. Sem. Univ. Hamburg*, **36**:9–15, 1971.
7. J. T. Oden, I. Babuska, and C. E. Baumann. A Discontinuous *hp* Finite Element Method for Diffusion Problems. *J. Comput. Phys.*, **146**:491–519, 1998.
8. R. Stenberg. On some Techniques for Approximating Boundary Conditions in the Finite Element Method. *J. Comput. Appl. Math*, **63**:139–148, 1995.
9. R. Stenberg. Mortaring by a Method of J. A. Nitsche. In S. Idelsohn, E. Oñate, and E. Dvorkin, editors, *Computational Mechanics-New Trends and Applications.* CIMNE, Barcelona, Spain, 1998.

The Utility of Modeling and Simulation in Determining Transport Performance Properties of Semiconductors

Bernardo Cockburn[1], Joseph W. Jerome[2], and Chi-Wang Shu[3]

[1] Department of Mathematics, University of Minnesota, Minneapolis, MN 55455
[2] Department of Mathematics, Northwestern University, Evanston, IL 60208
[3] Division of Applied Mathematics, Brown University, Providence, RI 02912

Abstract. The RKDG method has been effectively used in modeling and simulating semiconductor devices, where the underlying models are hydrodynamic in nature. These include classical as well as quantum models. In this paper, we survey and interpret some of these results. For classical transport, we review the simulation of a benchmark MESFET transistor by means of discontinuous Galerkin methods of degree one. For quantum transport, we report the success in simulation of the resonant tunneling diode. The principal features here are negative differential resistance and hysteresis.

1 Introduction

The goal of this work is to survey the effectiveness of continuum (hydrodynamic) models in one and two dimensions via discontinuous Galerkin methods, which were effectively used in [4] and in [5] for classical and quantum models, respectively. We choose for the underlying classical application that of charge transport in a MESFET transistor. This is a benchmark which has been studied intensively, and thus its characteristics are reliably determined (see [10,11,4,3]). For the quantum application, we select another benchmark, the resonant tunneling diode. For this device, Gardner developed a quantum hydrodynamic model in [8] (consult also for relevant references in the physics and device literature). Important characteristics of the hydrodynamic model include heat conduction, relaxation, and electrical forcing and heating terms. In particular, carrier transport occurs in a self-consistent electric field. The model is decidedly more complex than the standard gas dynamics model. The quantum model (QHD) includes perturbation terms in the pressure tensor and energy expression. These are characterized in the QHD model as third order derivative perturbations of the concentrations.

2 The Classical Hydrodynamic Model

The hydrodynamic model may be described as in [2,4]. A derivation is provided in [9] and an existence theorem for the reduced, two-carrier model is

given in [3]. It may be characterized as a second-order perturbation of a nonlinear hyperbolic system for n, the electron density, $\mathbf{p}$, the momentum density, and w, the energy density,

$$\partial_t n + \nabla\cdot(n\mathbf{v}) = 0, \tag{1}$$

$$\partial_t \mathbf{p} + \mathbf{v}\nabla\cdot\mathbf{p} + \mathbf{p}\cdot\nabla\mathbf{v} + \nabla(knT) = -en\mathbf{E} + (\partial_t \mathbf{p})_c, \tag{2}$$

$$\partial_t w + \nabla\cdot(\mathbf{v}w) + \nabla\cdot(\mathbf{v}knT + \mathbf{q}) = -en\mathbf{v}\cdot\mathbf{E} + (\partial_t w)_c, \tag{3}$$

where k is Boltzmann's constant and the velocity $\mathbf{v}$, the temperature T, and the heat flux $\mathbf{q}$ are given by

$$\mathbf{p} = mn\mathbf{v}, \tag{4}$$

$$w = \tfrac{3}{2}knT + \tfrac{1}{2}mn|\mathbf{v}|^2, \tag{5}$$

$$\mathbf{q} = -\nabla\cdot(\kappa\nabla\mathbf{T}), \tag{6}$$

where m is the effective electron mass. These equations are coupled with a Poisson equation defining the electric field $\mathbf{E}$:

$$\mathbf{E} = -\nabla\phi, \tag{7}$$

$$\nabla\cdot(\epsilon\nabla\phi) = -e(n_d - n), \tag{8}$$

where ϵ is the dielectric constant, and n_d is the doping density. The constant $e > 0$ is the electronic charge and κ is the heat conduction coefficient. The 'collision' terms are obtained by defining the momentum and energy relaxation times, $\tau_{\mathbf{p}}$ and τ_w, following [1] as

$$(\partial_t \mathbf{p})_c = -\frac{\mathbf{p}}{\tau_{\mathbf{p}}}, \qquad \tau_{\mathbf{p}} = m\frac{\mu_{n0}}{e}\frac{T_0}{T}, \tag{9}$$

$$(\partial_t w)_c = -\frac{w - \frac{3}{2}nT_0}{\tau_w}, \qquad \tau_w = \frac{\tau_{\mathbf{p}}}{2} + \frac{3}{2}\frac{\mu_{n0}}{ev_s^2}\frac{kTT_0}{T + T_0}, \tag{10}$$

where T_0 is the ambient temperature, $\mu_{n0} = \mu_{n0}(T_0, n_d)$ is the low field electron mobility, and $v_s = v_s(T_0)$ is the saturation velocity. Finally, κ is determined by the Wiedemann-Franz law

$$\kappa = \kappa_0\frac{\mu_{n0}}{e}k^2nT\left(\frac{T}{T_0}\right)^r.$$

In this paper, we take $r = -1$. We have selected a MESFET because of its acknowledged importance, particularly in microwave applications. It represents an application for which numerical methods are required to be robust over a wide parameter regime, although in this paper we restrict attention to ambient room temperature. We emphasize the importance of retention of the convective term, $\mathbf{p}\cdot\nabla\mathbf{v}$, in (2), if a robust model is desired. This is the term which permits shocks in the hydrodynamic model when present.

3 Numerical Method

3.1 General Description

To describe our numerical method, we first write the initial boundary value problem for $\mathbf{u} = (n, p_x, p_y, w)^t$ as follows:

$$\partial_t \mathbf{u} + \nabla \cdot \mathbf{F}(\mathbf{u}) = \mathbf{R}(\mathbf{u}), \quad \text{in } (0, t_f) \times \Omega, \tag{11}$$

$$\mathbf{u}(t = 0) = \mathbf{u}_0, \quad \text{on } \Omega, \tag{12}$$

$$\mathbf{B}\mathbf{u} = \mathbf{g}, \quad \text{on } (0, t_f) \times \partial\Omega, \tag{13}$$

where the flux $\mathbf{F} = (\mathbf{f}_x, \mathbf{f}_y)$ has the following components:

$$\mathbf{f}_x(\mathbf{u}) = v_x \mathbf{u} + (0, nT, 0, v_x nT)^t, \tag{14}$$

$$\mathbf{f}_y(\mathbf{u}) = v_y \mathbf{u} + (0, 0, nT, v_y nT)^t, \tag{15}$$

the right-hand side $\mathbf{R}$ is given by

$$\mathbf{R}(\mathbf{u}) = \xi_{\mathbf{E}}(\mathbf{u}) + \xi_c(\mathbf{u}) + \xi_{heat}(\mathbf{u}), \tag{16}$$

$$\xi_{\mathbf{E}}(\mathbf{u}) = (0, -e\,n\,E_x, -e\,n\,E_y, -e\,n\,\mathbf{v} \cdot \mathbf{E})^t, \tag{17}$$

$$\xi_c(\mathbf{u}) = \left(0, (\partial_t p_x)_c, (\partial_t p_y)_c, (\partial_t w)_c\right)^t, \tag{18}$$

$$\xi_{heat}(\mathbf{u}) = (0, 0, 0, \nabla \cdot (\kappa \nabla T))^t, \tag{19}$$

and $\mathbf{B}$ is a matrix-valued function.

An overview of the discretization of our equations is as follows. First, we triangulate our domain Ω with triangulations $\mathcal{T}_h$ made solely of rectangles R such that the intersection of two distinct rectangles of the triangulation $\mathcal{T}_h$ is either an edge, a vertex, or void. Then, for each $t \in (0, t_f]$, we take each of the components of our approximate solution $\mathbf{u}_h(t)$ in the space

$$V_h = \{p \in L^\infty(\Omega) : \ p|_R \ \text{ is linear}, \forall R \in \mathcal{T}_h\}. \tag{20}$$

We define each of the components of $\mathbf{u}_{0h}$ to be the L^2-projection of the corresponding component of $\mathbf{u}_0$ into V_h and discretize the equation (11) in space by using the Discontinuous Galerkin (DG) method. Since the functions of the space V_h are discontinuous, the mass matrix of the DG method is block-diagonal. Thus, the resulting discrete equations can be rewritten as the following ODE initial value problem:

$$\frac{d\mathbf{u}_h}{dt} = \mathbf{L}_h(\mathbf{u}_h, \mathbf{g}) + \mathbf{R}_h(\mathbf{u}_h), \quad t \in (0, t_f], \tag{21}$$

$$\mathbf{u}_h(t = 0) = \mathbf{u}_{0h}, \tag{22}$$

where $\mathbf{L}_h$ is the approximation of $-\nabla\cdot\mathbf{F}$. The exact solution of the above initial value problem gives an approximation which is formally second-order accurate in space; see [6]. Accordingly, a second-order accurate in time Runge-Kutta method must be used to discretize our ODE; see [6], [13], and [14]. Finally, a local projection $\Lambda\Pi_h$ is applied to the intermediate values of the Runge-Kutta discretization in order to enforce nonlinear stability. We give a short description of several components of the algorithm below but refer the reader to the cited papers for more details.

3.2 The Discontinuous Galerkin method

The general definition of the DG method in the case of a scalar $\mathbf{u}$ can be found in [6]. To define the method in our case, we simply have to apply the procedure for the scalar case component by component.

Let us denote by $\mathbf{u}^{\{k\}}$ the k-th component of the vector $\mathbf{u}$. Consider the equation for the k-th component of the system (11), multiply it by $v_h \in V_h$, integrate over each $R \in \mathcal{T}_h$, replace the exact solution $\mathbf{u}$ by its approximation $\mathbf{u}_h$, and formally integrate by parts to obtain

$$\begin{aligned}
&\frac{d}{dt}\int_R \mathbf{u}_h^{\{k\}}(t,x,y)\, v_h(x,y)dx\,dy \\
&+\sum_{e\in\partial R}\int_e \mathbf{F}^{\{k\}}(\mathbf{u}_h(t,x,y))\cdot\mathbf{n}_{e,R}\, v_h(x,y)d\,\Gamma(x,y) \\
&\qquad -\int_R \mathbf{F}^{\{k\}}(\mathbf{u}_h(t,x,y))\cdot\nabla v_h(x,y)dx\,dy \\
&=\int_R \mathbf{R}^{\{k\}}(\mathbf{u}_h(t,x,y))\; v_h(x,y)dx\,dy,\ \forall v_h\in V_h, \qquad (23)
\end{aligned}$$

where $\mathbf{n}_{e,R}$ is the outward unit normal to the edge e. Notice that $\mathbf{F}\cdot\mathbf{n} = \mathbf{f}_\mathbf{x}\mathbf{n}_\mathbf{x} + \mathbf{f}_\mathbf{y}\mathbf{n}_\mathbf{y}$ is a four-dimensional vector whose k-th component is $\mathbf{F}^{(\mathbf{k})}\cdot\mathbf{n} = \mathbf{f}_\mathbf{x}{}^{(\mathbf{k})}\mathbf{n}_\mathbf{x} + \mathbf{f}_\mathbf{y}{}^{(\mathbf{k})}\mathbf{n}_\mathbf{y}$. Notice also that $\mathbf{F}(\mathbf{u}_h(t,x,y))\cdot\mathbf{n}_{e,R}$ does not have a precise meaning, since $\mathbf{u}_h$ is discontinuous at $(x,y)\in e\in\partial R$. Thus, we replace $\mathbf{F}(\mathbf{u}_h(t,x,y))\cdot\mathbf{n}_{e,R}$ by a suitably chosen numerical flux $\mathbf{h}_{e,R}$, which depends on the two values of $\mathbf{u}_h$ on the edge e. The choice of this numerical flux is crucial since it is through the use of the numerical flux that the upwinding (or the artificial viscosity) which renders the method stable (without destroying its high-order accuracy) is introduced. In this paper, we choose the so-called local Lax-Friedrichs flux. Finally, we replace the integrals above by quadrature rules to obtain the discrete equations. In this way, we obtain a weak formulation which defines the operators $\mathbf{L}_h$ and $\mathbf{R}_h$.

3.3 The Local Projection $\Lambda\Pi_h$

The local projection (limiter) is devised to prevent the appearance of spurious oscillations in the approximate solution. The local averages are unchanged

to preserve the conservativity of the method, but the local variations in the x-direction and in the y-direction must be controlled to avoid the unphysical oscillations. One can alternatively take into account the local characteristic directions along which information travels with different speeds. Taking these characteristic directions into account results in a better control of the oscillations and in a higher quality of the approximation.

3.4 The Right-Hand Side $\mathbf{R}(\mathbf{u}_h)$

In this section we show how to evaluate the function $\mathbf{R}(\mathbf{u_h}) = \xi_{\mathbf{E}}(\mathbf{u_h}) + \xi_c(\mathbf{u_h}) + \xi_{heat}(\mathbf{u_h})$ for a given $\mathbf{u_h}$.

To evaluate $\xi_c(\mathbf{u_h})$, we simply use the equations (9–10) and (4–5). To evaluate $\xi_{\mathbf{E}}(\mathbf{u_h})$, we need a numerical method to obtain an approximation $\mathbf{E}_h$ to the electric field $\mathbf{E}$. The equations defining the electric field are the equations (7–8) and some boundary conditions we write as follows:

$$\phi = \phi_D, \qquad \text{on } \partial\Omega_D, \tag{24}$$

$$\mathbf{E}\cdot\mathbf{n} = \mathbf{0}, \qquad \text{on } \partial\Omega_N, \tag{25}$$

where $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$ and $\partial\Omega_D \cap \partial\Omega_N = \emptyset$. We discretize these equations with the lowest-order Raviart-Thomas mixed method which defines the approximation $(\mathbf{E}_h, \phi_h) \in \mathbf{U_h} \times \mathbf{W_h}$ as the solution of the following weak formulation:

$$(\nabla\cdot\mathbf{E}_h, w) = \left(\frac{e}{\epsilon}(n_d - n_h), w\right), \forall w \in W_h, \tag{26}$$

$$(\mathbf{E}_h, \mathbf{v}) - (\phi_h, \nabla\cdot\mathbf{v}) = -\langle\phi_D, \mathbf{v}\cdot\mathbf{n}\rangle_{\partial\Omega_D}, \forall \mathbf{v} \in \mathbf{U}_h, \tag{27}$$

where n_h is the approximate density given by the RKDG method, and

$$\mathbf{U}_h = \left\{\mathbf{v} \in \mathbf{H}(\nabla\cdot;\Omega) : \mathbf{v}|_R = (a_R^1 + a_R^2 x, a_R^3 + a_R^4 y)\right\},$$

$$a_R^i \in \mathbb{R},\ \forall R \in \mathcal{T}_h; \mathbf{v}\cdot\mathbf{n}|_{\partial\Omega_N} = 0, \tag{28}$$

$$W_h = \left\{w \in L^2(\Omega) : w|_R \quad \text{is a constant, } \forall R \in \mathcal{T}_h\right\}. \tag{29}$$

It can be shown that the above system has a unique solution in $\mathbf{U_h} \times \mathbf{W_h}$ whose approximation of the electric field is second-order accurate. We use Lagrange multipliers, which render the matrix of the resulting method a symmetric positive definite matrix. We invert it by using the conjugate gradient method with incomplete Choleski factorization as preconditioner. To evaluate $\xi_{heat}(\mathbf{u_h})$ we also use the Raviart-Thomas spaces; the procedure is analogous.

We remark that the procedure used here for evaluating the second derivative terms in $\xi_{heat}(\mathbf{u_h})$ is efficient only for the second order schemes, due to mass lumping. A more general approach which keeps the local property of the discontinuous Galerkin method and works for arbitrarily high order of accuracy is the local discontinuous Galerkin method in [7].

4 The Simulation of the MESFET

4.1 Basic MESFET Description

Next we describe a two dimensional MESFET of the size $0.6 \times 0.2\mu m^2$. The source and the drain each occupies $0.1\mu m$ at the upper left and the upper right, respectively, with a gate occupying $0.2\mu m$ at the upper middle (Fig. 1). The doping is defined by $n_d = 3 \times 10^5 \mu m^{-3}$ in $[0, 0.1] \times [0.15, 0.2]$ and in $[0.5, 0.6] \times [0.15, 0.2]$, and $n_d = 1 \times 10^5 \mu m^{-3}$ elsewhere. We apply, at the drain, voltage biases varying up to $vbias = 2$V. This bias has been described in [3] as a symmetry breaking parameter for the concentration and velocity, with respect to the center of the gate. The gate is a Schottky contact, with negative voltage bias up to $vgate = -0.8$V and very low concentration value $n = 3.8503 \times 10^{-8} \mu m^{-3}$ (following Selberherr [12]). The lattice temperature is taken as $T_0 = 300$ K. The mathematical model for the MESFET is the system (1–3), coupled to Poisson's electrostatic equation (7–8).

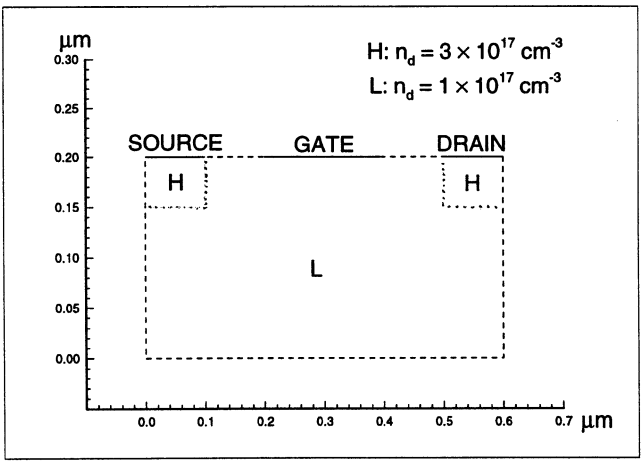


Fig. 1. Two dimensional MESFET. The geometry and the doping n_d.

4.2 Characteristics

We display the concentration, n, in Fig. 2 below, as well as the tangential velocity component, v_x, in Fig. 3 below, obtained from the discontinuous Galerkin method described above. Uniform rectangular meshes of 96×32 and 192×64 were employed for the simulations in which the method is run until the steady state is reached. The results shown are those obtained with the 192×64 mesh. The boundary conditions are determined as follows.

(i) At the source, gate, and drain: $n = n_d$, $v_x = 0\,\mu m/ps$.
(ii) At all other parts of the boundary: Homogeneous Neumann conditions are imposed.

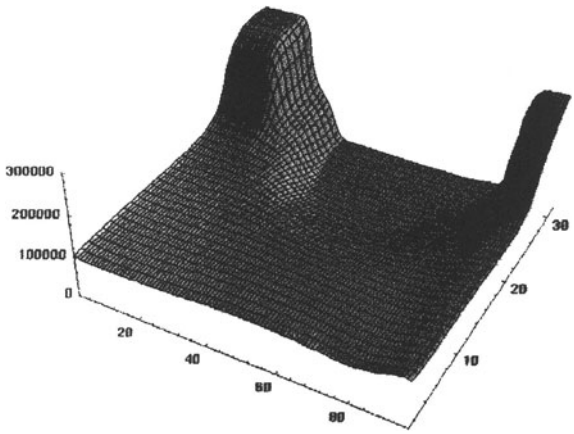


Fig. 2. Concentration n, per μ^3. Domain slightly rotated clockwise from Fig. 1. Axis units defined by (2) grid points per unit.

Notice the boundary layer for n at the drain, but not at the source. This is reasonable since the drain is an outflow boundary and the source is an inflow boundary. A rapid drop of n at the depletion region occurs near the gate. The normal velocity component at the gate appears to be negligible, while the horizontal component shows evidence of strong carrier movement toward the source beneath the left gate area, and strong movement toward the drain immediately to the left of the drain junction. Notice the cusps and strong gradients in the components of the velocity.

5 The Quantum Hydrodynamic Model

The quantum hydrodynamic model used in this paper was derived by Gardner in [8]. In this section, we shall review the basic characteristics of the model

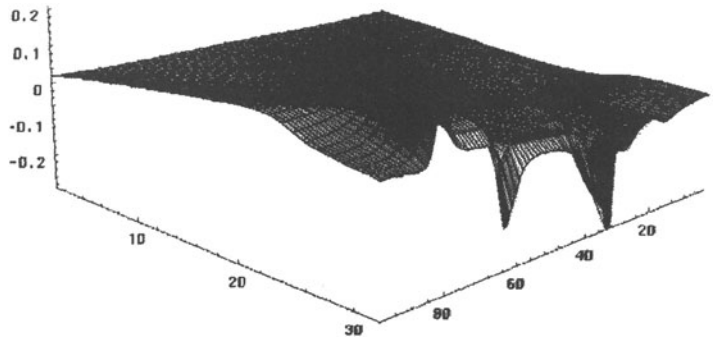


Fig. 3. Horizontal Velocity Component v_x, in μm/ps. See Fig. 2.

as it was described in [5]. An existence theorem for the reduced model was obtained in [15]. The model is also discussed in [9].

The QHD model has exactly the same structure as the classical hydrodynamic model (electrogasdynamics), where we now permit a non-isotropic stress tensor:

$$\frac{\partial n}{\partial t} + \frac{\partial}{\partial x_i}(nv_i) = 0, \tag{30}$$

$$\frac{\partial}{\partial t}(mnv_j) + \frac{\partial}{\partial x_i}(v_i mnv_j - P_{ij}) = -n\frac{\partial V}{\partial x_j} - \frac{mnv_j}{\tau_p} \tag{31}$$

$$\frac{\partial w}{\partial t} + \frac{\partial}{\partial x_i}(v_i w - v_j P_{ij} + q_i) = -nv_i\frac{\partial V}{\partial x_i} - \frac{\left(w - \frac{3}{2}nT_0\right)}{\tau_w} \tag{32}$$

in conjunction with Poisson's equation, (7–8), where P_{ij} is the stress tensor, $V = -e\phi$ is the potential energy, T_0 is the temperature of the semiconductor lattice in energy units (k is set equal to 1), Spatial indices i, j equal 1, 2, 3, and repeated indices are summed over. T is the electron temperature in energy units.

Quantum mechanical effects appear in the stress tensor and the energy density. Gardner derived the stress tensor and the energy density based upon the $O(\hbar^2)$ momentum-shifted thermal equilibrium Wigner distribution function:

$$P_{ij} = -nT\delta_{ij} + \frac{\hbar^2 n}{12m}\frac{\partial^2}{\partial x_i \partial x_j}\log(n) + O(\hbar^4) \tag{33}$$

$$w = \frac{3}{2}nT + \frac{1}{2}mnu^2 - \frac{\hbar^2 n}{24m}\nabla^2\log(n) + O(\hbar^4). \tag{34}$$

In one dimension, the QHD model requires eight boundary conditions. Well-posed boundary conditions for the resonant tunneling diode are $n = n_d$,

$\partial n/\partial x = 0$, and $\partial T/\partial x = 0$ at the left and right diode boundaries x_L and x_R, with a bias ΔV across the device: $V(x_L) = T\log(n/n_i)$ and $V(x_R) = T\log(n/n_i) + e\Delta V$, where n_i is the intrinsic electron concentration.

To exhibit hysteresis, we simulate a GaAs resonant tunneling diode with double $Al_{0.3}Ga_{0.7}As$ barriers (the barrier height $\mathcal{B} = 0.209$ eV). The doping density $n_d = 10^{18}$ cm^{-3} in the n^+ source and drain, and $n_d = 5 \times 10^{15}$ cm^{-3} in the n channel. The channel is 250 Å long, the barriers are 50 Å wide, and the well between the barriers is 50 Å wide. The device has 50 Å spacers between the barriers and the contacts (source and drain) to enhance negative differential resistance.

The current-voltage curve for the resonant tunneling diode is plotted in Fig. 4 for ΔV increasing from 0 volts to 0.22 volts (upper curve) and decreasing from 0.22 volts to 0 volts (lower curve). Note that hysteresis occurs predominantly in the region of negative differential resistance. The physical mechanism for hysteresis is that electrons "see" a different potential energy due to different accumulated electron charges in the diode when the applied voltage is decreasing than when the applied voltage is increasing.

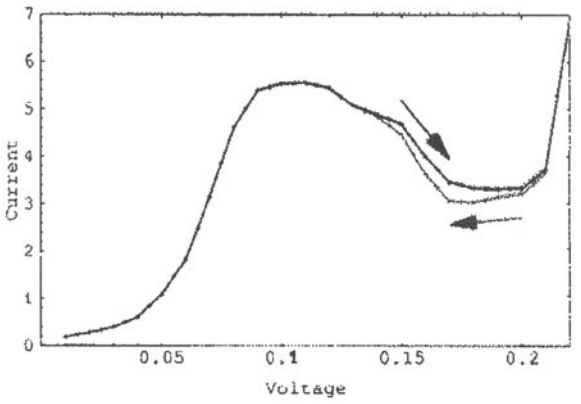


Fig. 4. Current-Voltage Curve.

Acknowledgments: We would like to thank Professor Umberto Ravaioli for his help in formulating the parameters for the MESFET transistor. The research of the first author is partially supported by NSF grant DMS-9807491. The research of the second author is partially supported by NSF grant DMS-9704458. The research of the third author is partially supported by NSF grant ECS-9627849 and ARO grant DAAG55-97-1-0318.

References

1. Baccarani, G., Wordeman, M.R.: An investigation of steady-state velocity overshoot effects in Si and GaAs devices. Solid State Electr., **28** (1985) 407–416
2. Bløtekjær, K.: Transport equations for electrons in two–valley semiconductors. IEEE Trans. Electron Devices, **17** (1970) 38–47
3. Chen, G.-Q., Jerome, J. W., Shu, C.-W., Wang, D.: Two carrier semiconductor device models with geometric structure and symmetry properties. In: J. Jerome (ed.) *Modelling and Computation for Applications in Mathematics, Science, and Engineering*, Clarendon Press, Oxford 1998, pp. 103–140
4. Chen, Z., Cockburn, B., Jerome, J. W., Shu, C.-W.: Mixed-RKGD finite element methods for the 2-D hydrodynamic model for semiconductor device simulation. VLSI DESIGN **3** (1995) 145–158
5. Chen, Z., Cockburn, B., Gardner, C., Jerome, J. W.: Quantum hydrodynamic simulation of hysteresis in the resonant tunneling diode. J. Comp. Phys. **117** (1995) 274–280
6. Cockburn, B., Hou, S., Shu, C.-W.: TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: the multidimensional case. Math. Comp. **54** (1990) 545–581
7. Cockburn, B., Shu, C.-W.: The local discontinuous Galerkin method for time-dependent convection-diffusion systems, SIAM J. Numer. Anal. **35** (1998) 2440–2463
8. Gardner, C.L.: The quantum hydrodynamic model for semiconductor devices. SIAM J. Appl. Math. **54** (1994) 409–427
9. Jerome, J. W.: *Analysis of Charge Transport: A Mathematical Study of Semiconductor Devices.* Springer-Verlag, Heidelberg, 1996
10. Jerome, J. W., Shu, C.-W.: Energy models for one-carrier transport in semiconductor devices. In: W.M. Coughran. J. Cole, P. Lloyd, and J.K. White (eds.) *Semiconductors, Part II. IMA Volumes in Mathematics and Its Applications*, vol. 59. Springer, New York 1994, pp. 185–207
11. Jerome, J. W., Shu, C.-W.: Transport effects & characteristic modes in the modeling & simulation of submicron devices. IEEE Trans. on Computer–Aided Design. **14** (1995) 917–923
12. Selberherr, S.: *Analysis and Simulation of Semiconductor Devices.* Springer-Verlag, Wien-New York, 1984
13. Shu, C.-W., Osher, S.J.: Efficient implementation of essentially non-oscillatory shock capturing schemes. J. Comp. Phys. **77** (1988) 439–471
14. Shu, C.-W., Osher, S.J.: Efficient implementation of essentially non-oscillatory shock capturing schemes, II. J. Comp. Phys. **83** (1989) 32–78
15. Zhang, B., Jerome, J. W.: On a steady-state quantum hydrodynamic model for semiconductors. Nonlinear Anal. **26** (1996) 845–856

A Discontinuous Galerkin Method for the Incompressible Navier-Stokes Equations

Ohannes Karakashian[1] and Theodoros Katsaounis[2]

[1] Dept. Of Mathematics, University of Tennessee, Knoxville TN 37996-1300, USA
[2] Dept. Of Mathematics, University Of Crete, Iraklion, Crete, Greece

Abstract. Approximations to solutions of the inhomogeneous boundary value problem for the Navier-Stokes equations are constructed via the discontinuous Galerkin method. The velocity field is approximated using piecewise polynomial functions that are totally discontinuous across interelement boundaries and which are pointwise divergence-free on each element (locally solenoidal). The pressure is approximated by standard continuous piecewise polynomial functions.

1 Introduction

We consider the stationary Navier-Stokes equations for viscous incompressible flow as given in the primitive variable formulation

$$-\nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } \Omega, \tag{1}$$

$$\operatorname{div}\mathbf{u} = 0 \quad \text{in } \Omega, \tag{2}$$

$$\mathbf{u} = \mathbf{g} \quad \text{on } \partial\Omega\,. \tag{3}$$

Here $\mathbf{u} = (u_1, \cdots, u_N) : \Omega \to R^N$ represents the velocity field, and $p : \Omega \to R$, the pressure; the function $\mathbf{f} = (f_1, \cdots, f_N) : \Omega \to R^N$ denotes the prescribed external body forces, $\mathbf{g} = (g_1, \cdots, g_n) : \partial\Omega \to R^N$, the admitted flux across the boundary $\partial\Omega$ and $\nu > 0$ is a constant measuring viscosity. Note that $\mathbf{g}$ must satisfy the compatibility condition $\int_{\partial\Omega} \mathbf{g}\cdot\mathbf{n}\, d\sigma = 0$.

One of the attributes of the numerical method presented herein is the use of totally discontinuous piecewise polynomial vector functions to approximate the velocity field $\mathbf{u}$. These functions satisfy the incompressibility condition (2) pointwise on each element of a partition of Ω. A special weak formulation is designed to account for the interelement jumps that ensue. For the pressure, standard continuous piecewise polynomial functions are used. For the method that we shall describe below, a range of theoretical issues including the stability and convergence of the approximations at the optimal rates were presented in [11]. These results extended those obtained in [5] for the Stokes problem. Also, in [13], the application of implicit Runge-Kutta methods to the corresponding time dependent problem were analyzed.

In this paper, we outline the method for the stationary problem (1)–(3) and give a summary of these results. In addition, a small number of numerical experiments displaying optimal convergence rates for the errors and other

features are presented. A more substantial array of numerical experiments will be provided in a forthcoming work [12].

2 The energy spaces

We shall next construct appropriate settings for the velocity, the pressure and their approximations as well as the Galerkin formulation. To begin, we consider partitions $\mathcal{T}_k = \{\Omega_1, \ldots, \Omega_{d_k}\}$ of Ω parametrized by $k > 0$. For simplicity, we shall use the generic name of element to denote Ω_i, which will be typically a triangle in 2-D or a tetrahedron in 3-D. We note that our formulation allows for more general shapes. In particular, the outlying elements may have a curved edge or face if Ω is not polygonal.

The formal setting for the velocity will be provided by the (mesh-dependent) "energy" space $\mathbf{E}_k = \mathbf{H}^2(\Omega_1) \times \cdots \times \mathbf{H}^2(\Omega_{d_k})$, where $\mathbf{H}^2$ is the Sobolev space of index 2 (cf. [1]). We may view $\mathbf{E}_k$ as a subspace of $\mathbf{L}^2(\Omega)$. In addition to the L^2 norm, we equip $\mathbf{E}_k$ with the mesh-dependent H^1-like norm

$$\|\mathbf{v}\|_{1,k} = \left\{ \sum_{i=1}^{d_k} \left(\|\nabla \mathbf{v}^{(i)}\|^2_{\partial\Omega_i} + \sum_{j \in \mathcal{N}_i} \tau_{ij} \left[k_i \left| \frac{\partial \mathbf{v}^{(i)}}{\partial n} \right|^2_{\partial\Omega_{i,j}} + k_i^{-1} \left| \mathbf{v}^{(i)} - \mathbf{v}^{(j)} \right|^2_{\partial\Omega_{i,j}} \right] + k_i \left| \frac{\partial \mathbf{v}^{(i)}}{\partial n} \right|^2_{\partial\Omega_i^e} + k_i^{-1} |\mathbf{v}^{(i)}|^2_{\partial\Omega_i^e} \right) \right\}^{1/2}$$

$$\begin{aligned}
k_i &= \text{diameter of } \partial\Omega_i, \\
\partial\Omega_{i,j} &= \partial\Omega_i \cap \partial\Omega_j \quad \text{if } \partial\Omega_i \text{ and } \partial\Omega_j \text{ are adjacent}, \\
\partial\Omega_i^e &= \partial\Omega_i \cap \partial\Omega \\
\mathbf{v}^{(i)} &\quad \text{restriction of } \mathbf{v} \text{ to } \partial\Omega_i, \\
\mathbf{v}^{(i)} - \mathbf{v}^{(j)} &\quad \text{jump in } \mathbf{v} \text{ across } \partial\Omega_{i,j}, \\
\frac{\partial \mathbf{v}^{(i)}}{\partial n} &\quad \text{normal derivative of } \mathbf{v}^{(i)} \text{ with respect to} \\
&\quad \text{the unit outward normal to } \partial\Omega_i \\
\mathcal{N}_i &= \{j : \partial\Omega_j \text{ is adjacent to } \partial\Omega_i\} \\
\tau_{ij} &= 1 \text{ if } i > j \text{ and } \tau_{ij} = 0 \text{ if } i \le j,
\end{aligned}$$

$$(\mathbf{u}, \mathbf{v})_D = \int_D \mathbf{u} \cdot \mathbf{v}\, dx = \sum_{i=1}^{N} \int_D u_i v_i dx, \quad \|\mathbf{v}\|_D = (\mathbf{v}, \mathbf{v})_D^{1/2},$$

$$(\mathbf{u}, \mathbf{v}) = (\mathbf{u}, \mathbf{v})_\Omega,$$

$$< \mathbf{u}, \mathbf{v} >_\Gamma = \int_\Gamma \mathbf{u} \cdot \mathbf{v}\, ds = \sum_{i=1}^{N} \int_\Gamma u_i v_i ds, \text{ edge or surface integrals},$$

$$|\mathbf{v}|_\Gamma = < \mathbf{v}, \mathbf{v} >_\Gamma^{1/2}.$$

To approximate the pressure, we use a partition $\mathcal{T}_h = \{\Omega_1^h, \ldots, \Omega_{d_h}^h\}$ of Ω possibly different from $\mathcal{T}_k$. In order to satisfy the Babuska-Brezzi stability condition establishing the compatibility of the velocity and pressure finite element spaces, we shall assume that $\mathcal{T}_k$ is possibly finer than $\mathcal{T}_h$ in the sense that every element Ω_ℓ^h is a union of members of $\mathcal{T}_k$.

Since the pressure is determined up to an additive constant only, it is convenient to work with quotient spaces X/R obtained by identifying all functions in the space X that differ by constants. Such a space is $L^2(\Omega)/R$. (Note that equivalently one could work with the space $L_0^2(\Omega) = \{q \in L^2(\Omega) : \int_\Omega q\,dx = 0\}$). $L^2(\Omega)/R$ is a Banach space when equipped with the quotient norm $\|q\|_{L^2(\Omega)/R} = \inf_{c \in R} \|q - c\|_{L^2(\Omega)}$. We shall also use the following mesh-dependent and L^2-like norm on the quotient space $H^1(\Omega)/R$

$$\|q\|_{0,h} = \left\{ \|q\|^2_{L^2(\Omega)/R} + \sum_{\ell=1}^{d_h} h_\ell^2 \|\nabla q^{(\ell)}\|^2_{\Omega_\ell^h} \right\}^{1/2},$$

where h_ℓ is the diameter of Ω_ℓ^h.

3 The finite element spaces

The set of vector functions

$$\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} y \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ x \end{pmatrix}, \begin{pmatrix} x \\ -y \end{pmatrix} \right\}$$

forms a basis for the space of linear solenoidal functions in R^2. Augmenting it by

$$\left\{ \begin{pmatrix} y^2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ x^2 \end{pmatrix}, \begin{pmatrix} x^2 \\ -2xy \end{pmatrix}, \begin{pmatrix} -2xy \\ y^2 \end{pmatrix} \right\},$$

gives a basis for quadratics. In R^3, for $r_1 = 2$, a basis is given by the set

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ x \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ x \end{pmatrix}, \begin{pmatrix} y \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ y \end{pmatrix},$$

$$\begin{pmatrix} z \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ z \\ 0 \end{pmatrix}, \begin{pmatrix} x \\ -y \\ 0 \end{pmatrix}, \begin{pmatrix} x \\ 0 \\ -z \end{pmatrix} \right\}.$$

It is typical in constructing finite element spaces to use affine transformations to map "master" basis functions to each element $\partial\Omega_i$. It turns out however that the incompressibility property is not preserved by general

affine transformations. Therefeore, the local basis functions are constructed by translations and scaling of the above functions. We denote the finite element space thus obtained by $\mathbf{V}_k^{r_1}$ where $r_1 - 1$ is the degree of the polynomials used. The fact that the spaces $\mathbf{V}_k^{r_1}$ possess optimal approximations properties is established in [5].

To approximate the pressure, we use spaces $P_h^{r_2}$ of continuous piecewise polynomial functions of degree $r_2 - 1, r_2 \geq 2$ defined on the partition $\mathcal{T}_h$. These spaces are quite standard cf. [8].

4 The discontinuous Galerkin formulations

We begin by defining the bilinear form $a_k^\gamma(\cdot,\cdot) : \mathbf{E}_k \times \mathbf{E}_k \to R$

$$
\begin{aligned}
a_k^\gamma(\mathbf{u},\mathbf{v}) = \sum_{i=1}^{d_k} \Bigg\{ & (\nabla \mathbf{u}^{(i)}, \nabla \mathbf{v}^{(i)})_{\partial\Omega_i} + \sum_{j \in \mathcal{N}_i} \tau_{ij} \Bigg[- \left\langle \frac{\partial \mathbf{u}^{(i)}}{\partial n}, \mathbf{v}^{(i)} - \mathbf{v}^{(j)} \right\rangle_{\partial\Omega_{i,j}} \\
& - \left\langle \frac{\partial \mathbf{v}^{(i)}}{\partial n}, \mathbf{u}^{(i)} - \mathbf{u}^{(j)} \right\rangle_{\partial\Omega_{i,j}} \\
& + \gamma k_i^{-1} \left\langle \mathbf{u}^{(i)} - \mathbf{u}^{(j)}, \mathbf{v}^{(i)} - \mathbf{v}^{(j)} \right\rangle_{\partial\Omega_{i,j}} \Bigg] - \left\langle \frac{\partial \mathbf{u}^{(i)}}{\partial n}, \mathbf{v}^{(i)} \right\rangle_{\partial\Omega_i^e} \\
& - \left\langle \frac{\partial \mathbf{v}^{(i)}}{\partial n}, \mathbf{u}^{(i)} \right\rangle_{\partial\Omega_i^e} + \gamma k_i^{-1} \left\langle \mathbf{u}^{(i)}, \mathbf{v}^{(i)} \right\rangle_{\partial\Omega_i^e} \Bigg\},
\end{aligned}
\tag{4}
$$

which constitutes a weak formulation for the Dirichlet integral $(\nabla \mathbf{u}, \nabla \mathbf{v})$. Indeed, if $\mathbf{u} \in \mathbf{H}^2(\Omega)$, then $\forall \mathbf{v} \in \mathbf{E}_k$

$$
a_k^\gamma(\mathbf{u},\mathbf{v}) = -(\Delta \mathbf{u}, \mathbf{v}) - \sum_{i=1}^{d_k} \left\langle \mathbf{u}^{(i)}, \frac{\partial \mathbf{v}^{(i)}}{\partial n} - \gamma k_i^{-1} \mathbf{v}^{(i)} \right\rangle_{\partial\Omega_i^e}. \tag{5}
$$

Some further comments on the nature of the form a_k^γ are in order

1. The first, second and fifth terms on the right side of (4) are byproducts of integration by parts and range over the interior and boundary edges of $\mathcal{T}_k$ respectively. The array τ_{ij} is used to ensure that each interior edge is visited only once. This device is also convenient as a method for relating the ordering of the edges in a natural way to the ordering of the elements $\partial\Omega_i$.
2. The third and sixth terms have been added to ensure symmetry of the form a_k^γ. Note that the third term is zero for smooth $\mathbf{u}$, while the sixth is a known quantity since $\mathbf{u}\,|_{\partial\Omega}$ is given. We note that the theoretical results remain valid if these terms are removed.
3. The fourth and seventh are so-called "penalty" terms which, upon choosing γ sufficiently large, induce coercivity of the form a_k^γ. The choice of γ is independent of the partition $\mathcal{T}_k$.

The next few results highlight the analysis presented in [5] and [11]. In particular, the role of the penalty parameter γ is exhibited.

Proposition 1. (i)

$$|a_k^\gamma(\mathbf{u},\mathbf{v})| \leq (1+\gamma)\,\|\mathbf{u}\|_{1,k}\;\|\mathbf{v}\|_{1,k}\,, \qquad \forall \mathbf{u},\mathbf{v} \in \mathbf{E}_k.$$

(ii) *There exist positive constants γ_0 and c_a such that for all $\gamma \geq \gamma_0$,*

$$a_k^\gamma(\mathbf{v},\mathbf{v}) \geq c_a \|\mathbf{v}\|_{1,k}^2\,, \qquad \forall \mathbf{v} \in \mathbf{V}_k^{r_1}.$$

The value of γ_0 depends on r_1 but is independent of the meshsize k. Indeed, the bilinear form a_k^γ is singular if γ is small. It is interesting to note that recently a class of related of methods which discard the penalty terms have been proposed cf. [6]. In these methods the bilinear form is made nonsingular by what essentially amounts to changing the sign of the 3rd term on the right side of (4).

Proposition 2. *There exists a constant $c > 0$ such that*

$$(\mathbf{v}, \nabla q) \leq c\,\|\mathbf{v}\|_{1,k}\|q\|_{0,h}\,, \quad \forall \mathbf{v} \in \mathbf{E}_k,\ \forall q \in H^1(\Omega).$$

Proposition 3. *Let $r_1 \geq 1$ and $r_2 \geq 2$ be given. Suppose $\mathcal{T}_k$ is sufficiently fine with respect to $\mathcal{T}_h$. Then, there exists a positive constant c, independent of k and h, such that*

$$\sup_{0 \neq \mathbf{v} \in \mathbf{V}_k^{r_1}} \frac{(\mathbf{v}, \nabla q)}{\|\mathbf{v}\|_{1,k}} \geq c\|q\|_{0,h}\,, \quad \forall q \in P_h^{r_2}. \tag{6}$$

This is the crucial Babuska-Brezzi (inf-sup) condition. Existence and convergence of the numerical approximations depend on it in an essential manner. It is a simple exercise in Linear Algebra to show that if (6) holds then we must necessarily have that $\dim \mathbf{V}_k^{r_1} \geq \dim P_h^{r_2} - 1$. In this sense, using discontinuous elements for the velocity in conjunction with continuous elements for the pressure constitutes a step in the right direction. Indeed, taking $\mathcal{T}_k$ fine with respect to $\mathcal{T}_h$ is a way of increasing the dimension of $\mathbf{V}_k^{r_1}$ with respect to the dimension of $P_h^{r_2}$. By the same token, (6) cannot hold for arbitrary choices of r_1 and r_2 without taking $\mathcal{T}_k$ finer than $\mathcal{T}_h$. However, our numerical experiments, all conducted with $r_1 = r_2$ or $r_1 = r_2 + 1$, suffered no apparent ill effects from taking $\mathcal{T}_k = \mathcal{T}_h$. We conjecture that (6) holds under these conditions.

At this point, we draw attention to similarities between our method and others in the literature. Indeed, in view of the fourth term in (4), our method can be termed as an "interior penalty" formulation. Such methods have been extensively studied in the context of elliptic and other types of problems [3],

[9], [4], [15], [2]. In addition, since the inf-sup property (6) holds for arbitrary choices of r_1 and r_2, provided of course we choose $\mathcal{T}_k$ finer than $\mathcal{T}_h$, relates our method in spirit to so-called stabilized methods cf. [7], [10].

We next construct a Galerkin approximation to the Stokes problem which is the system (1)–(3) minus the convective term $(\mathbf{u}\cdot\nabla)\mathbf{u}$. Multiplying (1) by $\mathbf{v}\in\mathbf{E}_k$ and integrating one obtains after using (5)

$$\nu a_k^\gamma(\mathbf{u},\mathbf{v})+(\mathbf{v},\nabla p)=(\mathbf{f},\mathbf{v})-\nu\sum_{i=1}^{d_k}\left\langle\frac{\partial\mathbf{v}^{(i)}}{\partial n}-\gamma k_i^{-1}\mathbf{v}^{(i)},\mathbf{g}\right\rangle_{\partial\Omega_i^e}.$$

Now multiplying (2) by $q\in H^1(\Omega)$ and integrating by parts, we see that

$$(\mathbf{u},\nabla q)=\langle\mathbf{g}\cdot\mathbf{n},q\rangle_{\partial\Omega}.$$

Combining the last two equations, we derive the following weak formulation for the Stokes problem

$$\nu a_k^\gamma(\mathbf{u},\mathbf{v})+(\mathbf{v},\nabla p)+(\mathbf{u},\nabla q)=F_S([\mathbf{v},q]),\quad\forall[\mathbf{v},q]\in\mathbf{E}_k\times H^1(\Omega),$$

where

$$F_S([\mathbf{v},q])=(\mathbf{f},\mathbf{v})-\nu\sum_{i=1}^{d_k}\left\langle\frac{\partial\mathbf{v}^{(i)}}{\partial n}-\gamma k_i^{-1}\mathbf{v}^{(i)},\mathbf{g}\right\rangle_{\partial\Omega_i^e}+\langle\mathbf{g}\cdot\mathbf{n},q\rangle_{\partial\Omega}.$$

Hence, we define the Galerkin approximation to the Stokes problem as the unique element $[\mathbf{u}_k,p_h]\in\mathbf{V}_k^{r_1}\times P_h^{r_2}/R$ satisfying

$$\nu a_k^\gamma(\mathbf{u}_k,\mathbf{v})+(\mathbf{v},\nabla p_h)+(\mathbf{u}_k,\nabla q)=F_S([\mathbf{v},q]),\quad\forall[\mathbf{v},q]\in\mathbf{V}_k^{r_1}\times P_h^{r_2}/R.$$

To handle the convective term $(\mathbf{u}\cdot\nabla)\mathbf{u}$ we define the trilinear form $b_1(\cdot,\cdot,\cdot):\mathbf{E}_k^3\to R$ by

$$b_1(\mathbf{u},\mathbf{v},\mathbf{w})=\sum_{i=1}^{d_k}\left\{\int_{\partial\Omega_i}u_\ell^{(i)}\frac{\partial v_m^{(i)}}{\partial x_\ell}w_m^{(i)}dx\right.$$
$$\left.-\sum_{j\in\mathcal{N}_i}\tau_{ij}\int_{\partial\Omega_{i,j}}u_\ell^{(i)}(v_m^{(i)}-v_m^{(j)})w_m^{(i)}n_\ell^{(i)}d\sigma\right\},$$

where we have adopted Einstein's summation convention for repeated indices for components of vectors. Following a well-known device of Témam, we introduce the skew-symmetric form

$$b(\mathbf{u},\mathbf{v},\mathbf{w})=\frac{1}{2}\left[b_1(\mathbf{u},\mathbf{v},\mathbf{w})-b_1(\mathbf{u},\mathbf{w},\mathbf{v})\right].$$

Note that we have $b(\mathbf{u},\mathbf{v},\mathbf{v})=0,\quad\forall\mathbf{u},\mathbf{v}\in\mathbf{E}_k$. additionally, the following consistency result holds

Proposition 4. *Suppose* $\mathbf{u}$ *is in* $\mathbf{H}^2(\Omega)$ *and satisfies* $\operatorname{div}\mathbf{u} = \mathbf{0}$ *in* Ω. *Then*

$$b(\mathbf{u},\mathbf{u},\mathbf{v}) = \int_\Omega ((\mathbf{u}\cdot\nabla)\mathbf{u})\mathbf{v}dx - \frac{1}{2}\int_{\partial\Omega}(\mathbf{u}\cdot\mathbf{n})(\mathbf{u}\cdot\mathbf{v})d\sigma, \quad \forall\mathbf{v}\in\mathbf{E}_k.$$

The Galerkin approximation of the stationary Navier-Stokes problem (1)–(3) is defined as the unique solution $[\mathbf{u}_k, p_h] \in \mathbf{V}_k^{r_1} \times P_h^{r_2}/R$ of

$$\begin{aligned}\nu a_k^\gamma(\mathbf{u}_k,\mathbf{v}) + (\mathbf{u}_k,\nabla q) + (\mathbf{v},\nabla p_h) + b(\mathbf{u}_k,\mathbf{u}_k,\mathbf{v}) \\ = F_{NS}[\mathbf{v},q], \quad \forall[\mathbf{v},q]\in\mathbf{V}_k^{r_1}\times P_h^{r_2}/R,\end{aligned}$$

with

$$F_{NS}[\mathbf{v},q] = F_S([\mathbf{v},q]) - \frac{1}{2}\langle\mathbf{g}\cdot\mathbf{n}, \mathbf{g}\cdot\mathbf{v}\rangle_{\partial\Omega}.$$

The convergence of the numerical approximations defined above is analyzed in [5] and [11]. Let $[\mathbf{u}, p]$ denote the solution of either the stationary Stokes or Navier-Stokes problems and assume it to be sufficiently smooth. For simplicity, suppose that $k = h$ and that $r_1 = r_2 = r$. Then, under certain conditions, (cf. [5], [11] for details) the Galerkin approximations $[\mathbf{u_k}, p_h]$ converge to $[\mathbf{u}, p]$ and satisfy the optimal rates

$$\|\mathbf{u} - \mathbf{u_k}\|_{L^2} + h\|\mathbf{u} - \mathbf{u_k}\|_{1,k} + h\|p - p_h\|_{0,h} \le ch^r. \tag{7}$$

5 Numerical results

For considerations of space, we restricted ourselves to the Stokes problem and relegated algorithmic and implementational issues to our forthcoming work [12]. Taking Ω to be the unit square, we imposed on it a regular partition consisting of right isoceles triangles, with the equal sides having length h. Again, we would like to stress the fact that the grids for the velocity and the pressure were identical. To study the errors, we used the function

$$\mathbf{u} = \frac{1}{\pi^2}(\sin\pi(x+y), -\sin\pi(x+y)), \quad p = \frac{1}{\pi^2}\sin\pi(x+y),$$

adjusting $\mathbf{f}$ and $\mathbf{g}$ so that $[\mathbf{u}, p]$ is a solution to (1) and (3).

In Table 1 and Table 2, we exhibit the errors $\|\mathbf{u} - \mathbf{u}_k\|_{L^2}$, $\|\mathbf{u} - \mathbf{u}_k\|_{1,k}$, $\|p - p_h\|_{0,h}$ respectively and the corresponding convergence rates. The first table corresponds to $r_1 = r_2 = 2$ and the second to $r_1 = 3$, $r_2 = 2$. The prevailing values of the Reynolds number $Re = 1/\nu$ and γ are as shown. The rates are seen to conform to theoretical predictions. Note that the rate for the L^2-error for the velocity in Table 1 is slightly larger than the predicted value of 2. Also, the errors for the pressure are the same for both values of r_1.

	$\|\|E(u)\|\|_{L^2}$		$\|\|E(u)\|\|_{1,k}$		$\|\|E(p)\|\|_{0,h}$	
h^{-1}	Error	Rate	Error	Rate	Error	Rate
6	0.254786E-02		0.218224E+00		0.117596E-01	
8	0.969968E-03	3.357	0.154482E+00	1.201	0.645897E-02	2.083
10	0.488330E-03	3.075	0.121078E+00	1.092	0.410444E-02	2.032
12	0.288787E-03	2.881	0.999404E-01	1.052	0.284273E-02	2.015
14	0.189057E-03	2.748	0.852051E-01	1.035	0.208610E-02	2.008
16	0.132241E-03	2.677	0.743051E-01	1.025	0.159624E-02	2.004
18	0.981661E-04	2.530	0.658993E-01	1.019	0.126083E-02	2.003
20	0.755572E-04	2.485	0.592077E-01	1.016	0.102108E-02	2.002
22	0.595668E-04	2.495	0.537598E-01	1.013	0.843775E-03	2.001
24	0.484666E-04	2.370	0.492338E-01	1.011	0.708952E-03	2.001

Table 1. $r_1 = r_2 = 2, Re = 1000, \gamma = 50$

	$\|\|E(u)\|\|_{L^2}$		$\|\|E(u)\|\|_{1,k}$		$\|\|E(p)\|\|_{0,h}$	
h^{-1}	Error	Rate	Error	Rate	Error	Rate
6	0.160837E-02		0.129356E+00		0.117676E-01	
8	0.595667E-03	3.453	0.702981E-01	2.120	0.646089E-02	2.084
10	0.291263E-03	3.206	0.451162E-01	1.988	0.410517E-02	2.032
12	0.165575E-03	3.098	0.315465E-01	1.962	0.284316E-02	2.015
14	0.103424E-03	3.053	0.233018E-01	1.965	0.208636E-02	2.008
16	0.689920E-04	3.032	0.178880E-01	1.980	0.159646E-02	2.004
18	0.490183E-04	2.902	0.140943E-01	2.024	0.126089E-02	2.003
20	0.366955E-04	2.748	0.115363E-01	1.901	0.102111E-02	2.002
22	0.277798E-04	2.920	0.958844E-02	1.940	0.843802E-03	2.001
24	0.214329E-04	2.981	0.807169E-02	1.979	0.709007E-03	2.000

Table 2. $r_1 = 3, r_2 = 2, Re = 1000, \gamma = 50$

A very interesting issue is the dependence of the errors on Re and γ. In Fig. 1 we show the level curves for the errors in terms of these two parameters. In these experiments, we took $h = 1/16$. Figures 1a and 1b correspond to $r_1 = r_2 = 2$ while figures 1c and 1d correspond to $r_1 = 3$, $r_2 = 2$. First, we observe that while the error for the velocity increases with Re, the opposite happens with the pressure. However, the variation in the latter is rather insignificant.

On the other hand, for a fixed value of Re, the shapes of the level curves for the velocity indicate the existence of an optimal value of γ in the sense of minimizing the error. This is in accordance with our experience using the bilinear form (4) in the context of the Dirichlet problem for Laplace's equation. It is also worth noting that, while the bilinear form a_k^γ is no longer positive definite for γ small, the errors increase monotonically to an asymptotic limit as $\gamma \to \infty$.

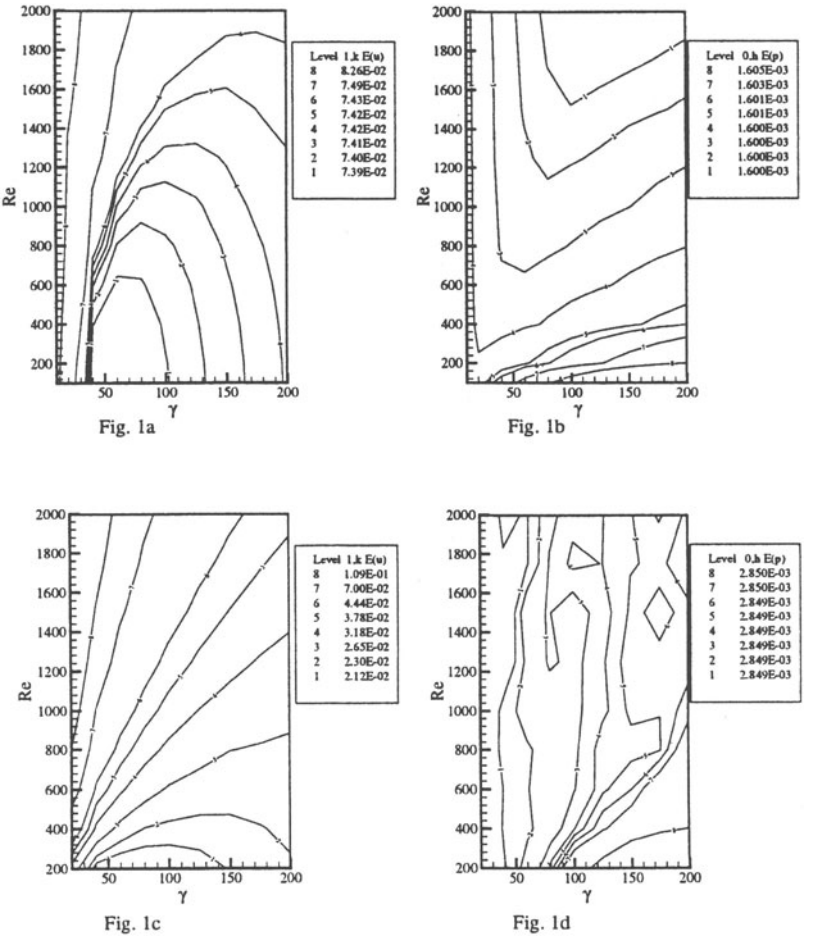


Fig. 1a Fig. 1b

Fig. 1c Fig. 1d

Fig. 1. Level Curves for the errors

References

1. R. Adams, Sobolev Spaces, Academic Press, New York, (1970).
2. D.N. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM J. Num. Anal. **19** (1982), 742–760.
3. I. Babuska, M. Zlamal, *Nonconforming elements in the finite element method with penalty*, SIAM J. Num. Anal. **10** (1973) 863–875.
4. G. Baker, *Finite element methods for elliptic equations using nonconforming elements*, Math. Comp. **31** (1977) 45–59.

5. G. Baker, W. Jureidini, O. Karakashian, *Piecewise solenoidal vector fields and the Stokes problem*, SIAM J. Num. Anal.**27** (1990) 1466–1485.
6. C.E. Baumann, J.T. Oden, *A discontinuous hp finite element method for convection–diffusion problems*, Comput. Methods Appl. Mech. and Engng., In press, special issue on Spectral, Spectral Element and hp methods in CFD, edited by G.E. Karniadakis, M. Ainsworth and C. Bernardi.
7. F. Brezzi, J. Douglas, Jr., *Stabilized mixed methods for the Stokes problem*, Num. Math. **53** (1988) 225–235.
8. P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam (1980).
9. J. Douglas Jr., T. Dupont, *Interior penalty procedures for elliptic and parabolic Galerkin methods*, Lecture Notes in Physics **58** Springer Verlag, Berlin (1976).
10. T.J.R. Hughes, L.P. Franca, *A new finite element formulation for computational fluid dynamics: VII. The Stokes problem with various well-posed boundary conditions: Symmetric formulations that converge for all velocity-pressure spaces*, computer methods in applied mechanics and engineering **65** (1987) 85–96.
11. O. Karakashian, W. Jureidini, *A nonconforming finite element method for the stationary Navier-Stokes equations*, SIAM J. Num. Anal. **35** (1998) 93–120.
12. O. Karakashian, T. Katsaounis, *Numerical simulation of incompressible fluid flow using locally solenoidal elements*, In Preparation.
13. T. Katsaounis, *On Fully discrete Galerkin approximations for the incompressible Navier-Stokes equations*, Ph.D. Thesis, (1994), University of Tennessee, Knoxville, Tennessee.
14. R. Témam, *Navier-Stokes equations, Theory and Numerical Analysis*, North-Holland, Amsterdam (1979).
15. M.F. Wheeler, *An elliptic collocation-finite element method with interior penalties*, SIAM. J. Num. Anal. **15** (1978), 152–161.

Full Convergence for Hyperbolic Finite Elements

Qun Lin

Institute of Systems Science, Academia Sinica,
Beijing 100080, China

Abstract. The error analysis for finite element methods for partial differential equations can be reduced to estimates of a few integral functionals. Some standard estimates obtained by means of a crude application of the Schwarz inequality do not capture the full order of accuracy of the method. By using a suitable integral identity, however, we can capture the full order of convergence when the mesh is nearly uniform and the exact solution is smooth enough. In this paper, we consider first order linear hyperbolic systems and show how to obtain full order of convergence for the standard Galerkin method and the streamline diffusion method; we also show how to obtain superconvergence in the derivative for the discontinuous Galerkin method. Finally, we show how to obtain full order of convergence for the Ying method for nonlinear scalar conservation laws.

1 Introduction

In this paper, we show that a suitable analysis of certain integral functionals can lead to a converge order higher than the one obtained by a simple application of the Schwarz inequality. The main idea is to rewrite the integral functional as the sum of a "small" term and a "big" term whose structure allows for a subtle *cancellation* to take place when the mesh is almost uniform. This results in a much better estimate of the integral functional and leads to full convergence order. Here, we apply this general idea to first order hyperbolic equations; for applications to other partial differential equations, we refer the reader to Lin and Yan [5].

The paper is organized as follows. In section 2, we illustrate our ideas in the one dimensional case for the sake of clarity and simplicity. We show how to obtain full convergence order for the Galerkin and the streamline diffusion methods applied to first-order hyperbolic equations; we also show how to obtain superconvergence in the derivative for the discontinuous Galerkin method. Finally, we show how to obtain the full order of convergence for the so-called Ying method for scalar nonlinear conservation laws; see Ying [8]. In section 3, we sketch the extension of these results to several space dimensions. Finally, in section 4, we end with some concluding remarks.

2 The one dimensional case

2.1 The Integral identity

Sharp estimates for finite element methods for hyperbolic equation depend on a *sharp* estimate for the following integral functional:

$$I(u,v) = \int_{x_0}^{x_n} (u - i_h u)' v,$$

where v is in the space V_h, the standard space of piecewise-linear functions associated with the mesh T_h:

$$x_0 < x_1 < \cdots < x_{i-1} < x_i < \cdots < x_n, \qquad h_i = \frac{1}{2}(x_i - x_{i-1}), \qquad h = \max h_i,$$

and where $i_h u \in V_h$ is the linear interpolate of u at the nodes.

A naïve application of Schwarz inequality gives only the following rough estimate:

$$| I(u,v) | \leq \|u - i_h u\|_1 \|v\|_0 = O(h)\|u\|_2 \|v\|_0.$$

However, we can prove [4] (see also [3]) the *integral identity*

$$\begin{aligned} I(u,v) =& \frac{1}{3}[-h_1^2 (u''v)(x_0) + h_n^2 (u''v)(x_n)] + \frac{1}{3}\sum_{i=1}^{n-1}(h_i^2 - h_{i+1}^2)(u''v)(x_i) \\ &+ \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} u'''(-\frac{1}{3}h_i^2 v + \frac{1}{6}(E^2)'v'), \end{aligned} \tag{1}$$

where $E = \frac{1}{2}[(x - x_{i-\frac{1}{2}})^2 - h_i^2]$. Such an identity gives, when T_h is almost uniform, the *sharp estimate*

$$| I(u,v) | \leq O(h^2)(\|u\|_3 \|v\|_0 + |(u''v)(x_0)| + |(u''v)(x_n)|). \tag{2}$$

In what follows, we show how to apply this inequality to obtain full convergence order estimates for several methods for first order hyperbolic equations.

2.2 The linear hyperbolic equation.

We start by considering the following boundary value problem:

$$u' + u = f \qquad \text{in } (x_0, x_n), \qquad u(x_0) = g.$$

The standard Galerkin method. The approximate solution given by the standard Galerkin method is defined as the function $u_h \in V_h$ satisfying the following weak formulation:

$$B(u_h, v) = \int_{x_0}^{x_n} f\, v, \qquad \forall v \in V_h, \qquad u_h(x_0) = g,$$

where

$$B(u, v) = \int_{x_0}^{x_n} u'v + \int_{x_0}^{x_n} uv.$$

It is easy to see that

$$B(v, v) = \int_{x_0}^{x_n} v'v + \|v\|_0^2 = \frac{1}{2}v(x_n)^2 - \frac{1}{2}v(x_0)^2 + \|v\|_0^2,$$

and that

$$\begin{aligned} B(u_h - i_h u, v) &= B(u - i_h u, v) \\ &= \int_{x_0}^{x_n} (u - i_h u)'v + \int_{x_0}^{x_n} (u - i_h u)v \\ &= I(u - i_h u, v) + \int_{x_0}^{x_n} (u - i_h u)v. \end{aligned}$$

Taking $v = u_h - i_h u$, we have $v(x_0) = 0$ and, making use of the sharp estimate (2), we get

$$\begin{aligned} \frac{1}{2}v(x_n)^2 + \|v\|_0^2 &= B(v, v) \\ &\leq \left| I(u - i_h u, v) \right| + \left| \int_{x_0}^{x_n} (u - i_h u)v \right| \\ &\leq \frac{1}{3}h^2 |u''(x_n)||v(x_n)| + Ch^2 \|u\|_3 \|v\|_0 \\ &\leq C\, h^2 (|u''(x_n)| + \|u\|_3) \sqrt{\frac{1}{2}|v(x_n)|^2 + \|v\|_0^2}. \end{aligned}$$

Hence

$$|v(x_n)| + \|v\|_0 \leq Ch^2 \|u\|_3,$$

and so,

$$\|u_h - u\|_0 \leq Ch^2 \|u\|_3,$$

for almost uniform meshes T_h. Notice that without the integral identity (1) we can only get the rough estimate

$$\|u_h - u\|_0 \leq Ch \|u\|_2.$$

The streamline diffusion method. The approximate solution given by the streamline diffusion method is defined as the solution $u_h \in V_h$ of the following weak formulation:

$$B(u_h, v) = \int_{x_0}^{x_n} f(hv' + v) \; \forall v \in V_h, \qquad u_h(x_0) = g,$$

where

$$B(u, v) = \int_{x_0}^{x_n} (u' + u)(hv' + v).$$

A simple computation gives that

$$B(v, v) = \int_{x_0}^{x_n} (v' + v)(hv' + v) = h\|v\|_1^2 + \|v\|_0^2 + \frac{1+h}{2}[v(x_n)^2 - v(x_0)^2],$$

and that

$$\begin{aligned} B(u_h - i_h u, v) &= B(u - i_h u, v) \\ &= \int_{x_0}^{x_n} ((u - i_h u)' + u - i_h u)(hv' + v) \\ &= h \int_{x_0}^{x_n} (u - i_h u)v' + \int_{x_0}^{x_n} (u - i_h u)'v + \int_{x_0}^{x_n} (u - i_h u)v. \\ &= h\, I(v, u - i_h u) + I(u - i_h u, v) + \int_{x_0}^{x_n} (u - i_h u)v. \end{aligned}$$

Taking T_h to be almost uniform and $v = u_h - i_h u$, we have that, thanks to the sharp estimate (2),

$$\begin{aligned} h\|v\|_1^2 + \|v\|_0^2 + \frac{1+h}{2} v(x_n)^2 = B(v, v) \qquad & \text{since } v(x_0) = 0, \\ \leq & h \left| I(v, u - i_h u) \right| + \left| I(u - i_h u, v) \right| + \left| \int_{x_0}^{x_n} (u - i_h u)v \right| \\ \leq & ch^3 \|u\|_2 \|v\|_1 + \frac{1}{3} h^2 |u''(x_n)| |v(x_n)| + ch^2 \|u\|_3 \|v\|_0 \\ \leq & ch^2 \|u\|_3 \sqrt{h\|v\|_1^2 + \|v\|_0^2 + \frac{1+h}{2} v(x_n)^2} \end{aligned}$$

Hence

$$\sqrt{h}\|v\|_1 + \|v\|_0 + |v(x_n)| \leq Ch^2 \|u\|_3$$

and so,

$$\|u_h - u\|_0 \leq Ch^2 \|u\|_3 \quad \text{and} \quad \|u_h - i_h u\|_0 \leq Ch^{3/2} \|u\|_3.$$

It is also possible to prove that

$$\|u_h^* - u\|_1 \leq Ch^{3/2} \|u\|_3,$$

where u_h^* is given by

$$u_h^* = I_{2h}\, u_h,$$

where I_{2h} is the quadratic interpolation operator on T_{2h}; we assume, of course, that the mesh T_h is obtained from T_{2h} by dividing its intervals by half. See [5] for details.

Without the integral identity (1) we can only get the rough estimates

$$\|u_h - u\|_0 \leq Ch^{3/2}\|u\|_2, \quad \|u_h - u\|_1 \leq Ch\|u\|_2.$$

Remark. The test function in the bilinear form can take a general form:

$$\lambda v' + \mu v, \quad \lambda \geq 0,\ \mu \geq 0.$$

We have then

$$B(v,v) = \lambda\|v\|_1^2 + \mu\|v\|_0^2 + \frac{\lambda+\mu}{2}v(x_n)^2,$$
$$B(u_h - i_h u, v) = (\mu - \lambda)I(u,v) + \mu(u - i_h u, v).$$

When $\lambda = \mu = 1$ or $\lambda = 1, \mu = 0$, we have

$$\|u_h - u\|_0 \leq Ch^2\|u\|_2, \qquad \|u_h - i_h u\|_1 \leq Ch^2\|u\|_2$$

and

$$\|u_h^* - u\|_1 \leq Ch^2\|u\|_2.$$

The discontinuous Galerkin method. This method can be understood more clearly in the $1-d$ case; its extension to several space dimensions is almost straightforward.

The following is the LeSaint - Raviart argument in $1-d$ case.

The approximate solution given by the discontinuous Galerkin method is defined as follows: For $u \in C^1$ and v in

$$W_h = \{w : w(x) = w(x_i^+) \quad \forall x \in (x_i, x_{i+1}), 0 \leq i \leq n-1\},$$

consider the following bilinear form:

$$B(u,v) = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} u'v(x_i^+) + \int_{x_0}^{x_n} uv$$
$$= \sum_{i=0}^{n-1} [u(x_{i+1}) - u(x_i)]v(x_i^+) + \int_{x_0}^{x_n} uv,$$

if $u \in W_h$:

$$B(u,v) = \sum_{i=0}^{n-1} [u(x_i^+) - u(x_i^-)]v(x_i^+) + \int_{x_0}^{x_n} uv.$$

Now, define the approximate solution given by the discontinuous Galerkin method as the function $u_h \in W_h$ that satisfies

$$B(u_h, v) = \int_{x_0}^{x_n} fv \quad \forall v \in W_h, \quad u_h(x_0^-) = g.$$

Then we have

$$\begin{aligned} B(v,v) &= \sum_{i=0}^{n-1}[v(x_i^+) - v(x_i^-)]v(x_i^+) + \|v\|_0^2 \\ &= \frac{1}{2}\sum_{i=0}^{n-1}[v(x_i^+) - v(x_i^-)]^2 + \frac{1}{2}v^2(x_n^-) - \frac{1}{2}v^2(x_0^-) + \|v\|_0^2. \end{aligned}$$

Define $i_h u \in W_h$ as follows:

$$(i_h u)(x) = u(x_{i+1}) \quad \forall x \in (x_i, x_{i+1}), \quad (i_h u)(x_0^-) = g. \tag{3}$$

Then

$$\begin{aligned} B(i_h u, v) &= \sum_{i=1}^{n-1}[(i_h u)(x_i^+) - (i_h u)(x_i^-)]v(x_i^+) + \int_{x_0}^{x_n} i_h u \cdot v \\ &= \sum_{i=1}^{n-1}[u(x_{i+1}) - u(x_i)]v(x_i^+) + \int_{x_0}^{x_n} i_h u \cdot v, \end{aligned}$$

and hence

$$B(u_h - i_h u, v) = B(u, v) - B(i_h u, v) = \int_{x_0}^{x_n} (u - i_h u)v.$$

Taking $v = u_h - i_h u$, we have $v(x_0^-) = 0$ and

$$\frac{1}{2}\sum_{i=0}^{n-1}[v(x_i^+) - v(x_i^-)]^2 + \frac{1}{2}v^2(x_n^-) + \|v\|_0^2 = B(v,v) \le Ch\|u\|_1\|v\|_0.$$

Hence

$$\|v\|_0 \le Ch\|u\|_1, \quad |v(x_n^-)| \le Ch\|u\|_1, \quad \sum_{i=0}^{n-1}[v(x_i^+) - v(x_i^-)]^2 \le Ch\|u\|_1,$$

and so

$$\begin{aligned} &\|u_h - u\|_0 \le Ch\|u\|_1, \quad |u_h(x_i^-) - u(x_i)| \le Ch\|u\|_1, \\ &\Big(\sum_{i=0}^{n-1} h_{i+1}\Big(\frac{u_h(x_{i+1}^-) - u_h(x_i^-)}{h_{i+1}} - \frac{u(x_{i+1}) - u(x_i)}{h_{i+1}}\Big)^2\Big)^{1/2} \le Ch^{\frac{1}{2}}\|u\|_1, \end{aligned} \tag{4}$$

since $v(x_i^+) = u_h(x_{i+1}^-) - u(x_{i+1})$. The last formula provides a a half order superconvergence result for the l_2- estimate of the difference quotient.

Remark. The right-end point interpolation (3) makes $B(u_h - i_h u, v)$ to be minimal. One can take any point interpolation, for example the Gauss point interpolation

$$(i_h u)(x) = u(x_{i+\frac{1}{2}}) \quad \forall x \in (x_i, x_{i+1}).$$

If we want to use LeSaint and Raviart's argument [2] (or to use the Bramble-Hilbert lemma) to prove the first order estimate (4), we must assume that T_h is uniform. Otherwise, we have only $\|u_h - u\|_0 \leq Ch^{1/2}\|u\|_1$.

An interesting phenomenon happens for the equation $u' = f$ approximated by the discontinuous Galerkin solution u_h:

$$u_h = i_h u \quad \text{or} \quad u_h(x_i^-) = u(x_i)$$

Let us recall that the same phenomenon has been proved by Strang and Fix [6] for $u'' = f$ approximated by the linear elements and by Krizek and Neittaanmaki [1] for a fourth order $1-d$ problem approximated by the cubic Hermite elements.

2.3 Nonlinear conservation laws.

Consider the following nonlinear scalar conservation law:

$$u_t + f(u)_x = 0, \qquad f(0) = 0.$$

Finite volume methods and finite element methods for this equation have called the attention of several authors; see the references in [8]. Recently, Xu and Ying [7] studied an explicit upwind FEM, which is monotone and hence first order accurate. Based on this work, Ying [8] constructed a second order scheme and studied its convergence by using our integral identity.

Let us describe Ying's second order scheme. This scheme is constructed by using the following parabolic regularization with a suitable defined artificial viscosity coefficient ε:

$$u_t + f(u)_x = (\varepsilon u_x)_x, \qquad \varepsilon > 0.$$

Xu-Ying's first order scheme:

$$u_j^{n+1} = u_j^n + \frac{\Delta t}{h}[J(u_j^n, u_{j+1}^n) - J(u_{j-1}^n, u_j^n)],$$

$$J(u_j^n, u_{j+1}^n) = \frac{(e^\psi u)_{j+1}^n - (e^\psi u)_j^n}{\int_{jh}^{(j+1)h} \frac{e^\psi}{\varepsilon} dx}, \qquad \psi = -\int_{jh}^{x} \frac{f(u^n)}{\varepsilon u^n} dx,$$

where $\varepsilon = O(h)$ as $\Delta t, h \to 0$.

Ying's second order scheme:

- Near *shock waves*, $|u_x| >> 1$, take $J(u_j^n, u_{j+1}^n)$ as the same as that in 1^{st} order scheme, $\varepsilon = O(h)$;
- On the *smooth region*, take

$$J(u_j^n, u_j^{n+1}) = \varepsilon \frac{u_j^{n+1} - u_j^n}{h} + \frac{1}{h} \int_{jh}^{(j+1)h} f(u^n) dx, \qquad \varepsilon = O(h^2).$$

The scheme can now be described by the following equations:

$$w_j^n = u_j^n + K_j^n, \quad \text{where } K_j^n = \frac{\Delta t}{h}[J(u_j^n, u_{j+1}^n) - J(u_{j-1}^n, u_j^n)],$$

$$u_j^{n+1} = u_j^n + \frac{1}{2}(K_j^n + L_j^n), \quad \text{where } L_j^n = \frac{\Delta t}{h}[J(w_j^n, w_{j+1}^n) - J(w_{j-1}^n, w_j^n)].$$

Then, it can be proven [8] that u_j^n converges in $L^p, p \geq 1$, to the entropy solution. For almost uniform meshes and a sufficiently smooth solution, a second order estimate in L^2 is proved under the stronger condition

$$\Delta t \leq C h^{4/3}.$$

The proof relies in a crucial way on the use of the integral identity (1). Let u be the exact solution and u_h the approximate solution. Set $v^n = i_h u^n - u_h^n$. The main estimate is concerned in the convection term:

$$f(u^n) - f(u_h^n) = f'(u^n)(u^n - u_h^n) + \frac{1}{2} f''(u^n)(u^n - u_h^n)^2 + \frac{1}{6} f'''(u^n - u_h^n)^3,$$

and

$$\begin{aligned} \int f'(u^n)(u^n - u_h^n) v_x^n &= \int f'(u^n)(u^n - i_h u^n) v_x^n + \int f'(u^n) v_n v_x^n \\ &= \int f'(u^n)(u^n - i_h u^n) v_x^n - \frac{1}{2} \int (f'(u^n))_x (v^n)^2 \\ &\leq C h^2 \|u^n\|_3 \|v^n\|_0 + C \|v^n\|_0^2 \end{aligned}$$

which leads to a second order estimate.

3 Extensions to the multi-dimensional case

In applications, we find hyperbolic equations in several space dimensions equations that must be numerically solved with high order. A simple model problem is the following:

$$\beta \cdot \nabla u + u = f \text{ in } \Omega, \qquad \text{where } \frac{1}{2} \nabla \cdot \beta + 1 \geq \theta > 0,$$

for which we have extended the *sharp* estimate introduced in the previous section.

In this case, a *sharp* estimate for finite element approximations depends on *sharp* estimates for a few integral functionals, like

$$\int_\Omega (u - i_h u)_x v, \int_\Omega (u - i_h u)_x v_x, \int_\Omega (u - i_h u)_x v_y,$$

defined on a finite element space V_h associated with the mesh T_h.

When V_h is the standard space of continuous bilinear functions, a simple application of Schwarz inequality gives the following estimates:

$$\int_\Omega (u - i_h u)_x v = O(h)\|u\|_2\|v\|_0, \qquad \int_\Omega (u - i_h u)_x v_y = O(h)\|u\|_2\|v\|_1.$$

However, when T_h is a Cartesian mesh, by using our integral identity, we get

$$\int_\Omega (u - i_h u)_x v_x = O(h^2)\|u\|_3\|v\|_1,$$

$$\int_\Omega (u - i_h u)_x v_y = O(h^2)\|u\|_3(\|v\|_1 + h^{-1}\sqrt{\int_\Gamma v^2|n\beta|}),$$

and if T_h is also almost uniform, we get

$$\int_\Omega (u - i_h u)_x v = O(h^2)\|u\|_3(\|v\|_0 + \sqrt{\int_\Gamma v^2|n\beta|}).$$

The above estimate can be extended to meshes that are smooth mappings of almost uniform Cartesian meshes (called from now on, 'generalized' uniform meshes) as follows:

$$\int_\Omega (u - i_h u)_x v = O(h^{1.5})(\|u\|_3 + \|u\|_{2,\infty})\|v\|_0.$$

Thanks to these *sharp* estimates, we get:

- For the standard Galerkin method:

  $\|u_h - u\|_0 = O(h^2)\|u\|_3$, when T_h is almost uniform,

  $\|u_h - u\|_0 = O(h^{1.5})(\|u\|_3 + \|u\|_{2,\infty})$, when T_h is generalized uniform.

- For the streamline diffusion method:

  $$\|u_h - u\|_0 = O(h^2)\|u\|_3, \quad \text{when } T_h \text{ is almost uniform.}$$

- For the discontinuous Galerkin method:

  $$\|u_h - u\|_0 = O(h)\|u\|_2,$$

and a half order superconvergence for the l_2- estimate of the difference quotient:

$$\sum_e 4h_e k_e (\frac{u^h(P_e) - u^h(Q_e)}{2h_e} - \frac{u(P_e) - u(Q_e)}{2h_e})^2 \leq Ch,$$

$$\sum_e 4h_e k_e (\frac{u^h(P_e) - u^h(R_e)}{2k_e} - \frac{u(P_e) - u(R_e)}{2k_e})^2 \leq Ch,$$

where $e = [x_e - h_e, x_e + h_e] \times [y_e - k_e, y_e + k_e]$, $P_e = (x_e + h_e, y_e + k_e)$, $Q_e = (x_e - h_e, y_e + k_e)$, $R_e = (x_e + h_e, y_e - k_e)$ and $C_1 h \leq h_e, k_e \leq C_2 h$.

Some of the above results also hold for the variable coefficient case $\mathrm{div}(\beta u) + u = f$. For the nonlinear scalar conservation law

$$u_t + \nabla \cdot f(u) = 0, \qquad u(x, 0) = 0,$$

and u_h the approximate solution determined by Ying's scheme, the use of the sharp estimates leads, see [8], to the following error estimate when T_h is almost uniform:

$$\|u_h(t) - u(t)\|_0 \leq Ch^2,$$

$\forall t \in [0, T]$, for arbitrary $T > 0$, provided that we take $\Delta t \leq Ch^{4/3}$.

4 Concluding remarks

We have shown that a careful study of integral functionals can lead to improved rates of convergence, for almost uniform meshes, for finite element methods for linear and nonlinear hyperbolic equations. The main ideas of our approach can also be applied to the error analysis of finite element methods for other types of equations.

References

1. M. Krizek and P. Neittaanmaki, *Finite element approximation of variational problems and applications*, Pitman Monographs and Surveys in Pure and Applied Mathematics, 50, Longman Scientific & Technical, Harlow; copublished in the U.S. with John Wiley & Sons, N.Y., 1990.
2. P. LeSaint and P. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical aspects of finite element in partial differential equations (C. de Boor, ed), Academic Press, N.Y., 1974.
3. Q. Lin and J. Liu, *High accuracy techniques for finite element methods*, Math. in Practice and Theory, 2 (1991), pp. 63–68.
4. Q. Lin and A. Zhou, *A rectangle test for the first order hyperbolic equation*, Proc. Sys. Sci. & Sys., Great Wall Culture Publ. Co., Hong Kong (1991), pp. 234–235.
5. Q. Lin and N. Yan, *Construction and Analysis of Efficient Finite Element Methods*, Hebei Univ. Pub. House (in Chinese), 1996.

6. G. Strang and G. Fix, *An analysis of the finite element method*, Prentice Hall, Englewood Cliffs, NJ, 1973.
7. J. Xu and L.-A. Ying, *Convergence of an explicit upwind finite element method to multi-dimensional conservation laws*, Beijing University Preprint, 1998.
8. L.-A. Ying, *A second-order explicit finite element scheme to multi-dimensional conservation laws and its convergence*, Beijing University Preprint, 1998.

A Conservative DGM for Convection-Diffusion and Navier-Stokes Problems

J. Tinsley Oden[1] and Carlos Erik Baumann[2]

[1] Texas Institute for Computational and Applied Mathematics
[2] Computational Mechanics Company, Inc.

Abstract. An *hp*−adaptive conservative Discontinuous Galerkin Method for the solution of convection-diffusion problems is reviewed. A distinctive feature of this method is the treatment of diffusion terms with a new variational formulation. This new variational formulation is *not* based on mixed formulations, thus having the advantage of not using flux vairables or *extended stencils and/or global matrices' bandwidth* when the flux variables are statically condensed at element level.

The variational formulation for diffusion terms produces a compact, locally conservative, higher-order accurate, and stable solver. The method supports *h*−, *p*−, and *hp*−approximations and can be applied to any type of domain discretization, including non-matching meshes. *A priori* error estimates and numerical experiments indicate that the method is robust and capable of delivering high accuracy.

1 Introduction

In this paper we review a discontinuous Galerkin method for the solution of second order convection-diffusion problems which is based on the classical discontinuous Galerkin approximation for convection terms [9,14,23,24] and on the technique developed in [2,5,31] for diffusion terms. This new methodology supports *h-, p-, and hp*-version approximations and can produce highly accurate solutions.

This discontinuous Galerkin method produces truly compact approximations, only the *nearest* neighbors' degrees of freedom interact with the degrees of freedom of any element. A different approach to the treatment of diffusion terms is to cast the second order problem as a first order system using a mixed formulation. This has been done for the solution of the Navier-Stokes equations by Bassi and Rebay [3,4], and by Lomtev, Quillen and Karniadakis in [25–28], and Warburton, Lomtev, Kirby and Karniadakis in [33]. A generalization of the mixed approach was proposed by Cockburn and Shu under the name of Local Discontinuous Galerkin method [13,18].

A disadvantage of using mixed methods is that for a problem in $\mathbb{R}^d$, for each variable subject to a second-order differential operator, d more variables and equations have to be introduced to obtain a first-order system of equations. In practical applications, however, it is customary to use static condensation of the diffusion fluxes at element level. This strategy removes

the diffusion related unknowns from the global problem, but produces non-compact approximations, the degrees of freedom of a given element have interaction with *non-nearest* neighbors' degrees of freedom. Related properties of the discrete systems of equations and associated bandwith are not favorable for the mixed approach.

The DGM reviewed in this paper has the following features:

- the method does not use auxiliary variables such as those used in mixed formulations;
- the method is robust, stable, and exhibits elementwise conservative approximations;
- the method still produces ***block diagonal*** global mass matrices with ***uncoupled*** blocks;
- the parameters affecting the rate of convergence and the limitations of this method are well established;
- the method is very well suited for adaptive control of error, and can deliver very high orders of accuracy when the exact solution is smooth;
- the cost of solution and implementation is low; and
- implementations of this method on multiprocessor machines can deliver high levels of parallel efficiency.

A detailed overview of the development and use of discontinuous Galerkin methods is presented by Cockburn, Karniadakis and Shu [15]. What follows is a short description of the literature related to our review paper.

The solution of second-order partial differential equations with discontinuous basis functions dates back the the early 1970's, Nitsche [29] introduced the concept of replacing the Lagrange multipliers used in hybrid formulations with averaged normal fluxes at the boundaries, and added stabilization terms to produce optimal convergence rates. A similar approach was used by Percell and Wheeler [32]. A different approach was the Global Element Method (GEM) of Delves and Hall [19], applications of this method were presented by Hendry and Delves in [21]. The GEM consists of the classical hybrid formulation for a Poisson problem with the Lagrange multiplier eliminated in terms of the dependent variables; namely, the Lagrange multiplier is replaced by the average flux across interelement boundaries, without the addition of penalty terms. A major disadvantage of the GEM is that the matrix associated with space discretizations of diffusion operators is indefinite, therefore the method can not be used to solve efficiently time dependent diffusion problems; and being indefinite, the linear systems associated with steady state problems needs special iterative schemes. A study of the eigenvalue spectrum associated with a number of discontinuous techniques for diffusion problems is presented by Hughes *et al.* [22]. In the same paper Hughes *et al.* present new developments in conservation properties of continuous and discontinuous FEMs. The interior penalty formulations presented by Wheeler *et al.* in [32] utilize the bilinear form of the GEM augmented with a penalty term which includes the jumps of the solution across elements. The disadvantages of the

last approach include the dependence of stability and convergence rates on the penalty parameter, the loss of the conservation property at element level, and a bad conditioning of the matrices. The DGM for diffusion operators reviewed in this paper is a modification of the GEM, which is free from the deficiencies that affect the GEM. More details on these formulations, and the relative merits of each one are presented in [5,31].

This paper is structured as follows: Section 2 introduces a model scalar convection-diffusion problem with the associated notation. Section 3 presents the associated discontinuous Galerkin approximation with *a priori* error estimation and Section 4 reviews the application of the method to the Navier-Stokes equations. Finally, numerical experiments are discussed in Section 5 and conclusions are collected in Section 6.

2 Scalar convection-diffusion problem

Let Ω be an open bounded Lipschitz domain in $\mathbb{R}^d$. We consider a model second-order convection diffusion problem characterized by the following scalar partial differential equation and boundary conditions

$$\boxed{-\nabla\cdot(\boldsymbol{A}\nabla u) + \nabla\cdot(\boldsymbol{\beta}\, u) + \sigma u = S \quad \text{in } \Omega \subset \mathbb{R}^d} \tag{1}$$

$$\boxed{\begin{aligned} u &= f \quad \text{on } \Gamma_{\mathrm{D}} \\ (\boldsymbol{A}\nabla u)\cdot\boldsymbol{n} &= g \quad \text{on } \Gamma_{\mathrm{N}} \end{aligned}} \tag{2}$$

where $\boldsymbol{\beta} \in (L^\infty(\Omega))^d$ is the mass flux vector, $\sigma \in L^\infty(\Omega), \sigma \geq 0$ a.e. in Ω, $S \in L^2(\Omega)$, and $\boldsymbol{A} \in (L^\infty(\Omega))^{d\times d}$ is a diffusivity matrix characterized as follows:

$$\begin{gathered} \boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{A}^T(\boldsymbol{x}), \\ \alpha_1 \boldsymbol{a}^T\boldsymbol{a} \geq \boldsymbol{a}^T\boldsymbol{A}(\boldsymbol{x})\boldsymbol{a} \geq \alpha_0 \boldsymbol{a}^T\boldsymbol{a}, \quad \alpha_1 \geq \alpha_0 > 0, \quad \forall \boldsymbol{a} \in \mathbb{R}^d, \end{gathered} \tag{3}$$

a.e. in Ω.

The boundary $\partial\Omega$ consists of disjoint parts, Γ_{D} on which Dirichlet conditions are imposed, and Γ_{N} on which Neumann conditions are imposed: $\Gamma_{\mathrm{D}} \cap \Gamma_{\mathrm{N}} = \emptyset$, $\Gamma_{\mathrm{D}} \cup \Gamma_{\mathrm{N}} = \partial\Omega$, and *meas* $\Gamma_{\mathrm{D}} > 0$. The inflow Γ_- and outflow Γ_+ parts of the boundary are defined as follows:

$$\Gamma_{\mathrm{D}} \supseteq \Gamma_- = \{\boldsymbol{x} \in \partial\Omega \mid (\boldsymbol{\beta}\cdot\boldsymbol{n})(\boldsymbol{x}) < 0 \text{ a.e.}\}, \quad \Gamma_+ = \partial\Omega \setminus \Gamma_-.$$

2.1 Regular partitions

Let us consider regular partitions of Ω [12,30]. Let $\mathcal{P} = \{\mathcal{P}_h(\Omega)\}_{h>0}$ be a family of regular partitions of $\Omega \subset \mathbb{R}^d$ into $N \doteq N(\mathcal{P}_h)$ subdomains Ω_e such

that for $\mathcal{P}_h \in \mathcal{P}$,

$$\overline{\Omega} = \bigcup_{e=1}^{N(\mathcal{P}_h)} \overline{\Omega}_e, \quad \text{and } \Omega_e \cap \Omega_f = \emptyset \quad \text{for } e \neq f.$$

The *interelement boundary* is defined as follows:

$$\Gamma_{\text{int}} = \bigcup_{\Omega_f, \Omega_e \in \mathcal{P}_h} (\partial\Omega_f \cap \partial\Omega_e). \tag{4}$$

On Γ_{int}, we define $\boldsymbol{n} = \boldsymbol{n}_e$ on $(\partial\Omega_e \cap \partial\Omega_f) \subset \Gamma_{\text{int}}$ for indices e, f such that $e > f$.

For $v|_{\Omega_e} \in H^{3/2+\epsilon}(\Omega_e)$ and $v|_{\Omega_f} \in H^{3/2+\epsilon}(\Omega_f)$, we introduce the *jump* operator $[.]$ defined on $\Gamma_{ef} = \overline{\Omega}_e \cap \overline{\Omega}_f \neq \emptyset$ as follows:

$$[v] = (\gamma_0 v)|_{\partial\Omega_e \cap \Gamma_{ef}} - (\gamma_0 v)|_{\partial\Omega_f \cap \Gamma_{ef}}, \qquad e > f, \tag{5}$$

and the *average* operator $\langle . \rangle$ for the normal flux is defined for $(\boldsymbol{A}\boldsymbol{\nabla} v)\cdot\boldsymbol{n} \in L^2(\Gamma_{ef})$, as

$$\langle (\boldsymbol{A}\boldsymbol{\nabla} v)\cdot\boldsymbol{n} \rangle = \frac{1}{2}\left(((\boldsymbol{A}\boldsymbol{\nabla} v)\cdot\boldsymbol{n})|_{\partial\Omega_e \cap \Gamma_{ef}} + ((\boldsymbol{A}\boldsymbol{\nabla} v)\cdot\boldsymbol{n})|_{\partial\Omega_f \cap \Gamma_{ef}} \right), \; e > f \tag{6}$$

where $\boldsymbol{A}$ is the diffusivity. Note that $\boldsymbol{n}$ represents the outward normal of the element with higher index.

3 Discontinuous Galerkin approximation

The discontinuous Galerkin formulation for convection-diffusion problems is built as an extension of the classical discontinuous Galerkin method for hyperbolic problems [16,17,23,24], with the diffusion operators treated as in [5,7,31]. We review definitions and formulations presented in [7].

3.1 Weak formulation

Let $W(\mathcal{P}_h)$ be the Hilbert space on the partition $\mathcal{P}_h$ defined as the completion of $H^{3/2+\epsilon}(\mathcal{P}_h)$ under the norm $\|.\|_W$ defined as follows (induced by (12) below)

$$\|u\|_W^2 = \|u\|_V^2 + \|u\|_\beta^2 + \left\| u\sigma^{\frac{1}{2}} \right\|_{0,\Omega}^2, \tag{7}$$

$$\|v\|_V^2 = \sum_{\Omega_e \in \mathcal{P}_h} \int_{\Omega_e} \boldsymbol{\nabla} v \cdot \boldsymbol{A}\, \boldsymbol{\nabla} v \, \mathrm{d}\mathbf{x} + |v|_{0,\Gamma_{\mathcal{P}_h}}^2, \tag{8}$$

$$\|u\|_\beta^2 = \left| u|\boldsymbol{\beta}|^{\frac{1}{2}} \right|_{0,\Omega}^2 + \sum_{\Omega_e \in \mathcal{P}_h} \left| \boldsymbol{\nabla} u \cdot \boldsymbol{\beta}/|\boldsymbol{\beta}|^{\frac{1}{2}} \right|_{0,\Omega_e}^2 + \left| u\,|\boldsymbol{\beta}\cdot\boldsymbol{n}|^{\frac{1}{2}} \right|_{0,\Gamma_+}^2$$
$$+ \left| h^\alpha u^-\,|\boldsymbol{\beta}\cdot\boldsymbol{n}|^{\frac{1}{2}} \right|_{0,\Gamma_{\text{int}}}^2 + \left| h^{-\alpha}[u]\,|\boldsymbol{\beta}\cdot\boldsymbol{n}|^{\frac{1}{2}} \right|_{0,\Gamma_{\text{int}}}^2, \tag{9}$$

$$|v|^2_{0,\Gamma_{\mathcal{P}_h}} = |h^{-\alpha}\, v|^2_{0,\Gamma_D} + |h^{\alpha}\, (\boldsymbol{A}\boldsymbol{\nabla} v)\cdot\boldsymbol{n}|^2_{0,\Gamma_D} + |h^{-\alpha}\, [v]|^2_{0,\Gamma_{\rm int}} + |h^{\alpha}\, \langle (\boldsymbol{A}\boldsymbol{\nabla} v)\cdot\boldsymbol{n}\rangle|^2_{0,\Gamma_{\rm int}}, \tag{10}$$

and

$$|v|^2_{0,\Gamma} = \int_\Gamma v^2 \, \mathrm{ds}, \quad \text{for} \quad \Gamma \in \{\Gamma_D, \Gamma_N, \Gamma_{\rm int}\}.$$

The terms $h^{\pm\alpha}$, with $\alpha = 1/2$, are introduced to minimize the mesh-dependence of the norm. In (10), the value of h is $h_e/(2\alpha_1)$ on Γ_D, and the average $(h_e + h_f)/(2\alpha_1)$ on that part of $\Gamma_{\rm int}$ shared by two generic elements Ω_e and Ω_f, and the constant α_1 is defined in (3). In (9), however, h is $h_e/2$ on Γ_D, and the average $(h_e + h_f)/2$ on $\partial\Omega_e \cap \partial\Omega_f$.

A consistent formulation of problem (1-2) is the following variational statement:

$$\boxed{\begin{array}{l}\text{Find } u \in W(\mathcal{P}_h) \text{ such that} \\ \qquad\qquad B(u,v) = L(v) \quad \forall\, v \in W(\mathcal{P}_h)\end{array}} \tag{11}$$

where

$$B(u,v) = \sum_{\Omega_e \in \mathcal{P}_h} \left\{ \int_{\Omega_e} [\boldsymbol{\nabla} v \cdot \boldsymbol{A}\, \boldsymbol{\nabla} u - (\boldsymbol{\nabla} v \cdot \boldsymbol{\beta})\, u + v\sigma u]\, \mathrm{dx} + \int_{\partial\Omega_e \setminus \Gamma_-} v u^-\, (\boldsymbol{\beta}\cdot\boldsymbol{n}_e)\, \mathrm{ds} \right\} + \int_{\Gamma_D} ((\boldsymbol{A}\boldsymbol{\nabla} v)\cdot\boldsymbol{n}\, u - v\, (\boldsymbol{A}\boldsymbol{\nabla} u)\cdot\boldsymbol{n})\, \mathrm{ds}$$
$$+ \int_{\Gamma_{\rm int}} (\langle (\boldsymbol{A}\boldsymbol{\nabla} v)\cdot\boldsymbol{n}\rangle [u] - \langle (\boldsymbol{A}\boldsymbol{\nabla} u)\cdot\boldsymbol{n}\rangle [v])\, \mathrm{ds}, \tag{12}$$

$$u^{\pm} = \lim_{\epsilon \to 0} u\,(\boldsymbol{x} \pm \epsilon\boldsymbol{\beta}), \quad \text{for } \boldsymbol{x} \in \Gamma_{\rm int},$$

and

$$L(v) = \sum_{\Omega_e \in \mathcal{P}_h} \int_{\Omega_e} vS\, \mathrm{dx} + \int_{\Gamma_D} (\boldsymbol{A}\boldsymbol{\nabla} v)\cdot\boldsymbol{n}\, f\, \mathrm{ds} + \int_{\Gamma_N} v\, g\, \mathrm{ds} - \int_{\Gamma_-} v f\, (\boldsymbol{\beta}\cdot\boldsymbol{n})\, \mathrm{ds}. \tag{13}$$

Remark: Note that $H^1_0(\Omega) \subset W(\mathcal{P}_h)$. Indeed, for $u, v \in H^1_0(\Omega)$, the bilinear and linear forms $B(u,v)$ and $L(v)$ reduce to those of the continuous Galerkin formulation, which is known to be unstable for not well resolved

convection-dominated problems. The use of discontinuous basis functions in combination with (12)-(13), however, produces a method with superior stability properties. It is proven in [7] that the formulation presented is globally and locally (elementwise) conservative. Section 4.2 in this paper includes a proof of the conservation property for the Navier-Stokes equations. □

3.2 Polynomial approximations on partitions

We review a well-known local approximation property of polynomial finite element approximations (see [1]). Let $\hat{\Omega}$ be a regular master element in $\mathbb{R}^d$, and let $\{F_{\Omega_e}\}$ be a family of invertible maps from $\hat{\Omega}$ onto Ω_e. For every element $\Omega_e \in \mathcal{P}_h$, the finite-dimensional space of real-valued shape functions $\hat{P} \subset H^m\left(\hat{\Omega}\right)$ is the space $P_{p_e}\left(\hat{\Omega}\right)$ of polynomials of degree $\leq p_e$ defined on $\hat{\Omega}$. Then we define

$$P_{p_e}(\Omega_e) = \left\{\psi \mid \psi = \hat{\psi} \circ F_{\Omega_e}^{-1},\ \hat{\psi} \in \hat{P} = P_{p_e}\left(\hat{\Omega}\right)\right\}. \tag{14}$$

Using the spaces $P_{p_e}(\Omega_e)$, we can define

$$W_p(\mathcal{P}_h) = \prod_{e=1}^{N(\mathcal{P}_h)} P_{p_e}(\Omega_e), \tag{15}$$

$N(\mathcal{P}_h)$ being the number of elements in $\mathcal{P}_h$.

The approximation properties of $W_p(\mathcal{P}_h)$ will be estimated using standard local approximation estimates (see [1]). Let $u \in H^s(\Omega_e)$; there exist a constant C depending on s and on the angle condition of Ω_e, but independent of u, $h_e = diam(\Omega_e)$, and p_e, and a polynomial u_p of degree p_e, such that for any $0 \leq r \leq s$ the following estimate hold:

$$\|u - u_p\|_{r,\Omega_e} \leq C\frac{h_e^{\mu-r}}{p_e^{s-r}}\|u\|_{s,\Omega_e}, \qquad s \geq 0, \tag{16}$$

where $\|.\|_{r,\Omega_e}$ denotes the usual Sobolev norm, and $\mu = min(p_e + 1, s)$.

3.3 Discontinuous Galerkin approximation

The variational formulation of our discontinuous Galerkin method (11) will be used as a basis to construct approximations to the exact solution in a finite dimensional space. The variational formulation in the space $W_p(\mathcal{P}_h)$ is the following:

$$\boxed{\begin{array}{l}\text{Find } u_{DG} \in W_p(\mathcal{P}_h) \text{ such that} \\ \qquad\qquad B(u_{DG}, v_h) = L(v_h)\ \forall\, v_h \in W_p(\mathcal{P}_h)\end{array}} \tag{17}$$

where $B(.,.)$ and $L(.)$ are defined in (12) and (13), respectively.

Note that all the properties of the discontinuous Galerkin method (11) also hold for the finite dimensional approximation (17); namely, solutions are elementwise conservative, mass matrices are block diagonal, and the space of dicontinuous functions provides the basis to obtain solutions with potentially good stability properties.

Stability is one of the most important characteristics of a method for the solution of convection-diffusion problems. The following section addresses this issue and provides *a priori* error estimation to solutions of (17).

3.4 *A priori* error estimation

***A priori* error estimates for pure diffusion cases:** We now review results presented in [31]. Assuming that

$$\inf_{\substack{u \in V_h \\ \|u\|_{V_h} = 1}} \sup_{\substack{v \in V_h \\ \|v\|_{V_h} \leq 1}} |B(u,v)| \geq \gamma_h \approx \mathrm{O}\left(p_{\max}^{-\kappa}\right), \tag{18}$$

where $\|.\|_{V_h}$ is the norm defined in (8), and $\kappa \geq 0$, an estimate of the global rates of convergence of the DGM (13) with $\beta \equiv 0$ and $\sigma \equiv 0$ is given by the following theorem:

Theorem 1. *Let the solution $u \in H^s(\mathcal{P}_h(\Omega))$, with $s > 3/2$, and assume that the value of the inf-sup parameter is $\gamma_h = C_p\, p_{\max}^{-\kappa}$ with $\kappa \geq 0$. If the approximation estimate (16) hold for the spaces $W_p(\mathcal{P}_h)$, then the error of the approximate solution u_{DG} can be bounded as follows:*

$$\|u - u_{DG}\|_V^2 \leq C p_{\max}^{2\kappa} \sum_{\Omega_e \in \mathcal{P}_h} \left(\frac{h_e^{\mu_e - 1 - \epsilon}}{p_e^{s-3/2-\epsilon}} \|u\|_{s,\Omega_e} \right)^2, \tag{19}$$

where $\mu_e = min(p_e + 1, s)$, $\epsilon \longrightarrow 0^+$, and the constant C depends on s and on the angle condition of Ω_e, but it is independent of u, h_e and p_e.

Remark: The error estimate (19) is a bound for the worst possible case, including all possible data. For a wide range of data, however, the error estimate (19) may be pessimistic, and the actual rate of convergence can be larger than that suggested by the above bound.

The value of the parameter κ depends on p_e and on d. For $d = 1$, $\kappa = 0$ regardless of p_e, as long as $p_e \geq 2$, whereas for $d > 1$ the value depends on the mesh regularity; numerical evaluation of the inf-sup condition suggests that for $p_e \geq 2$ the exponent $\kappa \approx 1.0 - 1.5$. □

Riviere and Wheeler [10,11] have presented the following results related to the discontinuous Galerkin method reviewed in this paper:

Theorem 2. *Let the solution $u \in H^s(\mathcal{P}_h(\Omega))$, with $s > 3/2$, then the error of the approximate solution u_{DG} can be bounded as follows:*

$$\left(|u - u_{DG}|_1^2 + \left\|\sigma^{1/2}(u - u_{DG})\right\|_1^2\right)^{1/2} \leq C\frac{h^{\mu-1}}{p^{s-3/2}}\left(\sum_{\Omega_e \in \mathcal{P}_h} \|u\|_{s,\Omega_e}^2\right)^{1/2} \quad (20)$$

where $\mu = min(p+1, s)$, and the constant C depends on α_1, and $\|\sigma\|_\infty$. This estimate was proven along with some other estimates related to the Non-Symmetric Interior Penalty Galerkin (NIPG) method of Wheeler and Riviere.

A priori **error estimates for convection-diffusion cases:** In this section we review results presented in [7]. The norm $\|.\|_{W_1}$ used in the error estimate is the following:

$$\|u\|_{W_1}^2 = \|u\|_{V_1}^2 + \|u\|_{\beta_1}^2, \quad (21)$$

$$\|u\|_{V_1}^2 = |u|_{0,\mathcal{P}_h}^2 + |h^\alpha u|_{0,\Gamma_D \cup \Gamma_N}^2 + \left|h^\delta \alpha_1^{-1} (\boldsymbol{A}\boldsymbol{\nabla} u)\cdot \boldsymbol{n}\right|_{0,\Gamma_D}^2$$
$$+ |h^\alpha [u]|_{0,\Gamma_{\text{int}}}^2 + \left|h^\delta \alpha_1^{-1} \langle (\boldsymbol{A}\boldsymbol{\nabla} u)\cdot \boldsymbol{n}\rangle\right|_{0,\Gamma_{\text{int}}}^2 + |h^\alpha \langle u\rangle|_{0,\Gamma_{\text{int}}}^2, \quad (22)$$

$$|u|_{0,\mathcal{P}_h}^2 = \sum_{\Omega_e \in \mathcal{P}_h} \int_{\Omega_e} u^2 \, \mathrm{dx}, \; |v|_{0,\Gamma}^2 = \int_\Gamma v^2 \, \mathrm{ds}, \text{ for } \Gamma \in \{\Gamma_D, \Gamma_N, \Gamma_{\text{int}}\},$$

and

$$\|u\|_{\beta_1}^2 = |h^\alpha u^+|_{0,\Gamma_-}^2 + |h^\alpha [u]|_{0,\Gamma_{\text{int}}}^2 + \|h u_\beta\|_{0,\Omega}^2, \quad (23)$$

with $\alpha = 1/2$, $\delta = 3/2$, and $u_\beta = |\boldsymbol{\beta}|^{-1}(\boldsymbol{\nabla} u \cdot \boldsymbol{\beta})$ when $|\boldsymbol{\beta}| > 0$, otherwise $u_\beta = 0$. In $\|.\|_{V_j}$ and $\|.\|_{\beta_j}$ the scaling parameter h is $h_e/2$ on $\partial\Omega_e \cap \partial\Omega$, and the average $(h_e + h_f)/2$ on $\partial\Omega_e \cap \partial\Omega_f$.

Solutions to convection-diffusion problems can exhibit features that range from those of diffusion dominated problems to those of pure convection problems. The error of diffusion dominated problems is better measured in the H^1-norm because the associated physics depends on the solution gradient, such as heat transfer, viscous stresses, etc; whereas the error in convection dominated transport is better measured in the L^2-norm, because the underlying physics depends more on the solution values than on the solution gradient.

The range $[0, 1)$ of local Peclet numbers (P_e) represents a situation in which diffusion effects are dominant, and for which the W-norm converges to the V-norm as $P_e \longrightarrow 0$. The analysis of stability in the V-norm for diffusion

dominated problems was presented in [5,31], where optimal h-convergence rates are presented.

For high P_e, where convection is important, the following *a priori* error estimate applies (see [7]). We assume that the reaction coefficient $\sigma = 0$, which is the worst case scenario from the point of view of stability.

Theorem 3. *Let the solution to (11) be* $u \in H^s(\mathcal{P}_h(\Omega))$, *with* $s > 3/2$, *and assume that there exists* $\kappa \geq 0$ *and* $C_p > 0$ *such that*

$$\inf_{\substack{u \in W_1 \\ \|u\|_{W_1} = 1}} \sup_{\substack{v \in W_1 \\ \|v\|_{W_1} \leq 1}} |B(u,v)| \geq C_p \, p_{\max}^{-\kappa}, \tag{24}$$

where $p_{\max} = \max_e(p_e)$. *If the approximation estimate (16) holds for the spaces* $W_p(\mathcal{P}_h)$, *then the error of the approximate solution* u_{DG} *is bounded as follows:*

$$\|u - u_{DG}\|_{W_1}^2 \leq C p_{\max}^{2\kappa} \sum_{\Omega_e \in \mathcal{P}_h} \left(\frac{h_e^{\mu_e - \epsilon}}{p_e^{s-3/2-\epsilon}} \|u\|_{s,\Omega_e} \right)^2, \tag{25}$$

where $\mu_e = min(p_e + 1, s)$, $\epsilon \longrightarrow 0^+$, *and the constant* C *depends on* s *and on the angle condition of the element, but it is independent of* u, h_e *and* p_e.

A proof of this result can be found in [7]. **Remark:** Numerical experiments presented in [7] indicate that $\kappa < 1.5$. □

4 Navier-Stokes problems

First we review a model problem and related notations in preparation for analyzing the discontinuous Galerkin formulation.

Let Ω be a bounded Lipschitz domain in $\mathbb{R}^d$.The governing equations for the conservation of mass, momentum, and energy can be written in vector form as follows:

$$\boxed{\begin{aligned} \frac{\partial \boldsymbol{U}}{\partial t} + \frac{\partial \boldsymbol{F}_i}{\partial x_i} &= \frac{\partial \boldsymbol{F}_i^v}{\partial x_i} + \boldsymbol{S}, \quad \text{in } \Omega \\ \boldsymbol{U}(\boldsymbol{x}, 0) &= \boldsymbol{U}_o(\boldsymbol{x}), \quad \text{at } t = 0 \end{aligned}} \tag{26}$$

where repeated indices are summed throughout their range, $\boldsymbol{U} = (u_1, ..., u_m) = \boldsymbol{U}(\boldsymbol{x}, t) \in \mathbb{R}^m$ is a vector of conservation variables with $m = d + 1$ (momentum and mass conservation) or $d + 2$ when the energy equation is included, $\boldsymbol{F}_i(\boldsymbol{U}) = (f_{1i}, ..., f_{mi}) \in \mathbb{R}^m$ and $\boldsymbol{F}_i^v(\boldsymbol{U}) = (f_{1i}^v, ..., f_{mi}^v) \in \mathbb{R}^m$ are the inviscid and diffusive flux vectors associated with the i$-th$ space coordinate, and $\boldsymbol{S}$ represents the body forces in the momentum equations and a source of heat (e.g. heat source due to viscous dissipation) in the energy equation. The system of equations (26) is accompanied by appropriate boundary conditions for each problem.

4.1 Inviscid and viscous flux vectors

The inviscid flux vectors $\boldsymbol{F}_i$ are homogeneous functions of degree one in the conservation variables $\boldsymbol{U}$; therefore the fluxes can be written as $\boldsymbol{F}_i = \boldsymbol{A}_i(\boldsymbol{U})\,\boldsymbol{U}$, where $\boldsymbol{A}_i(\boldsymbol{U})$ is the Jacobian matrix.

Let $\boldsymbol{F}_n(\boldsymbol{U})$ be the normal flux at any point on a boundary $\partial\Omega$ with outward normal $\boldsymbol{n}$; then

$$\boldsymbol{F}_n = \boldsymbol{F}_i\, n_i \qquad i \in [1, ..., d],$$

$$\boldsymbol{A}_n(\boldsymbol{U}) = \frac{\partial \boldsymbol{F}_n(\boldsymbol{U})}{\partial \boldsymbol{U}} = \boldsymbol{A}_i(\boldsymbol{U})\, n_i, \qquad \boldsymbol{A}_n(\boldsymbol{U}) \in \mathbb{R}^m \times \mathbb{R}^m.$$

The flux vector $\boldsymbol{F}_n(\boldsymbol{U})$ can be split into inflow and outflow components $\boldsymbol{F}_n^+$ and $\boldsymbol{F}_n^-$; for example

$$\boldsymbol{F}_n^{\pm}(\boldsymbol{U}) = \boldsymbol{R}\,\boldsymbol{\Lambda}^{\pm}\boldsymbol{R}^{-1}\,\boldsymbol{U}, \qquad \boldsymbol{\Lambda}^{\pm} = \frac{1}{2}\left(\boldsymbol{\Lambda} \pm |\boldsymbol{\Lambda}|\right),$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues of $\boldsymbol{A}_n$, and the columns of the matrix $\boldsymbol{R}$ are the corresponding eigenvectors. From a physical point of view, $\boldsymbol{F}_n^+$ and $\boldsymbol{F}_n^-$ represent the fluxes of mass, momentum, and energy leaving $(+)$ and entering $(-)$ the domain through $\partial\Omega$.

Given that the approximation of field variables may be discontinuous across internal surfaces in Ω or across $\partial\Omega$, let us define

$$\boldsymbol{U}^{\pm} = \lim_{\epsilon \to 0^+} \boldsymbol{U}(\boldsymbol{x} \pm \epsilon\, \boldsymbol{n}),$$

where $\boldsymbol{x}$ is a point at a boundary which can be real (e.g. bounding walls) or artificial (e.g. interelement, far-field). With this notation, $\boldsymbol{F}_n^+(\boldsymbol{U}^-)$ is the flux in the direction $\boldsymbol{n}$, and $\boldsymbol{F}_n^-(\boldsymbol{U}^+)$ is that in the opposite direction.

The projection of the viscous flux vectors $\boldsymbol{F}_i^v$ onto the normal $\boldsymbol{n}$ to a boundary is a linear functional of $\boldsymbol{U}$, and will be written in the following alternative forms

$$\boldsymbol{F}_i^v\, n_i = \boldsymbol{F}_n^v = \boldsymbol{D}_n\, \boldsymbol{U},$$

where for Newtonian flows the matrix $\boldsymbol{D}_n$ is a linear differential operator.

4.2 Space discretization with broken spaces

Let $\boldsymbol{V}(\mathcal{P}_h)$ be a broken space of admissible vectors of conservation variables $\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{x})$ having the necessary regularity conditions.

For a given initial data $\boldsymbol{U}_o$, and appropriate boundary conditions, the *space discretization* using the discontinuous Galerkin method can be stated

as follows:

$$
\begin{aligned}
&\text{Given } \boldsymbol{U}_o = \boldsymbol{U}_o(\boldsymbol{x}), \text{ for } t \in (0,T), \text{ find} \\
&\boldsymbol{U}(.,t) \in \boldsymbol{V}(\mathcal{P}_h) \times H^1(0,T) \text{ such that } \boldsymbol{U}(\boldsymbol{x},0) = \boldsymbol{U}_o(\boldsymbol{x}), \text{ and} \\
&\int_\Omega \boldsymbol{W}^T \frac{\partial \boldsymbol{U}}{\partial t}\, \mathrm{dx} + \sum_{\Omega_e \in \mathcal{P}_h} \int_{\partial\Omega_e} \boldsymbol{W}^T \left(\boldsymbol{F}^+_{n_e}(\boldsymbol{U}^-) + \boldsymbol{F}^-_{n_e}(\boldsymbol{U}^+)\right) \mathrm{ds} \\
&\qquad + \int_{\Gamma_{\mathrm{int}}} \left(\left\langle \boldsymbol{W}^T \boldsymbol{D}_n^T \right\rangle [\boldsymbol{U}] - \left[\boldsymbol{W}^T\right] \langle \boldsymbol{F}^v_n \rangle\right) \mathrm{ds} \\
&+ \int_{\Gamma_{\mathrm{D}}} \left(\boldsymbol{W}^T \boldsymbol{D}_n^T \boldsymbol{U} - \boldsymbol{W}^T \boldsymbol{F}^v_n\right) \mathrm{ds} + \sum_{\Omega_e \in \mathcal{P}_h} \int_{\partial\Omega_e} \frac{\partial \boldsymbol{W}^T}{\partial x_i} \left(\boldsymbol{F}^v_i - \boldsymbol{F}_i\right) \mathrm{dx} \\
&\qquad = \int_\Omega \boldsymbol{W}^T \boldsymbol{S}\, \mathrm{dx} + \int_{\Gamma_{\mathrm{D}}} \boldsymbol{W}^T \boldsymbol{D}_n^T \hat{\boldsymbol{U}}\, \mathrm{ds} + \int_{\Gamma_{\mathrm{N}}} \boldsymbol{W}^T \hat{\boldsymbol{F}}^v_n\, \mathrm{ds} \\
&\qquad \forall\, \boldsymbol{W} \in \boldsymbol{V}(\mathcal{P}_h)
\end{aligned}
\tag{27}
$$

here

$$
\boldsymbol{F}_{n_e}(\boldsymbol{U}) = \boldsymbol{F}_i(\boldsymbol{U})\, n_{e_i}, \qquad \boldsymbol{U}^\pm = \lim_{\epsilon \to 0^+} \boldsymbol{U}(\boldsymbol{x} \pm \epsilon\, \boldsymbol{n}_e),
$$

$$
\boldsymbol{F}^v_n(\boldsymbol{U}) = \boldsymbol{F}^v_i(\boldsymbol{U})\, n_i, \qquad \boldsymbol{F}^v_n(\boldsymbol{U}) = \boldsymbol{D}_n \boldsymbol{U}.
$$

$\boldsymbol{F}^\pm_n$ are known in closed form for the usual flux vector and flux difference splittings (see [5] and references therein).

It is important to observe that (27) reduces to the classical weak Galerkin approximation if we restrict $\boldsymbol{V}(\mathcal{P}_h)$ to a space of continuous functions.

We now prove that (27) renders a *conservative* formulation. To show that (27) is *globally conservative,* let us pick a test function $\boldsymbol{W} = (v_1, ..., v_m)$ such that

$$
v_i(\boldsymbol{x}) = 1, \qquad i = 1, ..., m \qquad \forall \boldsymbol{x} \in \Omega,
$$

by definition $\boldsymbol{W} \in \boldsymbol{V}(\mathcal{P}_h)$. Substituting $\boldsymbol{W}$ in (27), we get

$$
\begin{aligned}
&\int_\Omega \frac{\partial \boldsymbol{U}}{\partial t}\, \mathrm{dx} + \sum_{\Omega_e \in \mathcal{P}_h} \int_{\partial\Omega_e} \left(\boldsymbol{F}^+_{n_e}(\boldsymbol{U}^-) + \boldsymbol{F}^-_{n_e}(\boldsymbol{U}^+)\right) \mathrm{ds} - \int_{\Gamma_{\mathrm{D}}} \boldsymbol{F}^v_n\, \mathrm{ds} \\
&\qquad = \int_\Omega \boldsymbol{S}\, \mathrm{dx} + \int_{\Gamma_{\mathrm{N}}} \hat{\boldsymbol{F}}^v_n\, \mathrm{ds}.
\end{aligned}
\tag{28}
$$

For any pair of adjoining elements (Ω_e, Ω_f), the following identities hold:

$$
\boldsymbol{F}^+_{n_e}(\boldsymbol{U}^-) = -\boldsymbol{F}^-_{n_f}(\boldsymbol{U}^+), \quad \text{and} \quad \boldsymbol{F}^-_{n_e}(\boldsymbol{U}^+) = -\boldsymbol{F}^+_{n_f}(\boldsymbol{U}^-),
$$

substituting the above identities in (28), we obtain

$$\int_\Omega \frac{\partial \boldsymbol{U}}{\partial t}\,\mathrm{dx} + \int_{\partial\Omega} \left(\boldsymbol{F}_n^+(\boldsymbol{U}^-) + \boldsymbol{F}_n^-(\boldsymbol{U}^+)\right)\,\mathrm{ds} = \int_\Omega \boldsymbol{S}\,\mathrm{dx} + \int_{\Gamma_\mathrm{N}} \hat{\boldsymbol{F}}_n^v\,\mathrm{ds} + \int_{\Gamma_\mathrm{D}} \boldsymbol{F}_n^v\,\mathrm{ds}, \tag{29}$$

which shows that the formulation is globally conservative.

To show that the formulation is also *locally conservative,* we select a generic weighting function

$$\boldsymbol{W} = (v_1, ..., v_m) \in \boldsymbol{V}(\mathcal{P}_h) \quad \text{such that } v_i(\boldsymbol{x}) = \begin{cases} 1 & \boldsymbol{x} \in \Omega_e \\ 0 & \boldsymbol{x} \notin \Omega_e \end{cases} \qquad i = 1, ..., m,$$

and substituting $\boldsymbol{W}$ in (27), we get

$$\int_{\Omega_e} \frac{\partial \boldsymbol{U}}{\partial t}\,\mathrm{dx} + \int_{\partial\Omega_e} \left(\boldsymbol{F}_{n_e}^+(\boldsymbol{U}^-) + \boldsymbol{F}_{n_e}^-(\boldsymbol{U}^+)\right)\,\mathrm{ds} = \int_{\Omega_e} \boldsymbol{S}\,\mathrm{dx} + \int_{\partial\Omega_e\cap\Gamma_\mathrm{N}} \hat{\boldsymbol{F}}_n^v + \int_{\partial\Omega_e\cap\Gamma_\mathrm{D}} \boldsymbol{F}_n^v\,\mathrm{ds} + \int_{\partial\Omega_e\cap\Gamma_\mathrm{int}} \langle \boldsymbol{F}_n^v \rangle\,\mathrm{ds}, \tag{30}$$

which represents the conservation equations at *element* level when the interelement viscous forces are taken as the average $\langle \boldsymbol{F}_n^v \rangle$.

Remark: To insure local conservation at element level, the interelement viscous flux $F^v(\boldsymbol{U}^-, \boldsymbol{U}^+)$ shoud have the following property:

$$\text{if } \frac{\partial \boldsymbol{U}^-}{\partial n} = \frac{\partial \boldsymbol{U}^+}{\partial n}, \quad \text{then} \quad F^v(\boldsymbol{U}^-, \boldsymbol{U}^+) = \boldsymbol{F}_n^v(\boldsymbol{U}^-) = \boldsymbol{F}_n^v(\boldsymbol{U}^+). \tag{31}$$

This property is verified by the interelement viscous flux $F^v(\boldsymbol{U}^-, \boldsymbol{U}^+) = \langle \boldsymbol{F}_n^v \rangle$ used in (27), see (30). Formulations that add *stabilizing* terms based on the jump of the solution $[\boldsymbol{U}] = (\boldsymbol{U}^+ - \boldsymbol{U}^-)$ do not possess property (31), because in general, approximate solutions have non-zero jump across interelement boundaries. □

5 Numerical tests

Numerical verification of convergence rates in the $\|.\|_V$ norm for diffusion problems and in the $\|.\|_{W_1}$ norm for convection diffusion problems have been presented in [31,7].

Next we present solutions to Hemker's problem, a convection-diffusion problem with a turning point in the middle of the domain:

$$\begin{cases} \alpha \dfrac{\partial^2 u}{\partial x^2} + x \dfrac{\partial u}{\partial x} = -\alpha\pi^2 \cos(\pi x) - \pi x \sin(\pi x) & \text{on } [0,1] \\ u(-1) = -2, \quad u(1) = 0 \end{cases} \tag{32}$$

for which the exact solution is

$$u(x) = \cos(\pi x) + \mathrm{erf}(x/\sqrt{2\alpha})/\mathrm{erf}(1/\sqrt{2\alpha}).$$

Figure 1 and Fig. 2 show the solutions to the above problem ($\alpha = 10^{-10}$ and $h = 1/10$) obtained with continuous and discontinuous Galerkin, respectively.

Numerical experiments show that when the best approximation to the exact solution in the L^2-norm is better represented using discontinuous functions, the DG method performs better than the continuous Galerkin method. This is in general true for solutions to convection dominated problems with boundary layers. If the quantity of interest is the gradient of the solution at the boundary layers, the DG method provides an excellent error indicator, namely, Dirichlet boundary conditions are not well approximated. This error indicator allows to capture steep gradients using adaptive refinement.

Solutions to the incompressible Navier-Stokes equations can be obtained using the artificial compressibility technique. The test case selected is a popular benchmark for laminar viscous flows, the driven cavity problem described in [20] with *Rey* 7500.

A solution to this problem is obtained with a mesh of quadratic elements which is equivalent (in number of degrees of freedom) to a mesh of 60x60 linear elements (used for plotting results) as shown in Fig. 3, this figure also shows the pressure distribution on the background. Note that the pressure range shown is much smaller than the actual range, which is very wide because of the presence of singularities at the top corners of the cavity. The cutoff values $[p_{min}, p_{max}]$ applied to the range of pressure allow to observe small changes within the domain, excluding the areas adjacent to the top corners. Figure 4 shows the streamline pattern. This solution is very accurate, comparisons of velocity profiles through horizontal and vertical planes can be found in [6]. Summarizing, the numerical experiments confirm the stability and high accuracy that the method can deliver for the class of problems considered, even with the use of h refinements and p enrichments.

6 Conclusions

The salient properties of the discontinuous Galerkin formulation reviewed in this paper can be summarized as follows:

- diffusion terms are discretized with a variational formulation that is *not* based on mixed formulations, this is very advantageous because extra flux vectors (gradient of each scalar variable) are not required. When used, the flux vectors of mixed formulations increase considerably the bandwidth of the systems (when they are statically condensed at element level), and the formulations are not compact, the degrees of freedom of any given element interact with the degrees of freedom of non-nearest neighboors;

- the method is capable of solving convection-diffusion problems with an $hp-$approximation methodology. If the local regularity of the solution is high, the $p-$approximation can be used and the method delivers very high accuracy; otherwise the $h-$approximation can be used and the error is reduced by local refinement of the mesh;
- approximate solutions of unresolved flows (e.g. boundary layers) do not suffer from widespread oscillations, for these cases the treatment of interelement boundaries prevents the appearance and spreading of numerical oscillations;
- stability studies and numerical tests demonstrate cuantitatively and qualitatively the superiority of discontinuous over continuous Galerkin solutions for convection-diffusion problems;
- discrete representations are stable in the sense that the real part of the eigenvalues associated with the space discretization are strictly negative, a property that allows the use of time marching schemes for time-dependent problems and also allows to solve steady state problems with explicit schemes;
- an *a priori* error estimate for high Pe numbers indicates that the method delivers optimal $h-$convergence rate (accuracy); and
- contrasting other techniques that use artificial diffusion to improve the stability of continuous Galerkin approximations, the present discontinuous Galerkin method does not introduce such mesh-size dependent terms in the governing equations, which allows for approximation with unlimited order of accuracy; namely, the order of accuracy grows linearly with the order p of the basis functions.

The structure of this discontinuous Galerkin method, particularly the fact that the degrees of freedom of an individual element are coupled only with those of neighbors sharing a boundary makes this method easily parallelizable.

Stability analysis and *a priori* error estimates have been presented for the scalar case in [7,31]. Numerical evidence presented in [7,6,8,31] suggests that this discontinuous Galerkin formulation is highly reliable for obtaining numerical solutions to problems characterized by a wide range of fluid flow conditions. Remarkably, this formulation is stable even when the flow field is not well resolved, and does not produce the classical oscillations near sharp gradients (e.g. boundary layers) which are characteristic in classical H^1 approximations of under-resolved boundary layers.

Acknowledgement

The support of this work by the Army Research Office under grant DAAH04-96-0062 is gratefully acknowledged.

References

1. I. Babuska and M. Suri. The hp-version of the finite element method with quasiuniform meshes. *Mathematical Modeling and Numerical Analysis*, 21:199–238, 1987.
2. I. Babuška, J. Tinsley Oden, and C.E. Baumann. A discontinuous *hp* finite element method for diffusion problems: 1-D Analysis . To appear, *Computer and Mathematics with Applications*, also TICAM Report 97-22, 1999.
3. F. Bassi and R. Rebay. A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations. *J. Comp. Physics*, 1997. Submitted.
4. F. Bassi, R. Rebay, M. Savini, and S. Pedinotti. The Discontinuous Galerkin method applied to CFD problems. In *Second European Conference on Turbomachinery, Fluid Dynamics and Thermodynamics*. ASME, 1995.
5. C.E. Baumann. *An hp-Adaptive Discontinuous Finite Element Method for Computational Fluid Dynamics*. PhD dissertation, The University of Texas at Austin, Aug 1997.
6. C.E. Baumann and J. Tinsley Oden. A discontinuous *hp* finite element method for the solution of the Navier-Stokes equations. In *Tenth International Conference on Finite Elements in Fluids*, Jan. 5-8 1998.
7. C.E. Baumann and J. Tinsley Oden. A discontinuous *hp* finite element method for convection-diffusion problems. TICAM Report 97-23. *Comp. Meth. Appl. Mech. Eng., in press*, special issue on Spectral, Spectral Elements, and *hp* Methods in CFD, edited by G.E. Karniadakis, M. Ainsworth, and C. Bernardi., 1999.
8. C.E. Baumann and J. Tinsley Oden. A discontinuous *hp* finite element method for the Euler and Navier-Stokes equations. In press, *Int. J. Num. Meth. Fluids*, edited by J. Heinrich, 1999.
9. K.S. Bey. *An hp-Adaptive discontinuous Galerkin method for Hyperbolic Conservation Laws*. PhD dissertation, The University of Texas at Austin, May 1994.
10. B.Riviere and M.F.Wheeler. Part I. Improved Energy Estimates for Interior Penalty, Constrained and Discontinuous Galerkin Methods for Elliptic Problems. *TICAM report*, April 1999.
11. B.Riviere, M.F.Wheeler, and C.Baumann. Part II. Discontinuous Galerkin Method Applied to a Single Phase Flow in Porous Media. *TICAM report*, April 1999.
12. P.G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North-Holland, Amsterdam, 1978.
13. B. Cockburn. An introduction to the discontinuous galerkin method for convection-dominated problems. School of Mathematics, University of Minnesota, 1997.
14. B. Cockburn, S. Hou, and C.W. Shu. TVB Runge-Kutta local projection dicontinuous Galerkin finite element for conservation laws IV: The multi-dimensional case. *Math. Comp.*, 54:545, 1990.
15. B. Cockburn, G. Karniadakis, and C-W. Shu. An overview of the development of discontinuous Galerkin methods. In *International Symposium on Discontinous Galerkin Methods*, Lecture Notes in Computational Science and Engineering. Springer-Verlag, 1999.

16. B. Cockburn, S.Y. Lin, and C.W. Shu. TVB Runge-Kutta local projection dicontinuous Galerkin finite element for conservation laws III: One-dimensional systems. *J. Comp. Physics*, 84:90–113, 1989.
17. B. Cockburn and C.W. Shu. TVB Runge-Kutta local projection dicontinuous Galerkin finite element for conservation laws II: General framework. *Math. Comp.*, 52:411–435, 1989.
18. B. Cockburn and C.W. Shu. The Local Discontinuous Galerkin method for time dependent convection-diffusion systems. *SIAM J. Numer. Anal.*, 1997. Submitted.
19. L.M. Delves and C.A. Hall. An implicit matching principle for global element calculations. *J. Inst. Math. Appl.*, 23:223–234, 1979.
20. U. Ghia, K.N. Ghia, and C.T. Shin. High-Re Solutions for Incompressible Flow Using the Navier-Stokes Equations and a Multigrid Method. *Journal of Computational Phisics*, 48:387–411, 1982.
21. J.A. Hendry and L.M. Delves. The global element method applied to a harmonic mixed boundary value problem. *J. Comp. Phys.*, 33:33–44, 1979.
22. T. Hughes, G. Engel, L. Mazzei, and M. Larson. A comparison of discontinuous and continuous Galerkin methods. In *International Symposium on Discontinous Galerkin Methods*, Lecture Notes in Computational Science and Engineering. Springer-Verlag, 1999.
23. P. Lesaint and P.A. Raviart. On a finite element method for solving the neutron transport equation. In C. de Boor, editor, *Mathematical Aspects of Finite Elements in Partial Differential Equations*, pages 89–123. Academic Press, 1974.
24. P. Lesaint and P.A. Raviart. Finite element collocation methods for first-order systems. *Math. Comp.*, 33(147):891–918, 1979.
25. I. Lomtev and G.E. Karniadakis. A discontinuous Galerkin method for the Navier-Stokes equations. Submitted, *Int. J. Num. Meth. Fluids*, 1997.
26. I. Lomtev and G.E. Karniadakis. Simulations of viscous supersonic flows on unstructured meshes. AIAA-97-0754, 1997.
27. I. Lomtev, C.B. Quillen, and G.E. Karniadakis. Spectral/hp methods for viscous compressible flows on unstructured 2d meshes. To appear, *J. Comp. Phys.*, 1998.
28. I. Lomtev, C.W. Quillen, and G. Karniadakis. Spectral/hp methods for viscous compressible flows on unstructured 2d meshes. Technical report, Center for Fluid Mechanics Turbulence and Computation - Brown University, Box 1966, Providence RI 02912, Dec. 1996.
29. J. Nitsche. Über ein Variationsprinzip zur Lösung von Dirichlet Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind. *Abh. Math. Sem. Univ. Hamburg*, 36:9–15, 1971.
30. J. T. Oden and G. F. Carey. *Texas Finite Elements Series Vol. IV - Mathematical Aspects.* Prentice-Hall, 1983.
31. J. Tinsley Oden, I. Babuška, and C.E. Baumann. A discontinuous *hp* finite element method for diffusion problems. *Journal of Computational Physics*, (146):491–519, 1998.
32. P. Percell and M.F. Wheeler. A local residual finite element procedure for elliptic equations. *SIAM J. Numer. Anal.*, 15(4):705–714, August 1978.
33. T.C. Warburton, I. Lomtev, R.M. Kirby, and G.E. Karniadakis. A discontinuous Galerkin method for the Navier-Stokes equations on hybrid grids. Center for Fluid Mechanics 97-14, Division of Applied Mathematics, Brown University, 1997.

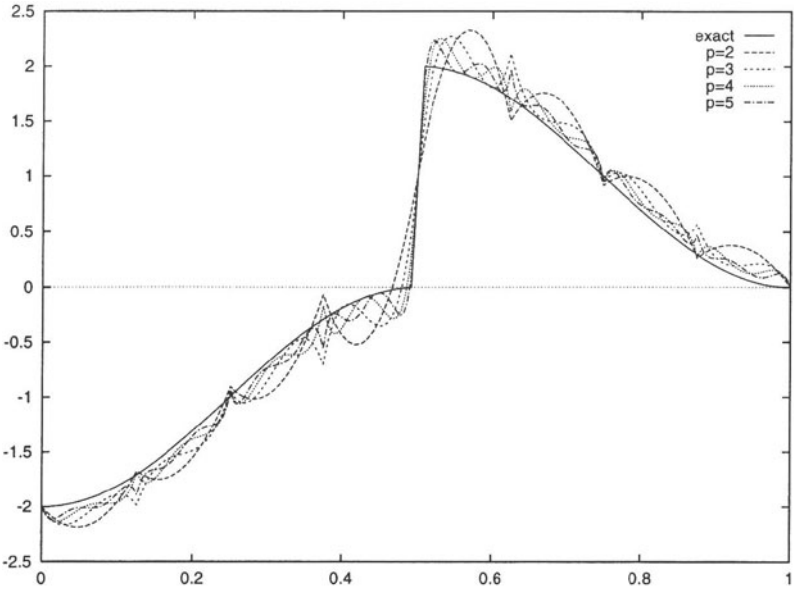


Fig. 1. Hemker problem: $\alpha = 10^{-10}$ and $h = 1/10$, Continuous Galerkin.

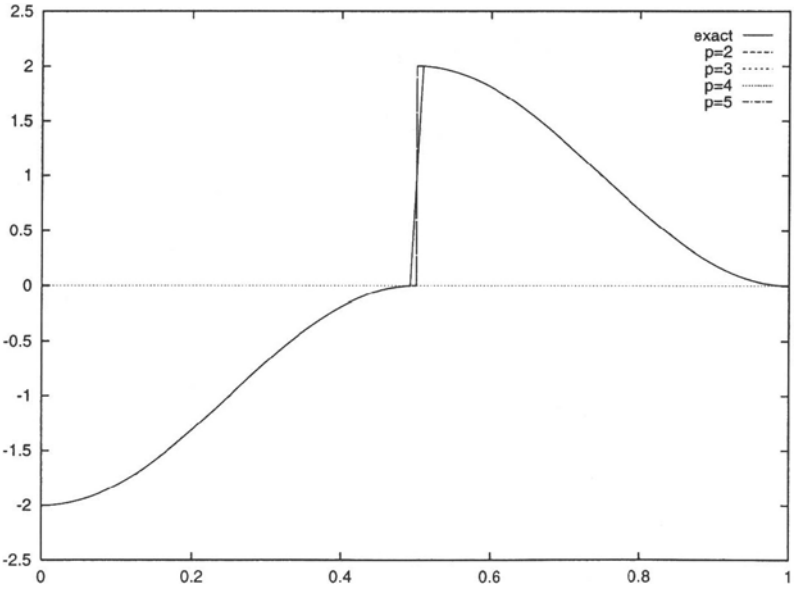


Fig. 2. Hemker problem: $\alpha = 10^{-10}$ and $h = 1/10$, Discontinuous Galerkin.

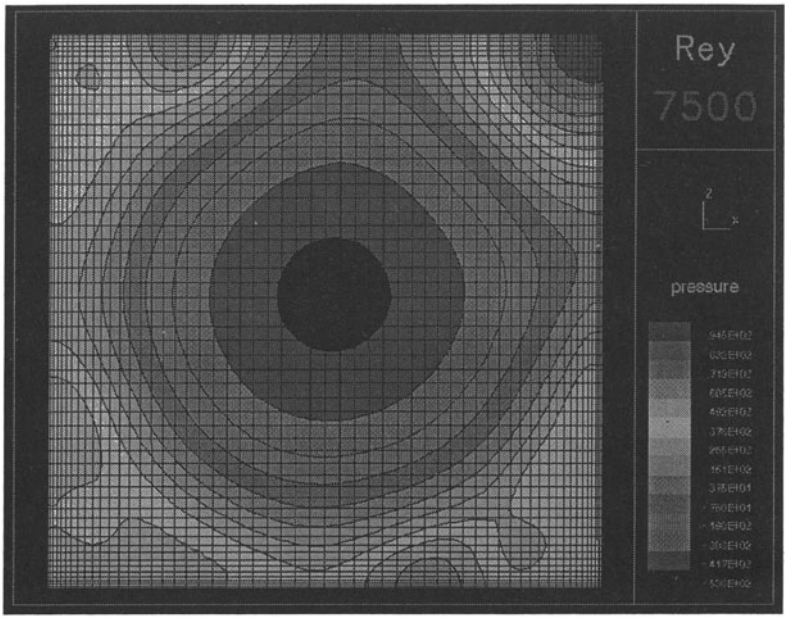


Fig. 3. Driven cavity at $Re = 7500$: Mesh and Pressure contours.

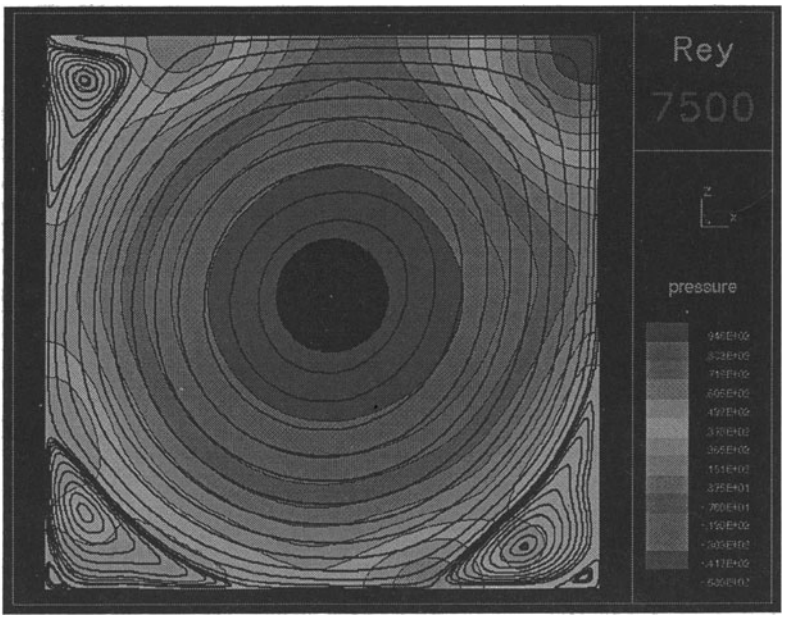


Fig. 4. Driven cavity at $Re = 7500$: Pressure contours and streamlines.

GMRES Discontinuous Galerkin Solution of the Compressible Navier-Stokes Equations

F. Bassi[1] and S. Rebay[2]

[1] Dipartimento di Energetica
Università degli Studi di Ancona
Via Brecce Bianche, 60100 Ancona, Italy
[2] Diparimento di Ingegneria Meccanica
Università degli Studi di Brescia
Via Branze 38, 25123 Brescia, Italy

Abstract

We present an implicit solution method for the compressible Navier-Stokes equations based on a Discontinuous Galerkin space discretization and on the implicit backward Euler time integration scheme. The linear system arising from the implicit time stepping scheme are solved with the preconditioned GMRES iterative method. Several preconditioners have been considered. We describe the features of the method and investigate its accuracy and performance by computing several classical 2-dimensional test cases.

1 Introduction

The Discontinuous Galerkin (DG) method is a recently developed higher-order accurate method which has been receiving great attention by several researchers, among others see e.g. [10,11,2,5,6,1,3], because of its several promising features. First of all, the DG method combines two key features which characterize the finite volume and the finite element method, the physics of wave propagation being accounted for by means of (approximate) Riemann solvers at element interfaces and accuracy being pursued by means of high-order polynomial approximations within elements. The method is therefore ideally suited to compute high-order accurate solution of the Navier–Stokes equations on general structured or unstructured grids. A second interesting feature of the DG method is the compactness of the scheme. The expansion coefficients of the numerical solution associated to a generic element are in fact coupled only with those associated to neighboring elements (that is the elements sharing a face). In the case of triangular (tetrahedral) elements, this means that coupling is introduced only among the unknowns associated to four (five) elements, respectively. This compactness results in sparse matrices which are very convenient for implicit time integration schemes (especially in 3D).

In this work we present a DG method which is a variant of that described in [4,3] exploiting the linear dependence of the viscous flux function

on the gradient of the conservation variables. The equations are integrated in time with the implicit backward Euler scheme. At each time step the coupled nonlinear equations resulting from the time stepping scheme are locally linearized and therefore reduced to a system of linear equations, which is iteratively solved with the preconditioned GMRES method. As preconditioner we have considered the standard incomplete LU factorization (ILU) with different levels of fill-in and a block diagonal preconditioner which turned out to be particularly well suited to the DG Galerkin discretization of the NS equations.

In the following we give a complete description of the method, with particular attention to the discretization of the viscous part of the NS equations. The performance of the method is displayed by computing the compressible laminar flow for two classical airfoil steady state calculations.

2 DG Space Discretization of the Navier-Stokes Equations

The compressible Navier-Stokes equations can be written in compact form as

$$\frac{\partial u}{\partial t} + \nabla \cdot \boldsymbol{f}_{\mathrm{c}}(u) + \nabla \cdot \boldsymbol{f}_{\mathrm{v}}(u, \nabla u) = 0,$$

where $u \in \mathbb{R}^{d+2}$ is the vector of the conservative variables, $\boldsymbol{f}_{\mathrm{c}}$ and $\boldsymbol{f}_{\mathrm{v}} \in \mathbb{R}^{d+2} \otimes \mathbb{R}^{d}$ are the inviscid and viscous flux functions, respectively, and d denotes the number of space dimensions. The viscous flux function is linear with respect to ∇u, i.e.

$$\boldsymbol{f}_{\mathrm{v}}(u, \nabla u) = \boldsymbol{\mathcal{A}}_{\mathrm{v}}(u) \nabla u, \quad \boldsymbol{\mathcal{A}}_{\mathrm{v}}(u) = \frac{\partial \boldsymbol{f}_{\mathrm{v}}(u, \nabla u)}{\partial \nabla u} \ . \tag{1}$$

A discrete version of the NS equations is obtained by subdividing Ω into a set of elements $\{e\}$ (triangles in this work), and by restricting u and ϕ to be polynomial functions inside each element. No global continuity is enforced on u and ϕ, which are therefore discontinuous at element interfaces. By splitting the integral over Ω into the sum of integrals over the elements e and by performing an integration by parts, we obtain the weak formulation

$$\sum_e \int_{\Omega_e} \phi \frac{\partial u}{\partial t} \, d\Omega + \sum_e \oint_{\partial \Omega_e} \phi \boldsymbol{n} \cdot \boldsymbol{f}(u, \nabla u) \, d\sigma - \sum_e \int_{\Omega_e} \nabla \phi \cdot \boldsymbol{f}(u, \nabla u) \, d\Omega = 0 \ . \tag{2}$$

Due to the discontinuous function approximation chosen, the flux functions appearing in the boundary integral of the previous equation is not uniquely defined and does not disappear for internal interfaces as in the standard

continuous finite element method. It is instead necessary to resort to an interface treatment which *weakly* enforces continuity at element interfaces (and the boundary conditions at boundary sides), thus providing a coupling between neighboring elements which would be otherwise completely missing. In general this is accomplished by replacing the physical normal flux $\boldsymbol{n} \cdot \boldsymbol{f}(u, \nabla u)$ with a numerical flux $h(u^-, \nabla u^-, u^+, \nabla u^+, \boldsymbol{n}^-)$. With the notation $(\cdot)^-$ and $(\cdot)^+$ we denote the interface value of any quantity associated to the two elements sharing the face, in which the normal unit vector $\boldsymbol{n}^-$ points outward from the element associated to the values $(\cdot)^-$.

To put in evidence the different role played by the contour integral for internal interfaces and for boundary sides (and to better reflect the way in which computations are actually performed), the sum of contour integrals appearing in Eq. (2) can be rearranged as the sums of internal interface integrals and of boundary side integrals

$$-\sum_e \int_{\Omega_e} \nabla\phi \cdot \boldsymbol{f}(u, \nabla u)\, d\Omega + \sum_f \int_{\sigma_f} (\phi^- h^- + \phi^+ h^+)\, d\sigma + \sum_b \int_{\sigma_b} \phi \boldsymbol{n}^- \cdot \boldsymbol{f}(u^*, \nabla u^*)\, d\sigma, \quad (3)$$

where the subscript f refer to internal interfaces, the subscript b to boundary sides, and

$$h^\pm = h(u^\pm, \nabla u^\pm, u^\mp, \nabla u^\mp, \boldsymbol{n}^\pm) \ .$$

Notice that for each interface integral of the second sum in Eq. (3) there are two numerical flux contribution which correspond to the two elements sharing a face. The flux function arguments u^* and ∇u^* appearing in the boundary integrals of Eq. (3) are defined as

$$u^* = u^b, \qquad \nabla u^* = \nabla u^-$$

to prescribe Dirichlet conditions u^b and as

$$u^* = u^-, \qquad \boldsymbol{n} \cdot \nabla u^* = \boldsymbol{n} \cdot \nabla u^b$$

to prescribe Neumann conditions $\boldsymbol{n} \cdot \nabla u^b$.

The formulation of the numerical flux function for the inviscid part of the NS equations is completely analogous to that commonly employed in upwind finite volume methods. In our computations we have used the van Leer flux difference splitting numerical flux as modified by Hänel [12]. The inviscid terms of the NS equations are therefore discretized as

$$\sum_e \int_{\Omega_e} \phi \frac{\partial u}{\partial t}\, d\Omega - \sum_e \int_{\Omega_e} \nabla\phi \cdot \boldsymbol{f}(u, \nabla u)\, d\Omega + \sum_f \int_{\sigma_f} (\phi^- h^- + \phi^+ h^+)\, d\sigma + \sum_b \int_{\sigma_b} \phi \boldsymbol{n} \cdot \boldsymbol{f}_c(u^b)\, d\sigma = 0 \ . \quad (4)$$

The formulation of a numerical flux function for the viscous part of the Navier-Stokes equations does not have a counterpart in the finite volume method, and it therefore will be described in greater detail. In order to accommodate the generalized Laplacian operator

$$\nabla \cdot \boldsymbol{f}_{\mathrm{v}} = \nabla \cdot (\mathcal{A}_{\mathrm{v}} \nabla u)$$

in weak variational form in a space of discontinuous functions, we reformulate the NS equations as the system

$$\begin{cases} \boldsymbol{f}_{\mathrm{v}} = \mathcal{A}_{\mathrm{v}} \nabla u \\ \partial_t u + \nabla \cdot \boldsymbol{f}_{\mathrm{c}} + \nabla \cdot \boldsymbol{f}_{\mathrm{v}} = 0, \end{cases} \tag{5}$$

in which the two coupled first order equations can be approximated with DG techniques similar to those already developed in the case of hyperbolic systems of conservation laws.

Let's begin from the first equation of the system. A "weakly continuized" viscous flux $\widetilde{\boldsymbol{f}_{\mathrm{v}}}$ is defined as

$$\begin{aligned} &\sum_e \int_{\Omega_e} \psi \cdot \widetilde{\boldsymbol{f}_{\mathrm{v}}}\, d\Omega = -\sum_e \int_{\Omega_e} u\, \nabla \cdot (\mathcal{A}_{\mathrm{v}}^T \psi)\, d\Omega \\ &+ \frac{1}{2} \sum_e \oint_{\partial\Omega_{e_i}} [\boldsymbol{n} \cdot (\mathcal{A}_{\mathrm{v}}^T \psi)]^- (u^+ + u^-)\, d\sigma + \sum_e \oint_{\partial\Omega_{e_b}} [\boldsymbol{n} \cdot (\mathcal{A}_{\mathrm{v}}^T \psi)]^- u^b\, d\sigma, \end{aligned} \tag{6}$$

where $\partial\Omega_{e_i}$ denotes the part of $\partial\Omega \notin \Gamma$ and $\partial\Omega_{e_b}$ denotes the part $\in \Gamma$. By summing and subtracting to the right hand side of Eq. (6) the expression

$$\frac{1}{2} \sum_e \oint_{\partial\Omega_{e_i}} [\boldsymbol{n} \cdot (\mathcal{A}_{\mathrm{v}}^T \psi)]^- u^-\, d\sigma + \sum_e \oint_{\partial\Omega_{e_b}} [\boldsymbol{n} \cdot (\mathcal{A}_{\mathrm{v}}^T \psi)]^- u^-\, d\sigma$$

and by "backward" integrating by parts, we can rewrite Eq. (6) as

$$\begin{aligned} \sum_e \int_{\Omega_e} \psi \cdot \boldsymbol{\delta}\, d\Omega = \sum_e \int_{\Omega_e} \psi \cdot \widetilde{\boldsymbol{f}_{\mathrm{v}}}\, d\Omega - \sum_e \int_{\Omega_e} \psi \cdot (\mathcal{A}_{\mathrm{v}} \nabla u)\, d\Omega = \\ \frac{1}{2} \sum_e \oint_{\partial\Omega_{e_i}} [\boldsymbol{n} \cdot (\mathcal{A}_{\mathrm{v}}^T \psi)]^- (u^+ - u^-)\, d\sigma \\ + \sum_e \oint_{\partial\Omega_{e_b}} [\boldsymbol{n} \cdot (\mathcal{A}_{\mathrm{v}}^T \psi)]^- (u^b - u^-)\, d\sigma\ . \end{aligned} \tag{7}$$

The above equation defines the auxiliary variable $\boldsymbol{\delta} = \widetilde{\boldsymbol{f}_{\mathrm{v}}} - \mathcal{A}_{\mathrm{v}} \nabla u$ as the difference between the weakly continuized viscous flux $\widetilde{\boldsymbol{f}_{\mathrm{v}}}$ — which takes into account the effect of the jump $u^+ - u^-$ at element interfaces and of the jump $u^b - u^-$ at the boundary — and the "internal" viscous flux $\mathcal{A}_{\mathrm{v}} \nabla u$. By

rearranging the sums of contour integrals as a sum over internal interfaces plus a sum over boundary faces, we can express $\boldsymbol{\delta}$ as

$$\sum_e \int_{\Omega_e} \boldsymbol{\psi} \cdot \boldsymbol{\delta}\, d\Omega = -\frac{1}{2} \sum_f \int_{\sigma_f} [(\boldsymbol{\mathcal{A}}_{\mathrm{v}}^T \boldsymbol{\psi})^- + (\boldsymbol{\mathcal{A}}_{\mathrm{v}}^T \boldsymbol{\psi})^+] \cdot [(u\boldsymbol{n})^- + (u\boldsymbol{n})^+]\, d\sigma + \sum_b \int_{\sigma_b} [\boldsymbol{n} \cdot (\boldsymbol{\mathcal{A}}_{\mathrm{v}}^T \boldsymbol{\psi})]^- \cdot (u^b - u^-)\, d\sigma \ . \quad (8)$$

We now consider the weak formulation of the last term of the second equation of the system, which, by substituting the physical viscous flux $\boldsymbol{f}_{\mathrm{v}}$ with the weakly continuized flux $\widetilde{\boldsymbol{f}_{\mathrm{v}}} = \boldsymbol{\mathcal{A}}_{\mathrm{v}} \boldsymbol{\nabla} u + \boldsymbol{\delta}$, and by rearranging the sums of contour integrals as sums of interface and boundary integrals, can be written in weak form as

$$\begin{aligned} &- \sum_e \int_{\Omega_e} \boldsymbol{\nabla}\phi \cdot (\boldsymbol{\mathcal{A}}_{\mathrm{v}} \boldsymbol{\nabla} u)\, d\Omega - \sum_e \int_{\Omega_e} \boldsymbol{\nabla}\phi \cdot \boldsymbol{\delta}\, d\Omega \\ &\quad + \frac{1}{2} \sum_f \int_{\sigma_f} [(\phi \boldsymbol{n})^- + (\phi \boldsymbol{n})^+] \cdot [(\boldsymbol{\mathcal{A}}_{\mathrm{v}} \boldsymbol{\nabla} u)^+ + (\boldsymbol{\mathcal{A}}_{\mathrm{v}} \boldsymbol{\nabla} u)^-]\, d\sigma \\ &\quad + \frac{1}{2} \sum_f \int_{\sigma_f} [(\phi \boldsymbol{n})^- + (\phi \boldsymbol{n})^+] \cdot (\boldsymbol{\delta}^+ + \boldsymbol{\delta}^-)\, d\sigma \\ &\quad + \sum_b \int_{\sigma_b} (\phi \boldsymbol{n})^- \cdot (\boldsymbol{\mathcal{A}}_{\mathrm{v}} \boldsymbol{\nabla} u)^b\, d\sigma + \sum_b \int_{\sigma_b} (\phi \boldsymbol{n})^- \cdot \boldsymbol{\delta}^b\, d\sigma \ . \end{aligned} \quad (9)$$

By virtue of Eq (8), in which we choose the test function $\boldsymbol{\psi} = \boldsymbol{\nabla}\phi$, the second volume integral sum of the above expression can be reformulated as a summation of interface and boundary integrals. The weak formulation of the second equations of system (5) can therefore be written as

$$\begin{aligned} &\mathbb{E} - \sum_e \int_{\Omega_e} \boldsymbol{\nabla}\phi \cdot (\boldsymbol{\mathcal{A}}_{\mathrm{v}} \boldsymbol{\nabla} u)\, d\Omega \\ &\quad + \frac{1}{2} \sum_f \int_{\sigma_f} [(\boldsymbol{\mathcal{A}}_{\mathrm{v}}^T \boldsymbol{\nabla}\phi)^- + (\boldsymbol{\mathcal{A}}_{\mathrm{v}}^T \boldsymbol{\nabla}\phi)^+] \cdot [(u\boldsymbol{n})^- + (u\boldsymbol{n})^+]\, d\sigma \\ &\quad + \frac{1}{2} \sum_f \int_{\sigma_f} [(\phi \boldsymbol{n})^- + (\phi \boldsymbol{n})^+] \cdot [(\boldsymbol{\mathcal{A}}_{\mathrm{v}} \boldsymbol{\nabla} u)^+ + (\boldsymbol{\mathcal{A}}_{\mathrm{v}} \boldsymbol{\nabla} u)^-]\, d\sigma \\ &\quad + \frac{1}{2} \sum_f \int_{\sigma_f} [(\phi \boldsymbol{n})^- + (\phi \boldsymbol{n})^+] \cdot (\boldsymbol{\delta}^+ + \boldsymbol{\delta}^-)\, d\sigma \\ &\quad - \sum_b \int_{\sigma_b} [\boldsymbol{n} \cdot (\boldsymbol{\mathcal{A}}_{\mathrm{v}} \boldsymbol{\nabla}\phi)]^- \cdot (u^b - u^-)\, d\sigma \\ &\quad + \sum_b \int_{\sigma_b} (\phi \boldsymbol{n})^- \cdot (\boldsymbol{\mathcal{A}}_{\mathrm{v}} \boldsymbol{\nabla} u)^b\, d\sigma + \sum_b \int_{\sigma_b} (\phi \boldsymbol{n})^- \cdot \boldsymbol{\delta}^b\, d\sigma = 0, \end{aligned} \quad (10)$$

where $\mathbb{E}$ denotes the DG discretization of the inviscid part of the NS equations as given in Eq. (4). The previous equations and Eq. (8) can be regarded as a system of two equations in the unknowns u and $\boldsymbol{\delta}$ which discretize the NS equations in mixed form (see e.g. [1]). Unfortunately a DG Galerkin method for the solution of the NS equations very similar to the one described so far displays an unsatisfactory convergence rate for polynomial approximations of odd order. A cure to this convergence problem is suggested in [4,3], and has been thoroughly investigated theoretically in a series of papers of Brezzi and coworkers [9,8,7].

Following the ideas introduced in [4,3], we therefore replace the variable $\boldsymbol{\delta}$ appearing in the third interface integral and in the third boundary integral of Eq. (10) with a locally defined variable $\boldsymbol{\delta}_f$ given by

$$\int_{\Omega_e{}^{\pm}} \boldsymbol{\psi} \cdot \boldsymbol{\delta}_f^{\pm} \, d\Omega = \frac{1}{2} \int_{\sigma_f} [\boldsymbol{n} \cdot (\boldsymbol{\mathcal{A}}_{\mathrm{v}}^T \boldsymbol{\psi})]^{\pm} (u^{\mp} - u^{\pm}) \, d\sigma \tag{11}$$

for internal interfaces and by

$$\int_{\Omega_e{}^{-}} \boldsymbol{\psi} \cdot \boldsymbol{\delta}_f^{-} \, d\Omega = \int_{\sigma_b} [\boldsymbol{n} \cdot (\boldsymbol{\mathcal{A}}_{\mathrm{v}}^T \boldsymbol{\psi})]^{-} \cdot (u^b - u^{-}) \, d\sigma \tag{12}$$

for boundary sides. The previously defined variable $\boldsymbol{\delta}$ and the newly defined variables $\boldsymbol{\delta}_f$ are related for each element e through the relation

$$\boldsymbol{\delta}|_e = \sum_{f \in e} \boldsymbol{\delta}_f,$$

where the summation includes the faces on the boundary of element e. The modified DG formulation of the NS equations can therefore be written as

$$\begin{aligned}
&\mathbb{E} - \sum_e \int_{\Omega_e} \boldsymbol{\nabla} \phi \cdot (\boldsymbol{\mathcal{A}}_{\mathrm{v}} \boldsymbol{\nabla} u) \, d\Omega \\
&\quad + \frac{1}{2} \sum_f \int_{\sigma_f} [(\boldsymbol{\mathcal{A}}_{\mathrm{v}}^T \boldsymbol{\nabla} \phi)^{-} + (\boldsymbol{\mathcal{A}}_{\mathrm{v}}^T \boldsymbol{\nabla} \phi)^{+}] \cdot [(u\boldsymbol{n})^{-} + (u\boldsymbol{n})^{+}] \, d\sigma \\
&\quad + \frac{1}{2} \sum_f \int_{\sigma_f} [(\phi \boldsymbol{n})^{-} + (\phi \boldsymbol{n})^{+}] \cdot [(\boldsymbol{\mathcal{A}}_{\mathrm{v}} \boldsymbol{\nabla} u)^{+} + (\boldsymbol{\mathcal{A}}_{\mathrm{v}} \boldsymbol{\nabla} u)^{-}] \, d\sigma \\
&\qquad + \frac{1}{2} \sum_f \int_{\sigma_f} [(\phi \boldsymbol{n})^{-} + (\phi \boldsymbol{n})^{+}] \cdot (\boldsymbol{\delta}_f^{+} + \boldsymbol{\delta}_f^{-}) \, d\sigma \\
&\qquad - \sum_b \int_{\sigma_b} [\boldsymbol{n} \cdot (\boldsymbol{\mathcal{A}}_{\mathrm{v}} \boldsymbol{\nabla} \phi)]^{-} \cdot (u^b - u^{-}) \, d\sigma \\
&\quad + \sum_b \int_{\sigma_b} (\phi \boldsymbol{n})^{-} \cdot (\boldsymbol{\mathcal{A}}_{\mathrm{v}} \boldsymbol{\nabla} u)^b \, d\sigma + \sum_b \int_{\sigma_b} (\phi \boldsymbol{n})^{-} \cdot \boldsymbol{\delta}_f^b \, d\sigma = 0 \ .
\end{aligned} \tag{13}$$

Notice that the second interface integral of the previous equation, i.e. that containing the sum $(\boldsymbol{\delta}_f^+ + \boldsymbol{\delta}_f^-)$, can be rewritten in the equivalent form

$$\frac{1}{2}\int_{\sigma_f} (\phi^+ - \phi^-)[(\boldsymbol{n}\cdot\boldsymbol{\delta}_f)^+ - (\boldsymbol{n}\cdot\boldsymbol{\delta}_f)^-]\, d\sigma,$$

thus showing that what is really needed in the computations is only the normal component $(\boldsymbol{n}\cdot\boldsymbol{\delta}_f)^\pm$ of $\boldsymbol{\delta}_f$. An equation giving the normal component $\boldsymbol{n}\cdot\boldsymbol{\delta}_f$ can be obtained from Eq. (11) for an interface — or from Eq. (12) for a boundary side — by considering test functions $\boldsymbol{\psi} = \phi\boldsymbol{n}$. For an internal interface, $\boldsymbol{n}\cdot\boldsymbol{\delta}_f$ on the face f is for example given by

$$\int_{e^\pm} \phi(\boldsymbol{n}\cdot\boldsymbol{\delta}_f)^\pm\, d\Omega = -\frac{1}{2}\int_{\sigma_f} \phi\left[\boldsymbol{n}\cdot(\boldsymbol{\mathcal{A}}_v^T\boldsymbol{n})\right]^\pm (u^\mp - u^\pm)\, d\sigma\ . \qquad (14)$$

Eq. (13) with the values of $\boldsymbol{\delta}_f$ expressed in terms of $u^\pm$ according to (14) provides a symmetric DG discretization of the viscous terms of the NS equations entirely in terms of the original variable u, which does not suffer the convergence problems related to the auxiliary variables $\boldsymbol{\delta}$ originally introduced to define the weakly continuized viscous flux function. In practice, even if starting from a formulation of the NS equations as a system of two first order equations, we have ended with a formulation in which the auxiliary variables can be computed on a local basis for the internal interfaces and for the boundary sides, with no requirement for additional storage at global level.

The DG discretization of the viscous terms here described bears some resemblance with that introduced by Baumann and Oden (see e.g. [5,6]), which can be obtained from Eq. 13 by removing the third interface and boundary integral summations (i.e. the terms involving $\boldsymbol{\delta}_f$) and by changing the sign of the first interface and boundary integral summations. Because of the different sign, the DG scheme here described leads to a symmetric discretization (for a symmetric matrix $\boldsymbol{\mathcal{A}}_v$) while that of Baumann and Oden does not. Baumann and Oden's scheme is however simpler since it does not require the terms containing $\boldsymbol{\delta}_f$.

The resulting scheme is characterized by a very compact support. In fact, the unknowns associated with element e are only coupled with the unknowns associated with the elements which share a face with e. This results in a discretized spatial operator which can be solved very efficiently and is therefore very well suited to be used with an implicit time integration scheme, as we are going to describe in the next section.

3 Implicit Time Integration

The space discretized equations are advanced in time with the implicit backward Euler time integration scheme. If the equations are linearized in time,

at each time step we have to solve a linear system $\boldsymbol{A}^n \boldsymbol{x}^n = \boldsymbol{b}^n$ of algebraic equations, where $\boldsymbol{x}^n$ is the solution increment $(\boldsymbol{U}^{n+1} - \boldsymbol{U}^n)$ and $\boldsymbol{b}^n = -\boldsymbol{R}^n$. The coefficient matrix $\boldsymbol{A}^n$ can be regarded as a sparse block matrix of $m \times m$ blocks, m being the number of unknown fields (ρ, ρE, ρu, ρv) times the number of degrees of freedom used to represent each field within an element. The number of nonzero blocks of a generic block row i is equal to the number of elements sharing a face with element i plus one. The linearization of the NS equations is simply accomplished by evaluating all the Jacobians, both inviscid and viscous, at time level n, thus reducing both the inviscid and the viscous parts of the NS equations to linear operators in u. The construction of the viscous linearized operator starting from Eq. (13) and (14) is a little bit involved but straightforward.

The linear systems are solved by means of the left preconditioned GMRES iterative solution algorithm. It is well known that the efficiency of the PGMRES algorithm is crucially dependent on the use of a preconditioner matrix $\boldsymbol{P}$. A reasonable compromise between efficiency and storage requirements has been found by using as preconditioner the $m \times m$ block diagonal part of $\boldsymbol{A}$ (see also [6], in which the linear system is solved by Jacobi and Gauss-Seidel iterations).

4 Numerical Results

In order to check the accuracy and the order of convergence of the DG discretization of the viscous term, we have considered the simple test problem

$$\boldsymbol{\nabla} \cdot (\boldsymbol{\mathcal{A}}_v \boldsymbol{\nabla} u) = s, \tag{15}$$

on the unit square with homogeneous Dirichlet conditions, where u is a vector unknown, and $\boldsymbol{\mathcal{A}}_v$ is a full constant matrix. The tests have been performed on 5 successively finer grids starting from an extremely coarse grid containing only 8 triangular elements. In all the cases the method converges with the expected order of accuracy provided that the grid is fine enough to capture the relevant feature of the solution (results not reported for brevity).

We next present the computational results for two laminar calculations around the NACA0012 airfoil at two flow regimes, namely $\mathrm{M}_\infty = 0.8$, $\alpha_\infty = 10°$, $\mathrm{Re} = 73$, and $\mathrm{M}_\infty = 0.5$, $\alpha_\infty = 0°$, $\mathrm{Re} = 5000$. The convergence history of the computations is monitored by the error indicator r_n, defined as

$$r_n = \log_{10} \left[||R(\boldsymbol{U}_P^n)||_{L_2} / ||R(\boldsymbol{U}_Q^0)||_{L_2} \right], \tag{16}$$

where $R(\boldsymbol{U}_P^n)$ denotes the residual at the n-th time step using polynomial approximation of order P (that is the polynomial order corresponding to the solution that we are monitoring), and $R(\boldsymbol{U}_Q^0)$ is the residual function computed for the initial data using polynomial approximation of order Q. This definition of the error indicator allows a comparison among residual histories of computations having different order of accuracy.

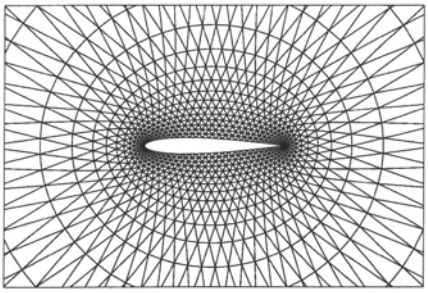


Fig. 1. Grid for the Naca0012 airfioil computations.

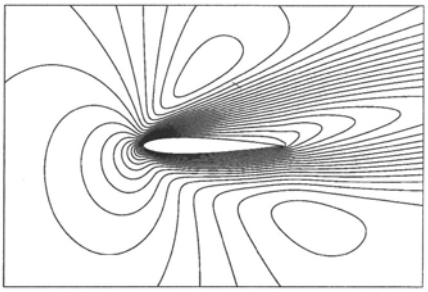


Fig. 2. Re = 73 Mach isolines computed with P3 elements.

In all the computations the CFL number is increased as the residual of the solution decreases according to equation $\mathrm{CFL}^{n+1} = \mathrm{CFL}_0 / r_n^{\beta}$ and subject to the additional user specified constraint $\mathrm{CFL}_0 \leq \mathrm{CFL} \leq \mathrm{CFL}_N$. All the test cases have been computed using $\beta = 1$ and $\mathrm{CFL}_N = 100$. The linear systems have been solved with the restarted GMRES(20) method implemented in the Sparskit2 library developed by Y. Saad and coworkers. All the computations have been performed on a grid containing 2048 triangular élements, 64 distributed along the airfoil and 16 in the direction normal to the airfoil (see Figure 1).

We have performed several test calculations using the ILU preconditioner with various levels of fill-in (up to 20) and the very simple block diagonal preconditioner described in the previous section. Quite surprisingly, for the NACA0012 computations the block diagonal preconditioner gives much better results than the ILU preconditioner both in terms of cpu time and in terms of the number of iterations needed to solve the linear system within a given level of accuracy. In fact, the ILU preconditioner displayed the same rate of convergence of the much simpler block diagonal preconditioner only allowing relatively large levels of fill-in (but with a consequently much greater cost both in terms of cpu time and memory required). This unexpected behavior could be due to a poor ordering of the unknowns, which does not affect the performance of the block diagonal preconditioner but could impair that of the ILU preconditioners. We have however not yet adopted any reordering strategy. We plan to address this issue in future work.

The solution computed with cubic elements for the Re = 73 test case is displayed in Figure 2, which show the accuracy which can obtained even on very coarse grids by using an high-order accurate method. The convergence history for the Re = 73 test case in terms of iterations and of cpu time is displayed in Figures 3 and 4. Notice that we have used as initial data for the high order computations the lower order solution previously computed. The solutions computed with linear, quadratic and cubic elements for the Re = 5000 test case are displayed in Figures 5, 6 and 7. Figure 8 shows a

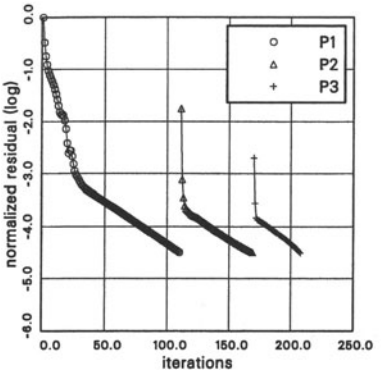


Fig. 3. Re = 73 residual history as a function of time steps.

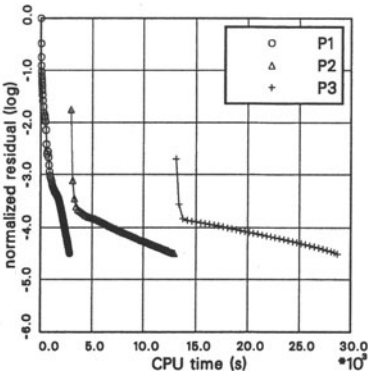


Fig. 4. Re = 73 residual history as a function of cpu time.

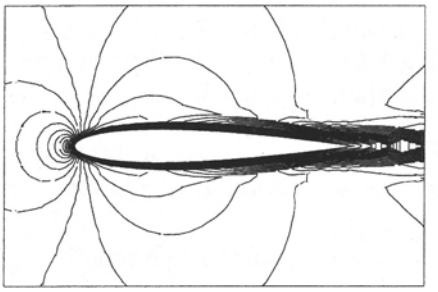


Fig. 5. Re = 5000 Mach isolines computed with P1 elements.

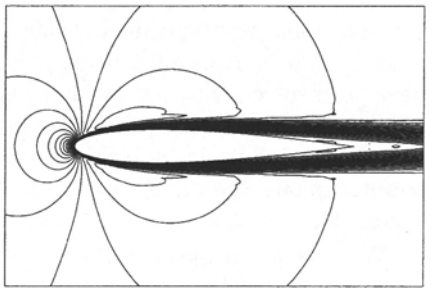


Fig. 6. Re = 5000 Mach isolines computed with P2 elements.

detail of the Mach isolines near the trailing edge and in the wake region behind the airfoil. This test case shows the improvement of the solution accuracy obtained by using high-order element. In fact P3 elements are needed to obtain an accurate solution on this grid (see e.g. [1]). The convergence history of the Re = 5000 test case, again in terms of both iteration number and cpu time, is given in Figures 9 and 10. Notice that the Re = 5000 test case requires less computational effort than the Re = 73 case, suggesting that the system of equations to be solved at each time step gets more ill-conditioned for lower Reynolds numbers.

5 Conclusions

We have here presented a method for the numerical solution of the compressible Navier-Stokes equations based on a DG spatial discretization and on the

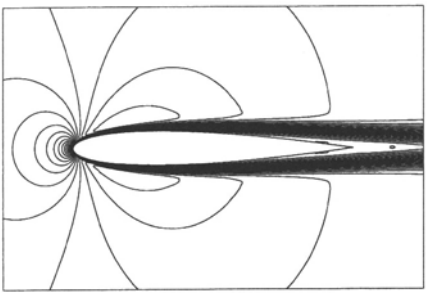


Fig. 7. Re = 5000 Mach isolines computed with P3 elements.

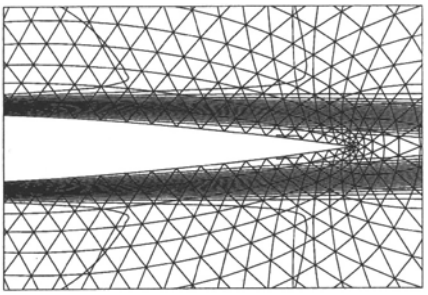


Fig. 8. Re = 5000 Mach isolines detail with P3 elements.

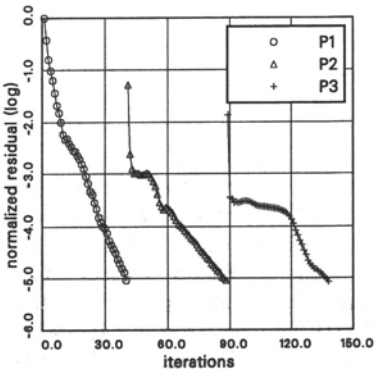


Fig. 9. Re = 5000 residual history as a function of time steps.

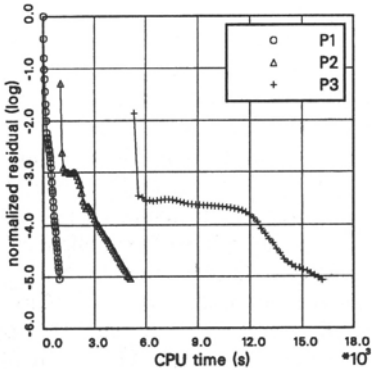


Fig. 10. Re = 5000 residual history as a function of cpu time.

fully implicit backward Euler time integration scheme. The linear systems arising from the implicit time discretization are solved with the preconditioned GMRES method. Various preconditioners have been considered, but the most efficient for the test cases attempted so far turned out to be a relatively simple block diagonal preconditioner. It is not clear at this stage of our investigation if the worst performance of ILU type preconditioners is due to a poor unknown ordering or if it is the block diagonal preconditioner which performs surprisingly well for DG discretizations (not excluding some other conjecture which we can not envisage at this moment). The results obtained so far are quite encouraging but we have considered only relatively simple test cases and we have not yet adopted any kind of reordering technique which could improve the performance of ILU preconditioners. We plan to work in this direction in the near future.

References

1. F. Bassi and S. Rebay. A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier–Stokes equations. *J. Comput. Phys.*, 131:267–279, 1997.
2. F. Bassi and S. Rebay. High-order accurate discontinuous finite element solution of the 2D Euler equations. *J. Comput. Phys.*, 138:251–285, 1997.
3. F. Bassi and S. Rebay. An implicit high-order discontinuous galerkin method for the steady state compressible Navier–Stokes equations. In K. D. Papailiou, D. Tsahalis, D. Périaux, C. Hirsh, and M. Pandolfi, editors, *Computational Fluid Dynamics 98, Proceedings of the Fourth European Computational Fluid Dynamics Conference*, volume 2, pages 1227–1233, Athens, Greece, September 5–7 1998. John Wiley and Sons.
4. F. Bassi, S. Rebay, G. Mariotti, S. Pedinotti, and M. Savini. A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows. In R. Decuypere and G. Dibelius, editors, *2nd European Conference on Turbomachinery Fluid Dynamics and Thermodynamics*, pages 99–108, Antwerpen, Belgium, March 5–7 1997. Technologisch Instituut.
5. C. E. Baumann and J. T. Oden. A discontinuous hp finite element method for convection-diffusion problems. to appear on CMAME, 1998.
6. C. E. Baumann and J. T. Oden. A discontinuous hp finite element method for the Euler and Navier-Stokes equations. to appear on IJNME, 1998.
7. F. Brezzi, G. Manzini, D. Marini, P. Pietra, and A. Russo. Discontinuous finite elements for diffusion problems. Technical Report 1112, IAN-CNR, via Ferrata, 1, 1998. submitted for publication to Numer. Meth. PDEs.
8. F. Brezzi, G. Manzini, D. Marini, P. Pietra, and A. Russo. Discontinuous galerkin approximations for elliptic problems. Technical Report 1110, IAN-CNR, via Ferrata, 1, 1998. to appear in "Atti del Convegno in Memoria di F. Brioschi, Istituto Lombardo di Scienze e Lettere, Milano, 22/23 ottobre 1997".
9. F. Brezzi, G. Manzini, D. Marini, P. Pietra, and P. Russo. Analisi delle proprietà di elementi finiti di tipo discontinuo. Technical Report 1107, IAN-CNR, via Ferrata, 1, 1998. relazione finale del progetto ENEL-MIGALE.
10. B. Cockburn, S. Hou, and C.-W. Shu. The Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: The multidimensional case. *Math. Comp.*, 54:454–581, 1990.
11. B. Cockburn and C.-W. Shu. The local discontinuous Galerkin method for time-dependent convection-diffusion systems. SIAM J. Numer. Anal. 35:2440-2463, 1998.
12. D. Hanel, R. Schwane, and G. Seider. On the accuracy of upwind schemes for the solution of the Navier–Stokes equations. AIAA Paper 87-1105 CP, AIAA, July 1987. Proceedings of the AIAA 8th Computational Fluid Dynamics Conference.

Explicit Finite Element Methods for Linear Hyperbolic Systems

RICHARD S. FALK[1] and GERARD R. RICHTER[2] *

[1] Department of Mathematics, Rutgers University, Piscataway, NJ 08854, falk@math.rutgers.edu
[2] Department of Computer Science, Rutgers University, Piscataway, NJ 08854, richter@cs.rutgers.edu

Abstract. Our focus is on explicit finite element discretization of transient, linear hyperbolic systems in arbitrarily many space dimensions. We propose several ways of generating suitable "explicit" meshes, and sketch an $O(h^{n+1/2})$ error estimate for a discontinuous Galerkin method. Continuous methods are also considered briefly. This paper parallels [2] in large part, while using a different approach in the analysis.

1 Introduction

The problem of interest to us here is a linear, symmetric hyperbolic system

$$\mathcal{L}\boldsymbol{u} \equiv \frac{\partial \boldsymbol{u}}{\partial t} + \sum_{i=1}^{N} A_i \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}_i} + B\boldsymbol{u} = \boldsymbol{f}, \qquad (\boldsymbol{x}, t) \in \Omega_T \equiv \Omega \times [0, T], \qquad (1)$$

where $\boldsymbol{u}$ is an m-vector and the matrices A_i are $m \times m$, symmetric, and constant. We assume Ω is a bounded polyhedral domain in R^N and denote its boundary by $\Gamma(\Omega)$. Likewise, we denote the boundary of the space-time domain Ω_T by $\Gamma(\Omega_T)$. Along $\Gamma(\Omega_T)$, the unit outer normal $\boldsymbol{n} = (\boldsymbol{n}_x, n_t) = (n_1, ..., n_N, n_t)$ has either $\boldsymbol{n}_x = 0$ or $n_t = 0$.

An appropriate set of initial and boundary conditions for (1) is

$$\begin{aligned} &\boldsymbol{u} = \boldsymbol{g} \qquad \text{at } t = 0, \\ &(D - \mathcal{N})\boldsymbol{u} = 0 \qquad \text{on } \Gamma(\Omega) \times [0, T], \end{aligned} \qquad (2)$$

where $D = \sum_{i=1}^{N} n_i A_i$ and $\mathcal{N} + \mathcal{N}^* \geq 0$. Problem (1)-(2) has the form of a Friedrichs system [4] for which a unique solution is guaranteed under certain restrictions.

An example of (1)-(2) is the wave equation in two space dimensions:

$$\begin{aligned} &w_{tt} - w_{xx} - w_{yy} = f, \\ &w, w_t \text{ given at } t = 0, \\ &w = 0 \text{ on } \Gamma(\Omega) \times [0, T]. \end{aligned}$$

* The authors were supported in part by NSF grant DMS-9704556 and DARPA grant 4-23685, respectively .

With $u_1 = w_x, u_2 = w_y, u_3 = w_t$, this can be written as

$$\boldsymbol{u}_t + \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} \boldsymbol{u}_x + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix} \boldsymbol{u}_y = \begin{pmatrix} 0 \\ 0 \\ f \end{pmatrix}.$$

For this system,

$$D = \begin{pmatrix} 0 & 0 & -n_1 \\ 0 & 0 & -n_2 \\ -n_1 & -n_2 & 0 \end{pmatrix},$$

and we may take

$$\mathcal{N} = \begin{pmatrix} 0 & 0 & n_1 \\ 0 & 0 & n_2 \\ -n_1 & -n_2 & 2 \end{pmatrix}.$$

Of the many previous finite element treatments of the general problem (1)-(2) and its related non-transient counterpart (e.g., [1], [5], [7], [8]), we know of none which is explicit, i.e., develops an approximate solution in an element by element fashion. This is our focus, in the setting of arbitrarily large m and N.

The key mesh requirement for explicitness is that

$$M \equiv n_t I + \sum_{i=1}^{N} n_i A_i$$

be *definite* on all interior (i.e, $\not\subset \Gamma(\Omega) \times [0,T]$) faces of each element K where $(\boldsymbol{n}_x, n_t)$ denotes the unit outer normal to K. We denote by $\Gamma_{\text{in}}(K)$ ($\Gamma_{\text{out}}(K)$) the portion of $\Gamma(K)$ for which M is negative (positive) definite. The above explicitness condition will hold if all element faces are inclined sufficiently toward the $\boldsymbol{x}$-hyperplane to make $\| \sum_{i=1}^{N} n_i A_i \| < |n_t|$. The sign of n_t will then indicate the direction of explicitness. In addition to the definiteness condition, we assume the eigenvalues $\lambda(M)$ of M are bounded away from zero:

$$|\lambda(M)| \geq \gamma > 0. \tag{3}$$

With an "explicit" mesh, we can discretize (1)-(2) via the following extension of the discontinuous Galerkin method:

$$a(\boldsymbol{u}_h, \boldsymbol{v}_h)_K = (\boldsymbol{f}, \boldsymbol{v}_h)_K, \qquad \text{all } \boldsymbol{v}_h \in \boldsymbol{S}_h(K), \tag{4}$$

$$a(\boldsymbol{u}, \boldsymbol{v})_K \equiv (\mathcal{L}\boldsymbol{u}, \boldsymbol{v})_K - \int_{\Gamma_{\text{in}}(K)} [\boldsymbol{u}]^T M \boldsymbol{v} + \frac{1}{2} \int_{\Gamma^*(K)} \boldsymbol{u}^T (\mathcal{N} - D) \boldsymbol{v}. \tag{5}$$

The approximation subspace $\boldsymbol{S}_h(K)$ is comprised of polynomials of total degree $\leq n$ over K or is an nth degree tensor product space, $\boldsymbol{u}_h \in \boldsymbol{S}_h(K)$ is the finite element approximation, and $(\ ,\ )_K$ denotes the $L^2(K)$ inner product; $\Gamma^*(K)$ denotes the intersection, if any, of $\Gamma(K)$ with $\Gamma(\Omega) \times [0,T]$. In general, $\boldsymbol{u}_h$ will be discontinuous on interelement boundaries. We denote by $\boldsymbol{u}_h^-$ and $\boldsymbol{u}_h^+$ its upstream and downstream limits, respectively, and use the notation $[\boldsymbol{u}_h] = \boldsymbol{u}_h^+ - \boldsymbol{u}_h^-$.

We will sketch the derivation of the following error estimate for $\boldsymbol{u}_h$ as defined by (4)-(5):

$$|\boldsymbol{u}_h^- - \boldsymbol{u}|^2_{\Gamma_{\text{out}}(\Omega_T)} + \|\boldsymbol{u}_h - \boldsymbol{u}\|^2_{\Omega_T} + h \sum_{K \subset \Omega_T} \|\mathcal{L}_0(\boldsymbol{u}_h - \boldsymbol{u})\|^2_K + \sum_{K \subset \Omega_T} |[\boldsymbol{u}_h - \boldsymbol{u}]|^2_{\Gamma_{\text{in}}(K)} \leq O(h^{2n+1}). \tag{6}$$

This is an extension of the standard error estimate, first given in [6], for the discontinuous Galerkin method. We use the notation $\|\cdot\|_{k,D}$ for the norm on $H^k(D)$, $D \subset \Omega_T$, omitting k when it has value zero, and denote "surface" L^2 norms (e.g., over $\Gamma_{\text{in}}(K)$) by $|\cdot|$. The principal part of $\mathcal{L}$ is denoted by $\mathcal{L}_0$, and C denotes a generic constant independent of h but which, in general, is different at each occurrence.

The estimate (6) is essentially the same as that obtained in [2]. We use a different approach here, however, establishing (6) directly without first showing existence and stability of $\boldsymbol{u}_h$. Also, we employ an additional test function ($\boldsymbol{v}_h = \mathcal{L}_0(\boldsymbol{u}_h - \boldsymbol{u}_I)$, where $\boldsymbol{u}_I$ is an interpolant of $\boldsymbol{u}$) not used in [2], which eliminates the need for a technical assumption made in [2] (that each element be convex or have "sufficiently many" faces in comparison to n, the degree of approximation). It also allows (6) to be obtained with an arbitrary optimal order interpolant.

An outline of the paper is as follows. In §2 we detail the requirements for an explicit mesh, and in §3 we describe some ways to generate such a mesh. In §4 we outline the derivation of the estimate (6) for the discontinuous Galerkin method. Finally, in §5, we briefly consider a pair of continuous explicit finite element methods for (1)-(2), which work well for the simplest case $\Omega \subset R^1$, but have significant shortcomings when generalized to higher dimension.

2 Requirements for explicitness

To elucidate the domain of dependence properties of (1), we consider the homogeneous equation

$$\mathcal{L}_0 \boldsymbol{u} = 0, \qquad \mathcal{L}_0 \equiv \frac{\partial}{\partial t} + \sum_{i=1}^{N} A_i \frac{\partial}{\partial x_i}$$

in a generic polyhedral element K in the interior of the space-time domain. Integrating against $\boldsymbol{u}$, we get

$$\oint_{\Gamma(K)} \boldsymbol{u}^T M \boldsymbol{u} = 0, \qquad M = n_t I + \sum_{i=1}^{N} n_i A_i. \tag{7}$$

We require that M be *definite* on each face of $\Gamma(K)$. This will be the case if K can be chosen so that $\| \sum_{i=1}^{N} n_i A_i \| < |n_t|$ on $\Gamma(K)$. A sufficient condition for this is

$$\frac{|\boldsymbol{n}_x|}{|n_t|} \leq \frac{1}{\sqrt{N} \max_i \|A_i\|}. \tag{8}$$

The sign of n_t will then indicate the inflow and outflow portions of $\Gamma(K)$, and the *flow* direction will be that of increasing t. We also require the inflow and outflow portions of $\Gamma(K)$ to be connected; otherwise explicit development of a solution will not be possible.

Assuming M is definite on each face of $\Gamma(K)$, we may recast (7) as

$$\int_{\Gamma_{\text{out}}(K)} \boldsymbol{u}^T M \boldsymbol{u} = \int_{\Gamma_{\text{in}}(K)} \boldsymbol{u}^T (-M) \boldsymbol{u},$$

where $\Gamma_{\text{out}}(K)$ ($\Gamma_{\text{in}}(K)$) denotes the portion of $\Gamma(K)$ where M ($-M$) is positive definite. Applying (3),

$$|\boldsymbol{u}|^2_{\Gamma_{\text{out}}(K)} \leq C\gamma^{-1} \left(\max_{\Gamma_{\text{in}}(K)} \|M\| \right) |\boldsymbol{u}|^2_{\Gamma_{\text{in}}(K)}.$$

Thus $\boldsymbol{u} \equiv 0$ on $\Gamma_{\text{in}}(K)$ implies $\boldsymbol{u} \equiv 0$ on $\Gamma_{\text{out}}(K)$. Now for an arbitrary point $(\boldsymbol{x}^*, t^*) \in K$, we may construct a smaller polyhedral element $K' \subset K$ such that $\Gamma_{\text{in}}(K') \subset \Gamma_{\text{in}}(K)$ and $(\boldsymbol{x}^*, t^*) \in \Gamma_{\text{out}}(K')$. Integrating against $\boldsymbol{u}$ over K', we infer that $\boldsymbol{u}(\boldsymbol{x}^*, t^*) = 0$ if $\boldsymbol{u} \equiv 0$ on $\Gamma_{\text{in}}(K)$. Thus $\boldsymbol{u} \equiv 0$ on $\Gamma_{\text{in}}(K)$ implies $\boldsymbol{u} \equiv 0$ throughout K.

In a similar way, we may obtain a local stability result for a discrete model of (1). Suppose

$$\mathcal{L}_0 \boldsymbol{u}_h = \boldsymbol{f}_h \quad \text{in } K,$$

where $\boldsymbol{u}_h \in \boldsymbol{S}_h(K)$ is given on $\Gamma_{\text{in}}(K)$. Reasoning as before, we conclude that if $\boldsymbol{f}_h \equiv 0$ in K and $\boldsymbol{u}_h \equiv 0$ on $\Gamma_{\text{in}}(K)$, then $\boldsymbol{u}_h \equiv 0$ in K. Since $\boldsymbol{u}_h$ in K may be regarded as the solution of a linear algebraic system with data $\boldsymbol{f}_h$ and $\boldsymbol{u}_h$ on $\Gamma_{\text{in}}(K)$, we infer that $\|\boldsymbol{u}_h\|_K$ can be bounded by a linear combination of $|\boldsymbol{u}_h|_{\Gamma_{\text{in}}(K)}$ and $\|\boldsymbol{f}_h\|_K$. Applying the appropriate scaling, we get for this bound:

$$\|\boldsymbol{u}_h\|_K \leq C \left(\sqrt{h} |\boldsymbol{u}_h|_{\Gamma_{\text{in}}(K)} + h \|\boldsymbol{f}_h\|_K \right).$$

Equivalently, for any $\boldsymbol{v}_h \in \boldsymbol{S}_h(K)$,

$$\|\boldsymbol{v}_h\|_K \leq C \left(\sqrt{h} |\boldsymbol{v}_h|_{\Gamma_{\text{in}}(K)} + h \|\mathcal{L}_0 \boldsymbol{v}_h\|_K \right). \tag{9}$$

We will use this bound later.

We briefly consider the wave equation example. Here

$$M = \begin{pmatrix} n_t & 0 & -n_1 \\ 0 & n_t & -n_2 \\ -n_1 & -n_2 & n_t \end{pmatrix},$$

whose eigenvalues are $\lambda = n_t, n_t \pm \sqrt{n_1^2 + n_2^2} = n_t \pm |\boldsymbol{n}_x|$. Thus M will be definite if $|\boldsymbol{n}_x| < |n_t|$. Condition (8) is more restrictive: $|\boldsymbol{n}_x| \leq |n_t|/\sqrt{2}$.

3 Mesh construction

We now consider the problem of generating an explicitly configured mesh of polyhedral space-time elements for $\Omega \times [0, T]$. As our starting point, we assume an appropriate face-conforming mesh of elements $\mathcal{T}_h$ is given for the spatial domain Ω. Let $\mathcal{X}_h = \{\boldsymbol{x}_i\}$ denote the nodes of $\mathcal{T}_h$ and $\mathcal{N}(\boldsymbol{x}_i)$ the set of neighboring vertices that share a common element with $\boldsymbol{x}_i$. The space-time mesh will be created incrementally, in the direction of increasing t. Its forward extent at $\boldsymbol{x}_i \in \mathcal{X}$ at any stage in its development will be denoted by $t_{\max}(\boldsymbol{x}_i)$. Each space-time element in our construction will be centered about a particular $\boldsymbol{x}_i$, will have $\boldsymbol{x}_i$ and $\mathcal{N}(\boldsymbol{x}_i)$ as $\boldsymbol{x}$ coordinates of its vertices, and will advance $t_{\max}(\boldsymbol{x}_i)$ to its next value while leaving $t_{\max}(\boldsymbol{x}_j), j \neq i$, unaltered. To elucidate the parallelism possibilities, we shall assign each spatial vertex $\boldsymbol{x}_i \in \mathcal{X}$ a "color" $C(\boldsymbol{x}_i) \in \{1, 2, ...\}$ subject to the condition $\boldsymbol{x}_j \in \mathcal{N}(\boldsymbol{x}_i) \implies C(\boldsymbol{x}_j) \neq C(\boldsymbol{x}_i)$.

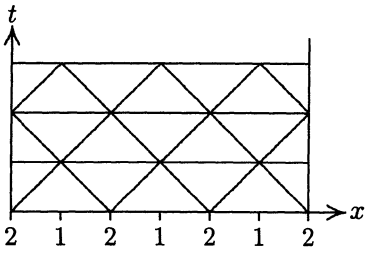


Fig. 3.1a

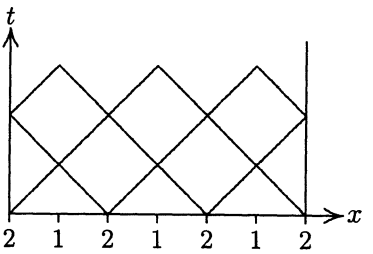


Fig. 3.1b

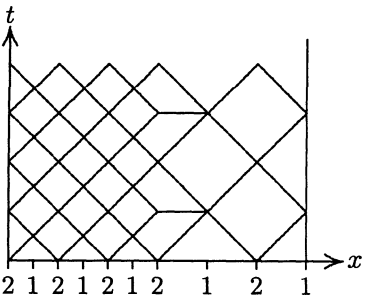


Fig. 3.1c

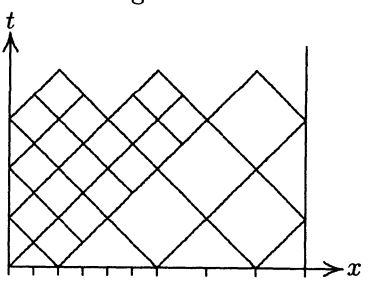


Fig. 3.1d

We first consider the case $\Omega = [0, 1] \subset R^1$, with Ω divided into uniform subintervals of width h. As indicated in Fig. 3.1a, two colors suffice for the spatial nodes $\{\boldsymbol{x}_i\}$. In the first step, we may advance $t_{\max}$ at nodes of color 1 to $t = \Delta t$. Assuming Δt is chosen sufficiently small in comparison to h to ensure explicitness, the PDE solution can be developed concurrently in all such space-time elements centered about vertices of color 1. If in steps 2, 3, 4, 5, 6, 7, ..., we advance $t_{\max}$ at nodes

of color $2, 2, 1, 1, 2, 2, \ldots$ to $t = \Delta t, 2\Delta t, 2\Delta t, 3\Delta t, 3\Delta t, 4\Delta t, \ldots$, respectively, we obtain the space-time mesh shown in Fig. 3.1a. A second alternative is to follow step 1 above by steps $2', 3', 4', 5', \ldots$ in which $t_{\max}$ at nodes of color $2, 1, 2, 1, \ldots$ is advanced to $t = 2\Delta t, 3\Delta t, 4\Delta t, 5\Delta t, \ldots$, resulting in the mesh of Fig. 3.1b. This mesh is comprised of a single generic element, a rhombus, and is twice as efficient at "consuming space" as the first scheme. One could, of course, bring the solution back to a common t at a subsequent time if desired.

Next suppose our spatial discretization of $\Omega = [0, 1]$ is nonuniform. We consider a simple case of a two-for-one mesh refinement in Fig 3.1c. The space-time mesh depicted there results if, in steps $1, 2, 3, 4, \ldots$, nodes of color $1, 2, 1, 2, \ldots$ are advanced to their maximum t values consistent with explicitness, but coarse mesh nodes are not updated in steps 3, 4 (also $7, 8, \ldots$). In general, one would like the frequency of update to vary inversely to the spatial grid size. This mesh illustrates the possibility of achieving two potentially desirable objectives: an explicit mesh, and a locally varying time step tailored to the degree of spatial refinement needed. (The more common, more rigid, alternative is to not have any spatial variation in time step). Another possible mesh generation technique would be to start with a uniform coarse mesh, consisting of congruent rhombuses and refine on a four-for-one basis where needed. Fig. 3.1d illustrates this for a case of a moving mesh.

We now turn to the more interesting case $\Omega \subset R^2$. Suppose, initially, that our spatial mesh consists of equilateral triangles of side length h. This illustrated in Fig. 3.2, where a 3-coloring of the corresponding nodes is also indicated. In analogy with the $\Omega \subset R^1$ case, we may, in steps 1,2,3,4,5,6,..., advance nodes of color 1,2,3,3,2,1, ... to $t = \Delta t, \Delta t, \Delta t, 2\Delta t, 2\Delta t, 2\Delta t, \ldots$. Now, however, the elements so generated in each step are each unions of 6 tetrahedra and have either 7 or 12 (in steps $2, 5, 8, \ldots$) faces. For this scheme $|n_x|/|n_t| \leq (2\Delta t)/(\sqrt{3}h)$.

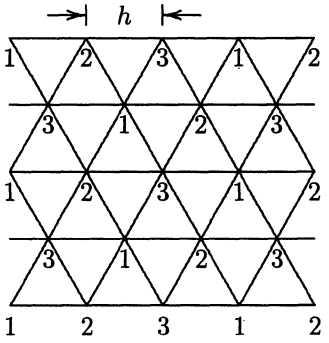


Fig. 3.2

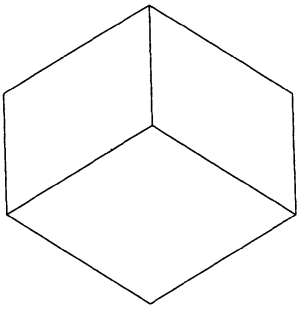


Fig. 3.3

A second alternative is: in steps 1,2,3,4,5,6,..., advance nodes of color 1,2,3,1,2,3, ... to $t = \Delta t, 2\Delta t, 3\Delta t, 4\Delta t, 5\Delta t, 6\Delta t, \ldots$. For this scheme, $|n_x|/|n_t| \leq 2\Delta t/h$; thus the maximum time step Δt for this scheme must be smaller than the previous one. Overall, however, the second scheme uses fewer elements to fill a given volume in the space-time domain. Moreover, the second scheme uses a a single generic element (apart from boundary effects) which is, in fact, a hexahedron with two opposite vertices lying along a line of constant x, parallel to the t-axis, as pictured in Fig. 3.3. Thus it may be viewed as a higher dimensional analog of the mesh depicted in Fig. 3.1b. Perhaps the simplest way to generate an explicit mesh for the case $\Omega \subset R^2$ would be use a coarse mesh of such hexahedra and then refine locally on an eight-for-one basis as appropriate. This scheme, as well as the first one, generalize readily to higher dimension.

4 Analysis

We shall restrict our attention here to interior elements K for which $\Gamma^*(K) = \emptyset$ in (5). The more general case is dealt with in [2]. We also assume the mesh is quasiuniform and nondegenerate (allowing the use of inverse inequalities).

We begin by giving a pair of identities for the bilinear form $a(u, v)$ defined in (5). By integrating $a(u, v)$ by parts, then performing some manipulations on the $\Gamma_{\text{in}}(K)$ integrals, we get:

$$a(u, v)_K = -a(v, u)_K + ((B + B^*)u, v)_K + \oint_{\Gamma(K)} (u^-)^T M v^- - \int_{\Gamma_{\text{in}}(K)} [u]^T M [v]. \tag{10}$$

Taking $v = u$ in (10), then using (3), we obtain:

$$\begin{aligned} a(u, u)_K &= \frac{1}{2} \oint_{\Gamma(K)} (u^-)^T M u^- + \frac{1}{2} \int_{\Gamma_{\text{in}}(K)} [u]^T (-M)[u] \\ &\quad + \frac{1}{2}((B + B^*)u, u)_K \\ &\geq \frac{1}{2} \oint_{\Gamma(K)} (u^-)^T M u^- + \frac{\gamma}{2} |[u]|^2_{\Gamma_{\text{in}}(K)} + \frac{1}{2}((B + B^*)u, u)_K. \end{aligned} \tag{11}$$

We now assume the continuous problem and its discrete counterpart have solutions u and u_h, respectively, and estimate the difference between the two. From the derived estimate, it will follow that if the continuous problem has a solution, then u_h is well-defined for h sufficiently small. It will be convenient to use an interpolant $u_I \in S_h(K)$ for u that, we assume, will give optimal order accuracy if u is sufficiently smooth, i.e.,

$$\|u_I - u\|_K + \sqrt{h} |u_I - u|_{\Gamma(K)} \leq C h^{n+1} \|u\|_{n+1,K}. \tag{12}$$

Subtracting $a(u, v_h)_K = (f, v_h)_T$ from (4) and introducing u_I, we obtain:

$$a(e_h, v_h)_K = a(u - u_I, v_h)_K, \qquad e_h \equiv u_h - u_I. \tag{13}$$

In what follows, we shall denote by $I(K)$ the "inflow" elements to K, lying immediately upstream from K.

The basic ingredients of the error estimate for u_h are expressed in the following:

Lemma 1. *(i) The choice $v_h = e_h$ in (13) yields, for arbitrary $\epsilon > 0$:*

$$\begin{aligned}\frac{1}{2}\oint_{\Gamma(K)} \left((u_h - u)^-\right)^T M(u_h - u)^- + \frac{\gamma}{2}|[e_h]|^2_{\Gamma_{in}(K)} \\ \leq \epsilon\left(|[e_h]|^2_{\Gamma_{in}(K)} + h\|\mathcal{L}_0 e_h\|^2_K\right) + C\left(\|e_h\|^2_K + \epsilon^{-1}h^{2n+1}\|u\|^2_{n+1,K\cup I(K)}\right).\end{aligned} \tag{14}$$

(ii) The choice $v_h = \mathcal{L}_0 e_h$ in (13) yields:

$$\|\mathcal{L}_0 e_h\|^2_K \leq C\left(\|e_h\|^2_K + h^{-1}|[e_h]|^2_{\Gamma_{in}(K)} + h^{2n}\|u\|^2_{n+1,K\cup I(K)}\right). \tag{15}$$

(iii) e_h satisfies:

$$\|e_h\|^2_K \leq C\left(h|e_h^-|^2_{\Gamma_{in}(K)} + h|[e_h]|^2_{\Gamma_{in}(K)} + h^2\|\mathcal{L}_0 e_h\|^2_K\right). \tag{16}$$

Proof. (i) By taking $v_h = e_h$ in (13), applying (10) and (11), then the Schwarz inequality, arithmetic-geometric mean inequality, and inverse inequalities, we get the following bounds:

$$a(e_h, e_h)_K \geq \frac{1}{2}\oint_{\Gamma(K)} (e_h^-)^T M e_h^- + \frac{\gamma}{2}|[e_h]|^2_{\Gamma_{in}(K)} - C\|e_h\|^2_K,$$

$$a(u - u_I, e_h)_K = -\left((\mathcal{L}_0 e_h + Be_h, u - u_I)_K - \int_{\Gamma_{in}(K)} [e_h]^T M(u - u_I)^+\right)$$

$$+ ((B + B^*)(u - u_I), e_h)_K + \oint_{\Gamma(K)} \left((u - u_I)^-\right)^T M e_h^-$$

$$- \int_{\Gamma_{in}(K)} [u - u_I]^T M[e_h]$$

$$\leq \oint_{\Gamma(K)} \left((u - u_I)^-\right)^T M e_h^- + \epsilon\left(|[e_h]|^2_{\Gamma_{in}(K)} + \|e_h\|^2_K + h\|\mathcal{L}_0 e_h\|^2_K\right)$$

$$+ C\epsilon^{-1}\left(h^{-1}\|u - u_I\|^2_K + |[u_h - u_I]|^2_{\Gamma_{in}(K)} + |(u - u_I)^+|^2_{\Gamma_{in}(K)}\right).$$

Combining these bounds, then completing the square on the $\Gamma(K)$ integral, we get (14)

(ii) Using similar techniques, for $v_h = \mathcal{L}_0 e_h$ in (13), we get:

$$\begin{aligned} a(e_h,\mathcal{L}_0 e_h)_K &= (\mathcal{L}_0 e_h + Be_h, \mathcal{L}_0 e_h)_K - \int_{\Gamma_{\text{in}}(K)} [e_h]^T M(\mathcal{L}_0 e_h)^+ \\ &\geq \|\mathcal{L}_0 e_h\|_K^2 - \|B\|\|e_h\|_K \|\mathcal{L}_0 e_h\|_K - C|[e_h]|_{\Gamma_{\text{in}}(K)} \frac{\|\mathcal{L}_0 e_h\|_K}{\sqrt{h}} \\ &\geq \frac{1}{2}\|\mathcal{L}_0 e_h\|_K^2 - C(\|e_h\|_K^2 + h^{-1}|[e_h]|_{\Gamma_{\text{in}}(K)}^2), \end{aligned}$$

$$\begin{aligned} a(\boldsymbol{u} - \boldsymbol{u}_I, \mathcal{L}_0 e_h)_K &= (\mathcal{L}(\boldsymbol{u} - \boldsymbol{u}_I), \mathcal{L}_0 e_h)_K - \int_{\Gamma_{\text{in}}(K)} [\boldsymbol{u} - \boldsymbol{u}_I]^T M(\mathcal{L}_0 e_h)^+ \\ &\leq C\|\boldsymbol{u} - \boldsymbol{u}_I\|_{1,K} \|\mathcal{L}_0 e_h\|_K + C|[\boldsymbol{u} - \boldsymbol{u}_I]|_{\Gamma_{\text{in}}(K)} \frac{\|\mathcal{L}_0 e_h\|_K}{\sqrt{h}} \\ &\leq \frac{1}{4}\|\mathcal{L}_0 e_h\|_K^2 + C\left(\|\boldsymbol{u} - \boldsymbol{u}_I\|_{1,K}^2 + h^{-1}|[\boldsymbol{u} - \boldsymbol{u}_I]|_{\Gamma_{\text{in}}(K)}^2\right). \end{aligned}$$

The result of these bounds is (15)

(iii) We may provide for a jump discontinuity in v_h across $\Gamma_{\text{in}}(K)$ in (9) by writing $|v_h^+|_{\Gamma_{\text{in}}(K)} \leq |v_h^-|_{\Gamma_{\text{in}}(K)} + |[v_h]|_{\Gamma_{\text{in}}(K)}$. Applying the resulting bound to u_h, we get (16)

Multiplying (14)-(16) by $1, \mu h, \nu$, respectively, then adding, then applying the bound $|e_h^-|_{\Gamma_{\text{in}}(K)} \leq |(u_h - u)^-|_{\Gamma_{\text{in}}(K)} + Ch^{n+1/2}\|u\|_{n+1,I(K)}$, gives:

$$\begin{aligned} \frac{1}{2}\oint_{\Gamma(K)} \left((u_h - u)^-\right)^T M(u_h - u)^- + (\frac{\gamma}{2} - \epsilon - C\mu - C\nu h)|e_h]|_{\Gamma_{\text{in}}(K)}^2 \\ +(\mu h - \epsilon h - C\nu h^2)\|\mathcal{L}_0 e_h\|_K^2 + (\nu - C - C\mu h)\|e_h\|_K^2 \\ \leq C\left(h|(u_h - u)^-|_{\Gamma_{\text{in}}(K)}^2 + (1 + \epsilon^{-1})h^{2n+1}\|u\|_{n+1,K\cup I(K)}^2\right). \end{aligned}$$

We next take $\mu < \gamma/4C$ to allow coercivity of $|[e_h]|_{\Gamma_{\text{in}}(K)}^2$, then take $\nu > C\mu h + C + 1$ to coerce $\|e_h\|_K^2$, then choose ϵ small enough to coerce $|[e_h]|_{\Gamma_{\text{in}}(K)}^2$ and $\|\mathcal{L}_0 e_h\|_K^2$ for h sufficiently small. We can write the result as follows:

Lemma 2. *There exist positive constants γ_1 and γ_2 such that for h sufficiently small,*

$$\begin{aligned} \oint_{\Gamma(K)} \left((u_h - u)^-\right)^T M(u_h - u)^- + \|e_h\|_K^2 + \gamma_1|[e_h]|_{\Gamma_{in}(K)}^2 + \gamma_2 h\|\mathcal{L}_0 e_h\|_K^2 \\ \leq C\left(h|(u_h - u)^-|_{\Gamma_{in}(K)}^2 + h^{2n+1}\|u\|_{n+1,K\cup I(K)}^2\right). \end{aligned} \tag{17}$$

Assuming sufficient smoothness in u and applying this bound over all $K \subset \Omega_T$ then yields (3)(cf. [2]).

5 Continuous explicit finite element methods

We briefly consider the possibility of explicitly generating a continuous finite element method over an appropriate mesh. The generic form of such a method is

$$\begin{aligned} &u_h \in S_h(K) \\ &(\mathcal{L}u_h, v_h)_K = (f, v_h)_K, \qquad \text{all } v_h \in T_h(K) \end{aligned}$$

Here $T_h(K)$ must have dimension less than that of $S_h(K)$ because u_h will already be known on $\Gamma_{\text{in}}(K)$ at the time when it is to be computed in K. A potential advantage of a continuous method is the smaller number of degrees of freedom in u_h, hence fewer unknowns to be solved for.

We first mention a method [9] which can be applied over a mesh of triangles like that depicted in Fig. 3.1a.

$$\begin{aligned} &u_h \in P_n(K) \\ &(\mathcal{L}u_h, v_h)_K = (f, v_h)_K, \qquad \text{all } v_h \in P_{n-\rho(K)}(K). \end{aligned} \tag{18}$$

Here P_n consists of polynomials of total degree $\leq n$ and $\rho(K)$ denotes the number of inflow sides that K has (either one or two). This method typically gives $O(h^{n+1})$ convergence, like the discontinuous Galerkin method; an analysis appears in [3].

The method (18) extends directly to higher dimension over simplices K. For the case $\Omega \subset R^2$, the elements are tetrahedra which may have either 1, 2, or 3 inflow faces (i.e., $\rho(K) = 1$, 2 or 3). Thus there are three possible test spaces for (18), and, not surprisingly, no analysis. In addition, n must be at least 2 (otherwise two of the three possible test spaces in (18) will be void), and an explicit tetrahedral mesh seems impractical to construct and manage. Thus (18) does not look promising for $N \geq 2$.

We also mention a continuous method for $\Omega \subset R^1$ due to Winther [10], which can be applied over a mesh of parallelograms like that depicted in Fig. 3.1b. It is:

$$\begin{aligned} &u_h \in \Pi_n(K) \\ &(\mathcal{L}u_h, v_h)_K = (f, v_h)_K, \qquad \text{all } v_h \in \Pi_{n-1}(K). \end{aligned} \tag{19}$$

Here $\Pi_n(K)$ is a tensor product space of polynomials of degree $\leq n$ in coordinates ξ, η aligned with the parallelogram sides. Optimal order error estimates are derived in [10]. This method, too, extends immediately to higher dimension. However, a simple calculation reveals that for the simplest case of (1), $u_t = 0$, in two space dimensions and time, with $n = 1$ (linear approximation), (19) has an algebraic instability arising from a nondecaying spurious root of multiplicity 2. This casts doubt on the usefullness of (19) for $N \geq 2$.

By contrast, the discontinuous Galerkin method is stable regardless of N and very flexible in terms of applicability.

References

1. Du, Q., Gunzburger, M., Layton, W.: A low dispersion, high accuracy finite element method for first order hyperbolic systems in several space variables. Hyperbolic partial differential equations V, Comput. Math. Appl. **15** (1988) 447–457

2. Falk, R.S., Richter, G.R.: Explicit finite element methods for symmetric hyperbolic equations. SIAM J. Numer. Anal., to appear.
3. Falk, R.S., Richter, G.R.: Analysis of a continuous finite element method for hyperbolic equations. SIAM J. Numer. Anal. **24** (1987) 257–278
4. Friedrichs, K. O.: Symmetric positive linear differential equations. Comm. Pure Appl. Math. **11** (1958) 333–418
5. Johnson, C., Nävert, U., Pitkäranta, J.: Finite element methods for linear hyperbolic problems. Comput. Methods Appl. Mech. Engrg. **45** (1984) 285–312
6. Johnson, C., Pitkäranta, J.: An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. Math. Comp. **46** (1986) 1–26
7. Layton, W.: High-accuracy finite-element methods for positive symmetric systems. Hyperbolic partial differential equations III, Comput. Math. Appl. Part A **12** (1986) 565–579
8. Lesaint, P.: Finite element methods for symmetric hyperbolic equations. Numer. Math. **21** (1973) 244–255
9. Reed, W. H., Hill, T. R.: Triangular mesh methods for the neutron transport equation. Los Alamos Scientific Laboratory Technical Report LA-UR-73-479 (1973)
10. Winther, R.: A stable finite element method for first-order hyperbolic systems Math. Comp. **36** (1981) 65–86

hp-DGFEM for Partial Differential Equations with Nonnegative Characteristic Form

Endre Süli[1], Christoph Schwab[2], and Paul Houston[1] *

[1] University of Oxford, Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, UK
[2] Seminar for Applied Mathematics, ETH Zürich, CH-8092 Zürich, Switzerland

Abstract. We develop the error analysis for the *hp*-version of a discontinuous finite element approximation to second-order partial differential equations with nonnegative characteristic form. This class of equations includes classical examples of second-order elliptic and parabolic equations, first-order hyperbolic equations, as well as equations of mixed type. We establish an a priori error bound for the method which is of optimal order in the mesh size h and 1 order less than optimal in the polynomial degree p. In the particular case of a first-order hyperbolic equation the error bound is optimal in h and $1/2$ an order less than optimal in p.

1 Introduction

Discontinuous Galerkin Finite Element Methods (DGFEM) were introduced over quarter of a century ago for the numerical solution of first-order hyperbolic problems [14,11] and as nonstandard techniques for the approximation of second-order elliptic equations [12] (see also [13] for a historical survey). Although subsequently much of the research in the field of numerical analysis of partial differential equations has concentrated on the development and the analysis of conforming finite element methods for self-adjoint elliptic problems, stabilised continuous finite element methods for convection-diffusion equations, and finite difference and finite volume methods for hyperbolic problems, recent years have witnessed renewed interest in discontinuous schemes. This paradigm shift was stimulated by several factors: the desire to handle, within the finite element framework, nonlinear hyperbolic problems (see [6] and [7]) which are known to exhibit discontinuous solutions even when the data are perfectly smooth; the need to treat convection-dominated diffusion problems without excessive numerical stabilisation; the computational convenience of discontinuous finite element methods due to a large degree of locality; and the necessity to accommodate high-order *hp*-adaptive finite element discretisations in a flexible manner (see [5]).

The aim of this paper is to extend the error analysis of the *hp*-DGFEM, developed in our earlier work [8] for first-order hyperbolic equations, to general second-order partial differential equations with nonnegative characteristic form. In [8] an error bound, optimal both in terms of the local mesh

* E. Süli and P. Houston acknowledge the financial support of the EPSRC (Grant GR/K76221).

size h and the local polynomial degree p, was derived for the hp-DGFEM supplemented by a streamline-diffusion type stabilisation involving a stabilisation parameter δ of size h/p. Here, we focus on the more subtle situation when $\delta = 0$, corresponding to no stabilisation. By exploiting theoretical tools similar to those in [8], we derive an error bound for the resulting scheme that is of optimal order in terms of the mesh size h and 1 order less than optimal in the polynomial degree p. For convection-dominated diffusion equations, suboptimality in p is compensated by the fact that the leading term in the error bound is multiplied by a small number, proportional to the square root of the norm of the diffusion matrix. Indeed, in the case of a first-order hyperbolic equation, our error bound collapses to one that is h-optimal, with a loss of only $1/2$ an order in p. The approximation technique adopted in the present paper involves a discontinuity-penalisation device based on the ideas of Nitsche [12], Wheeler [18] and Arnold [1], albeit with a slight (but important) modification which permits us to pass to the hyperbolic limit with inactive discontinuity-penalisation. The error analysis of the hp-DGFEM discretisation considered here can also be viewed as an extension of the work of Baumann [3], Oden, Babuška and Baumann [13], and Riviere and Wheeler [15] for a self-adjoint elliptic problem.

2 Model Problem and Discretisation

Given that Ω is a bounded Lipschitz domain in $\mathbb{R}^d$, $d \geq 2$, we consider the linear second-order partial differential equation

$$\mathcal{L}u \equiv -\sum_{i,j=1}^{d} \partial_j \left(a_{ij}(x)\, \partial_i u\right) + \sum_{i=1}^{d} b_i(x)\, \partial_i u + c(x)u = f(x) \ , \qquad (1)$$

where f is a real-valued function belonging to $L^2(\Omega)$, and the real-valued coefficients a, b, c have the following properties:

$$\begin{aligned} a(x) &= \{a_{ij}(x)\}_{i,j=1}^{d} \in L^\infty(\Omega)^{d\times d}_{\mathrm{sym}} \ , \\ b(x) &= \{b_i(x)\}_{i=1}^{d} \in W^{1,\infty}(\Omega)^d, \quad c(x) \in L^\infty(\Omega) \ . \end{aligned} \qquad (2)$$

We shall suppose throughout that the characteristic form associated with the principal part of the differential operator $\mathcal{L}$ is nonnegative; namely,

$$\boldsymbol{\xi}^T a(x)\, \boldsymbol{\xi} \geq 0 \quad \forall \boldsymbol{\xi} \in \mathbb{R}^d \text{ and a.e. } x \in \bar{\Omega} \ . \qquad (3)$$

For simplicity, we shall assume that the entries of the matrix a are piecewise continuous on $\bar{\Omega}$; this hypothesis is sufficiently general to cover most situations of practical relevance. Let $\boldsymbol{\mu}(x) = \{\mu_i(x)\}_{i=1}^{d}$ denote the unit outward normal vector to $\Gamma = \partial\Omega$ at $x \in \Gamma$ and define the following subsets of Γ:

$$\Gamma_0 = \{x \in \Gamma : \ \boldsymbol{\mu}^T a(x) \boldsymbol{\mu} > 0\} \ ,$$

$$\Gamma_- = \{x \in \Gamma \backslash \Gamma_0 : \ \boldsymbol{b} \cdot \boldsymbol{\mu} < 0\} \quad \text{and} \quad \Gamma_+ = \{x \in \Gamma \backslash \Gamma_0 : \ \boldsymbol{b} \cdot \boldsymbol{\mu} \geq 0\} \ .$$

The sets $\Gamma_{\mp}$ will be referred to as the inflow and outflow boundary, respectively. With these definitions we have that $\Gamma = \Gamma_0 \cup \Gamma_- \cup \Gamma_+$. We shall further decompose Γ_0 into two connected parts, $\Gamma_{\rm D}$ and $\Gamma_{\rm N}$, and supplement the partial differential equation (1) with the following boundary conditions:

$$u = g_{\rm D} \quad \text{on } \Gamma_{\rm D} \cup \Gamma_- \quad \text{and} \quad \boldsymbol{\mu}^T a \nabla u = g_{\rm N} \text{ on } \Gamma_{\rm N} \ . \tag{4}$$

We note that (1), (4) includes a range of physically relevant problems, such as the mixed boundary value problem for an elliptic equation corresponding to the case when (3) holds with strict inequality, as well as the case of a linear transport problem associated with the choice of $a \equiv 0$ on $\bar{\Omega}$. Our aim here is to develop, in a unified manner, the a priori error analysis of the *hp*-version of a discontinuous finite element approximation to (1), (4).

2.1 Finite element spaces

Let $\mathcal{T}$ be a subdivision of Ω into open element domains κ such that $\bar{\Omega} = \cup_{\kappa \in \mathcal{T}} \bar{\kappa}$. We shall assume that the family of subdivisions $\mathcal{T}$ is shape regular and that each $\kappa \in \mathcal{T}$ is a smooth bijective image of a fixed master element $\hat{\kappa}$, that is, $\kappa = \mathcal{F}_\kappa(\hat{\kappa})$ for all $\kappa \in \mathcal{T}$ where $\hat{\kappa}$ is either the open unit simplex or the open unit hypercube in $\mathbb{R}^d$. For an integer $r \geq 1$, we denote by $\mathcal{P}_r(\hat{\kappa})$ the set of polynomials of total degree $< r$ on $\hat{\kappa}$; when $\hat{\kappa}$ is the unit hypercube, we also consider $\mathcal{Q}_r(\hat{\kappa})$, the set of all tensor-product polynomials of degree $< r$ in each coordinate direction. Thus, to $\kappa \in \mathcal{T}$ we assign an integer $p_\kappa \geq 1$, collect the p_κ and $\mathcal{F}_\kappa$ in the vectors $\mathbf{p} = \{p_\kappa : \kappa \in \mathcal{T}\}$ and $\mathbf{F} = \{\mathcal{F}_\kappa : \kappa \in \mathcal{T}\}$, respectively, and consider the finite element space

$$S^{\mathbf{p}}(\Omega, \mathcal{T}, \mathbf{F}) = \{u \in L^2(\Omega) : \ u|_\kappa \circ \mathcal{F}_\kappa \in \mathcal{R}_{p_\kappa}(\hat{\kappa}) \quad \forall \kappa \in \mathcal{T}\} \ ,$$

where $\mathcal{R}$ is either $\mathcal{P}$ or $\mathcal{Q}$. Given the subdivision $\mathcal{T}$ and $s > 0$, the associated broken Sobolev space $H^s(\Omega, \mathcal{T})$ is defined by

$$H^s(\Omega, \mathcal{T}) = \prod_{\kappa \in \mathcal{T}} H^s(\kappa) = \{u \in L^2(\Omega) : \ u|_\kappa \in H^s(\kappa) \quad \forall \kappa \in \mathcal{T}\} \ .$$

In the next section, we formulate the *hp*-DGFEM approximation of (1), (4).

2.2 The numerical method

Discretisation of the Low-Order Terms. Let us begin by considering the first-order partial differential operator $\mathcal{L}_b$ defined by

$$\mathcal{L}_b w = \boldsymbol{b} \cdot \nabla w + cw \ .$$

Given that κ is an element in the partition $\mathcal{T}$, we denote by $\partial\kappa$ the union of open faces of κ. This is non-standard notation in that $\partial\kappa$ is a subset of the boundary of κ. Let $x \in \partial\kappa$ and suppose that $\boldsymbol{\mu}(x)$ denotes the unit outward

normal vector to $\partial\kappa$ at x. With these conventions, we define the inflow and outflow parts of $\partial\kappa$, respectively, by

$$\partial_-\kappa = \{x \in \partial\kappa : \ \boldsymbol{b}(x)\cdot\boldsymbol{\mu}(x) < 0\} \ , \qquad \partial_+\kappa = \{x \in \partial\kappa : \ \boldsymbol{b}(x)\cdot\boldsymbol{\mu}(x) \geq 0\} \ .$$

For each $v \in H^1(\Omega, \mathcal{T})$ and any $\kappa \in \mathcal{T}$, we denote by v^+ the interior trace of v on $\partial\kappa$ (the trace taken from within κ). Now consider an element κ such that the set $\partial_-\kappa\backslash\Gamma_-$ is nonempty; then for each $x \in \partial_-\kappa\backslash\Gamma_-$ (with the exception of a set of $(d-1)$-dimensional measure zero) there exists a unique element κ', depending on the choice of x, such that $x \in \partial_+\kappa'$. If $\partial_-\kappa\backslash\Gamma_-$ is nonempty for some $\kappa \in \mathcal{T}$, then we can also define the outer trace v^- of v on $\partial_-\kappa\backslash\Gamma_-$ relative to κ as the inner trace v^+ relative to those elements κ' for which $\partial_+\kappa'$ has intersection with $\partial_-\kappa\backslash\Gamma_-$ of positive $(d-1)$-dimensional measure. Further, we introduce the oriented jump of v across $\partial_-\kappa\backslash\Gamma_-$:

$$\lfloor v \rfloor = v^+ - v^- \ .$$

Supposing that $v, w \in H^1(\Omega, \mathcal{T})$, we define, as in [10], for example,

$$B_b(w,v) = \sum_{\kappa\in\mathcal{T}} \int_\kappa (\mathcal{L}_b w) v \ \mathrm{d}x \tag{5}$$
$$- \sum_{\kappa\in\mathcal{T}} \int_{\partial_-\kappa\backslash\Gamma_-} (\boldsymbol{b}\cdot\boldsymbol{\mu}) \lfloor w \rfloor \, v^+ \ \mathrm{d}s - \sum_{\kappa\in\mathcal{T}} \int_{\partial_-\kappa\cap\Gamma_-} (\boldsymbol{b}\cdot\boldsymbol{\mu}) w^+ \, v^+ \ \mathrm{d}s \ ,$$

and we put

$$\ell_b(v) = \sum_{\kappa\in\mathcal{T}} \int_\kappa f v \, \mathrm{d}x - \sum_{\kappa\in\mathcal{T}} \int_{\partial_-\kappa\cap\Gamma_-} (\boldsymbol{b}\cdot\boldsymbol{\mu}) \, g v^+ \ \mathrm{d}s \ .$$

Discretisation of the Leading Term. Let us suppose that the elements in the partition have been numbered in a certain way, regardless of the flow direction. We denote by $\mathcal{E}$ the set of element interfaces (edges for $d = 2$ or faces for $d = 3$) associated with the subdivision $\mathcal{T}$. Since hanging nodes are permitted in the DGFEM, $\mathcal{E}$ will be understood to consist of the smallest interfaces in $\partial\kappa$. With this notation, let Γ_{int} denote the union of all interfaces $e \in \mathcal{E}$. Given that $e \in \mathcal{E}$, there exist indices i and j such that $i > j$ and κ_i and κ_j share the interface e; we define the (numbering-dependent) jump of $v \in H^1(\Omega, \mathcal{T})$ across e and the mean value of v on e, respectively, by

$$[v] = v|_{\partial\kappa_i\cap e} - v|_{\partial\kappa_j\cap e} \quad \text{and} \quad \langle v \rangle = \frac{1}{2}\left(v|_{\partial\kappa_i\cap e} + v|_{\partial\kappa_j\cap e}\right) \ .$$

We note that, in general, $[v]$ is distinct from $\lfloor v \rfloor$ in that the latter depends on the sign of the normal flux on an element boundary, while the former is only dependent on the element numbering. With each face $e \in \mathcal{E}$ we associate

the normal vector $\boldsymbol{\nu}$ which points from κ_i to κ_j; on boundary faces, we put $\boldsymbol{\nu} = \boldsymbol{\mu}$. Finally, we introduce, as in [13], the bilinear form

$$B_a(w,v) = \sum_{\kappa\in\mathcal{T}} \int_\kappa a(x)\nabla w\cdot\nabla v\,\mathrm{d}x + \int_{\Gamma_\mathrm{D}} \{w((a\nabla v)\cdot\boldsymbol{\nu}) - ((a\nabla w)\cdot\boldsymbol{\nu})v\}\,\mathrm{d}s$$
$$+ \int_{\Gamma_\mathrm{int}} \{[w]\langle(a\nabla v)\cdot\boldsymbol{\nu}\rangle - \langle(a\nabla w)\cdot\boldsymbol{\nu}\rangle[v]\}\,\mathrm{d}s\ , \tag{6}$$

associated with the principal part of the partial differential operator $\mathcal{L}$, and the linear functional

$$\ell_a(v) = \int_{\Gamma_\mathrm{D}} g_\mathrm{D}((a\nabla v)\cdot\boldsymbol{\nu})\,\mathrm{d}s + \int_{\Gamma_\mathrm{N}} g_\mathrm{N} v\,\mathrm{d}s\ .$$

Discontinuity-Penalisation Term. Let $\bar{a} = ||a||_2$, with $||\cdot||_2$ denoting the matrix norm subordinate to the l^2 vector norm on $\mathbb{R}^d$, and let $\bar{a}_\kappa = \bar{a}|_\kappa$. To each e in $\mathcal{E}$ which is a common face of elements κ_i and κ_j in $\mathcal{T}$ we assign the nonnegative function $\langle\bar{a}p^2\rangle_e = (p^2_{\kappa_i}\bar{a}_{\kappa_i}|_e + p^2_{\kappa_j}\bar{a}_{\kappa_j}|_e)/2$. Letting $\mathcal{E}_\mathrm{D}$ denote the set of all faces contained in Γ_D, to each $e\in\Gamma_\mathrm{D}$ we assign the element $\kappa\in\mathcal{T}$ with that face and define $\langle\bar{a}p^2\rangle_e = p^2_\kappa\bar{a}_\kappa|_e$. Consider the function σ defined on $\Gamma_\mathrm{int}\cup\Gamma_\mathrm{D}$ by $\sigma(x) = K\langle\bar{a}p^2\rangle_e/|e|$ for $x\in e$ and $e\in\mathcal{E}\cup\mathcal{E}_\mathrm{D}$, where $|e| = \mathrm{meas}_{d-1}(e)$ and K is a positive constant (whose value is irrelevant for the present analytical study, so we put $K=1$), and introduce the bilinear form and the linear functional, respectively, by

$$B_s(w,v) = \int_{\Gamma_\mathrm{D}} \sigma wv\,\mathrm{d}s + \int_{\Gamma_\mathrm{int}} \sigma[w][v]\,\mathrm{d}s\ ,\quad \ell_s(v) = \int_{\Gamma_\mathrm{D}} \sigma g_\mathrm{D} v\,\mathrm{d}s\ . \tag{7}$$

We highlight the fact that since the weight-function σ involves the norm of the matrix a, in the hyperbolic limit of $a\equiv 0$ the bilinear form $B_s(\cdot,\cdot)$ and the linear functional ℓ_s both vanish. This is a desirable property, since linear hyperbolic equations may possess solutions that are discontinuous across characteristic hypersurfaces, and penalising discontinuities across faces which belong to these would seem unnatural.

It is also worth noting here that, conceptually, the bilinear form $B_a(\cdot,\cdot)+B_s(\cdot,\cdot)$ should be regarded as a single entity, rather than a sum of two separate bilinear forms; the same comment applies to $\ell_a(\cdot)+\ell_s(\cdot)$. Although more convenient from the point of view of the presentation, separation into B_a, ℓ_a on the one hand and B_s, ℓ_s on the other is somewhat artificial and can only be justified on historical grounds (see [12,18,1]).

Definition of the Method. Finally, we define the bilinear form $B_\mathrm{DG}(\cdot,\cdot)$ and the linear functional $\ell_\mathrm{DG}(\cdot)$, respectively, by

$$B_\mathrm{DG}(w,v) = B_a(w,v) + B_b(w,v) + B_s(w,v)\ ,$$
$$\ell_\mathrm{DG}(v) = \ell_a(v) + \ell_b(v) + \ell_s(v)\ .$$

The *hp*-DGFEM approximation of (1), (4) is: find $u_{\mathrm{DG}} \in S^{\mathbf{p}}(\Omega, \mathcal{T}, \mathbf{F})$ such that

$$B_{\mathrm{DG}}(u_{\mathrm{DG}}, v) = \ell_{\mathrm{DG}}(v) \qquad \forall v \in S^{\mathbf{p}}(\Omega, \mathcal{T}, \mathbf{F}) \ . \tag{8}$$

In the next section we state the key properties of this method. Before we do so, however, we note that in the definitions of the bilinear forms and linear functionals above and in the arguments which follow it has been tacitly assumed that $a \in C(\kappa)$ for each $\kappa \in \mathcal{T}$, that the fluxes $(a\nabla u)\cdot\boldsymbol{\nu}$ and $(\boldsymbol{b}\cdot\boldsymbol{\mu})u$ are continuous across element interfaces, and that u is continuous in an (open) neighbourhood of the subset of Ω where a is not identically equal to zero. If the problem under consideration violates these properties, the scheme and the analysis have to be modified accordingly.

3 Analytical Results

Our first result concerns the positivity of the bilinear form $B_{\mathrm{DG}}(\cdot,\cdot)$ and the existence and uniqueness of a solution to (8).

Theorem 1. *Suppose that, in addition to the conditions (2) and (3), the function* $\gamma \equiv c - \frac{1}{2}\nabla\cdot\boldsymbol{b}$ *is nonnegative on* $\bar{\Omega}$. *Then,*

$$|||w|||^2_{\mathrm{DG}} \equiv B_{\mathrm{DG}}(w,w) = D + \sum_{\kappa\in\mathcal{T}} E_\kappa + \frac{1}{2}\sum_{\kappa\in\mathcal{T}} F_\kappa \ , \tag{9}$$

where

$$D \equiv \int_{\Gamma_{\mathrm{D}}} \sigma w^2 \, ds + \int_{\Gamma_{\mathrm{int}}} \sigma [w]^2 \, ds \ , \quad E_\kappa \equiv \|\sqrt{a}\nabla w\|^2_{L^2(\kappa)} + \|\sqrt{\gamma} w\|^2_{L^2(\kappa)} \ ,$$

$$F_\kappa \equiv \int_{\partial_-\kappa\cap\Gamma_-} |\boldsymbol{b}\cdot\boldsymbol{\mu}| w_+^2 \, ds + \int_{\partial_+\kappa\cap\Gamma_+} |\boldsymbol{b}\cdot\boldsymbol{\mu}| w_+^2 \, ds + \int_{\partial_-\kappa\setminus\Gamma_-} |\boldsymbol{b}\cdot\boldsymbol{\mu}| [w]^2 \, ds \ ,$$

with $\sqrt{a}$ *denoting the (nonnegative) square-root of the matrix* a, *and* σ *as in the definition of the discontinuity-penalisation. Furthermore, given that either* a *is positive definite or* $\gamma > 0$ *on each element* κ *in the partition* $\mathcal{T}$, *the hp-DGFEM (8) has a unique solution* u_{DG} *in* $S^{\mathbf{p}}(\Omega, \mathcal{T}, \mathbf{F})$.

Proof. We begin by proving (9). First, we note that, trivially,

$$B_s(w,w) = \int_{\Gamma_{\mathrm{D}}} \sigma w^2 \, \mathrm{d}s + \int_{\Gamma_{\mathrm{int}}} \sigma [w]^2 \, \mathrm{d}s \ .$$

Further, as $(\boldsymbol{b}\cdot\nabla w)w = \frac{1}{2}\boldsymbol{b}\cdot\nabla(w^2)$, after integration by parts we have that

$$B_b(w,w) = \frac{1}{2}\sum_{\kappa\in\mathcal{T}} F_\kappa + \sum_{\kappa\in\mathcal{T}} \int_\kappa |\sqrt{\gamma(x)} w(x)|^2 \, \mathrm{d}x \ .$$

Finally, we observe that

$$B_a(w,w) = \sum_{\kappa\in\mathcal{T}} \int_\kappa |\sqrt{a(x)}\nabla w(x)|^2 \,\mathrm{d}x \ .$$

Upon adding these three identities, we arrive at (9).

To complete the proof of the lemma, we note that if either a is positive definite or $\gamma > 0$ on each element κ in the partition $\mathcal{T}$, then $B_{\mathrm{DG}}(w,w) > 0$ for all w in $S^{\mathbf{p}}(\Omega,\mathcal{T},\mathbf{F})\backslash\{0\}$, and hence we deduce the uniqueness of the solution u_{DG}. Further, since the linear space $S^{\mathbf{p}}(\Omega,\mathcal{T},\mathbf{F})$ is finite-dimensional, the existence of the solution to (8) follows from the fact that its homogeneous counterpart has the unique solution $u_{\mathrm{DG}} \equiv 0$. □

Our second result provides a bound on the discretisation error. For simplicity, we shall assume that the entries of the matrix a are constant on each element $\kappa \in \mathcal{T}$ (with possible discontinuities across faces $e \in \mathcal{E}$) and $\boldsymbol{b}$ is a constant vector. We require the following approximation result [2,16].

Lemma 2. *Suppose that $u \in H^{k_\kappa}(\kappa)$, $k_\kappa \geq 0$, $\kappa \in \mathcal{T}$, and let $\Pi_{hp}u$ denote the orthogonal projection of u in $L^2(\Omega)$ onto the finite element space $S^{\mathbf{p}}(\Omega,\mathcal{T},\mathbf{F})$. Then, there exists a constant C dependent on k_κ and the angle condition of κ, but independent of u, $h_\kappa = diam(\kappa)$ and p_κ, such that*

$$\|u - \Pi_{hp}u\|_{L^2(\kappa)} \leq C\frac{h_\kappa^{\tau_\kappa}}{p_\kappa^{k_\kappa}} \|u\|_{H^{k_\kappa}(\kappa)} \ , \tag{10}$$

where $\tau_\kappa = \min(p_\kappa, k_\kappa)$, $\kappa \in \mathcal{T}$.

Next, we state our main result, regarding the accuracy of the method (8).

Theorem 3. *Assume that there exists a positive constant γ_0 such that $\gamma \geq \gamma_0$ on each element κ in the partition $\mathcal{T}$. Then, assuming that $u \in H^{k_\kappa}(\kappa)$, $k_\kappa \geq 2$, for $\kappa \in \mathcal{T}$, the solution $u_{\mathrm{DG}} \in S^{\mathbf{p}}(\Omega,\mathcal{T},\mathbf{F})$ of (8) obeys the error bound*

$$|||u - u_{\mathrm{DG}}|||_{\mathrm{DG}}^2 \leq C\sum_{\kappa\in\mathcal{T}} \left(\bar{a}_\kappa \frac{h_\kappa^{2(\tau_\kappa-1)}}{p_\kappa^{2(k_\kappa-2)}} + \bar{b}\frac{h_\kappa^{2(\tau_\kappa-1/2)}}{p_\kappa^{2(k_\kappa-1)}}\right)\|u\|_{H^{k_\kappa}(\kappa)}^2 \ , \tag{11}$$

where $\tau_\kappa = \min(p_\kappa, k_\kappa)$ and $\bar{b}$ is the l^2 vector norm of $\boldsymbol{b}$.

Proof. Let us decompose $e = u - u_{\mathrm{DG}}$ as $e = \eta + \xi$ where $\eta = u - \Pi_{hp}u$, $\xi = \Pi_{hp}u - u_{\mathrm{DG}}$, and Π_{hp} is as in Lemma 2. Then, by virtue of Theorem 1,

$$|||\xi|||_{\mathrm{DG}}^2 = B_{\mathrm{DG}}(\xi,\xi) = B_{\mathrm{DG}}(e-\eta,\xi) = -B_{\mathrm{DG}}(\eta,\xi) \ ,$$

where we have used the Galerkin orthogonality property: $B_{\mathrm{DG}}(u-u_{\mathrm{DG}},\xi) = 0$ which follows from (8) with $v = \xi$ and the definition of the boundary value problem (1), (4), given the assumed smoothness of u. Thus, we deduce that

$$|||\xi|||_{\mathrm{DG}}^2 \leq |B_a(\eta,\xi)| + |B_b(\eta,\xi)| + |B_s(\eta,\xi)| \ .$$

Now, from (7) we have that

$$|B_s(\eta,\xi)| \le |||\xi|||_{\mathrm{DG}} \left(\int_{\Gamma_D} \sigma|\eta|^2 \,\mathrm{d}s + \int_{\Gamma_{\mathrm{int}}} \sigma[\eta]^2 \,\mathrm{d}s \right)^{1/2} . \tag{12}$$

As $\nabla \cdot \boldsymbol{b} = 0$ on each $\kappa \in \mathcal{T}$, after integration by parts, we obtain

$$B_b(\eta,\xi) = \sum_{\kappa} \int_{\kappa} c\eta\xi \,\mathrm{d}x - \sum_{\kappa\in\mathcal{T}} \int_{\kappa} \eta(\boldsymbol{b}\cdot\nabla\xi) \,\mathrm{d}x + \sum_{\kappa\in\mathcal{T}} \int_{\partial_+\kappa\cap\Gamma_+} (\boldsymbol{b}\cdot\boldsymbol{\mu})\eta^+\xi^+ \,\mathrm{d}s$$
$$+ \sum_{\kappa\in\mathcal{T}} \int_{\partial_+\kappa\setminus\Gamma_+} (\boldsymbol{b}\cdot\boldsymbol{\mu})\eta^+\xi^+ \,\mathrm{d}s + \sum_{\kappa\in\mathcal{T}} \int_{\partial_-\kappa\setminus\Gamma_-} (\boldsymbol{b}\cdot\boldsymbol{\mu})\eta^-\xi^+ \,\mathrm{d}s . \tag{13}$$

Denoting by $S_4 + S_5$ the sum of the last two (of the five) terms in (13), we find, after shifting the 'indices' in the summation in S_4, that

$$|S_4 + S_5| \le \sum_{\kappa\in\mathcal{T}} \left(\int_{\partial_-\kappa\setminus\Gamma_-} |\boldsymbol{b}\cdot\boldsymbol{\mu}||\eta^-|^2 \,\mathrm{d}s \right)^{1/2} \left(\int_{\partial_-\kappa\setminus\Gamma_-} |\boldsymbol{b}\cdot\boldsymbol{\mu}||\xi|^2 \,\mathrm{d}s \right)^{1/2} .$$

Also, since $\boldsymbol{b}$ is a constant vector, $\int_{\kappa} \eta(\boldsymbol{b}\cdot\nabla\xi) \,\mathrm{d}x = 0$. Thus, (13) yields

$$|B_b(\eta,\xi)| \le C|||\xi|||_{\mathrm{DG}} \left(\|\eta\|^2_{L^2(\Omega)} + \sum_{\kappa\in\mathcal{T}} \int_{\partial_+\kappa\cap\Gamma_+} |\boldsymbol{b}\cdot\boldsymbol{\mu}||\eta^+|^2 \,\mathrm{d}s \right.$$
$$\left. + \sum_{\kappa\in\mathcal{T}} \int_{\partial_-\kappa\setminus\Gamma_-} |\boldsymbol{b}\cdot\boldsymbol{\mu}||\eta^-|^2 \,\mathrm{d}s \right)^{1/2} , \tag{14}$$

where C is a positive constant, as in the statement of the theorem.

Next,

$$|B_a(\eta,\xi)| \le I + II + III ,$$

where

$$I \equiv \left| \sum_{\kappa\in\mathcal{T}} \int_{\kappa} a\nabla\eta\cdot\nabla\xi \,\mathrm{d}x \right| , \quad II \equiv \left| \int_{\Gamma_{\mathrm{D}}} \{\eta((a\nabla\xi)\cdot\boldsymbol{\nu}) - ((a\nabla\eta)\cdot\boldsymbol{\nu})\xi\} \,\mathrm{d}s \right| ,$$

$$III \equiv \left| \int_{\Gamma_{\mathrm{int}}} \{[\eta]\langle(a\nabla\xi)\cdot\boldsymbol{\nu}\rangle - \langle(a\nabla\eta)\cdot\boldsymbol{\nu}\rangle[\xi]\} \,\mathrm{d}s \right| .$$

Now,

$$I^2 \le |||\xi|||^2_{\mathrm{DG}} \sum_{\kappa\in\mathcal{T}} \|\sqrt{a}\nabla\eta\|^2_{L^2(\kappa)} ,$$

$$II^2 \le C|||\xi|||^2_{\mathrm{DG}} \sum_{\kappa\,:\,\partial\kappa\cap\Gamma_{\mathrm{D}}\neq\emptyset} \left(\frac{\bar{a}_\kappa p_\kappa^2}{h_\kappa} \|\eta\|^2_{L^2(\partial\kappa\cap\Gamma_{\mathrm{D}})} + \frac{\bar{a}_\kappa h_\kappa}{p_\kappa^2} \|\nabla\eta\|^2_{L^2(\partial\kappa\cap\Gamma_{\mathrm{D}})} \right) ,$$

$$III^2 \le C|||\xi|||^2_{\mathrm{DG}} \sum_{\kappa\,:\,\partial\kappa\cap\Gamma=\emptyset} \left(\frac{\bar{a}_\kappa p_\kappa^2}{h_\kappa} \|[\eta]\|^2_{L^2(\partial\kappa)} + \frac{\bar{a}_\kappa h_\kappa}{p_\kappa^2} \|\nabla\eta\|^2_{L^2(\partial\kappa)} \right) .$$

Collecting the bounds on the terms I, II and III gives

$$|B_a(\eta,\xi)| \le C|||\xi|||_{\mathrm{DG}} \Bigg(\sum_{\kappa\in\mathcal{T}} \|\sqrt{a}\nabla\eta\|^2_{L^2(\kappa)} + \sum_{\kappa\,:\,\partial\kappa\cap\Gamma_{\mathrm{D}}\neq\emptyset} \left(\frac{\bar{a}_\kappa p_\kappa^2}{h_\kappa}\|\eta\|^2_{L^2(\partial\kappa\cap\Gamma_{\mathrm{D}})} + \frac{\bar{a}_\kappa h_\kappa}{p_\kappa^2}\|\nabla\eta\|^2_{L^2(\partial\kappa\cap\Gamma_{\mathrm{D}})} \right) + \sum_{\kappa\,:\,\partial\kappa\cap\Gamma=\emptyset} \left(\frac{\bar{a}_\kappa p_\kappa^2}{h_\kappa}\|[\eta]\|^2_{L^2(\partial\kappa)} + \frac{\bar{a}_\kappa h_\kappa}{p_\kappa^2}\|\nabla\eta\|^2_{L^2(\partial\kappa)} \right) \Bigg)^{1/2} . \quad (15)$$

The required result now follows by noting that

$$|||u-u_{\mathrm{DG}}|||_{\mathrm{DG}} \le |||\eta|||_{\mathrm{DG}} + |||\xi|||_{\mathrm{DG}} ,$$

using the estimates (12), (14) and (15) to bound $|||u-u_{\mathrm{DG}}|||_{\mathrm{DG}}$ in terms of $|||\eta|||_{\mathrm{DG}}$ and other norms of η, and applying Lemma 2, together with the Trace Inequality

$$\|v\|^2_{L^2(e)} \le C\left(\|v\|_{L^2(\kappa)}\|\nabla v\|_{L^2(\kappa)} + h_\kappa^{-1}\|v\|^2_{L^2(\kappa)} \right) , \quad v\in H^1(\kappa),\ e\subset\partial\kappa ,$$

to estimate norms of η and $\nabla\eta$ over $\partial\kappa\cap\Gamma_{\mathrm{D}}$ and $\partial\kappa$ in terms of norms of over κ, $\kappa\in\mathcal{T}$. The argument is fairly standard, so we omit the details. □

We note that in the purely hyperbolic case of $a\equiv 0$ the error bound in Theorem 3 collapses to $O(h^{\tau-1/2}/p^{k-1})$; in the DG-norm, this is optimal with respect to h, while in p it is $1/2$ an order below the *hp*-optimal bound established in [8]. In fact, for $a\equiv 0$, the error bound of Theorem 3 is identical to the p-suboptimal *hp* error estimate of Bey and Oden [4], except that there a streamline-diffusion type stabilisation was included with stabilisation parameter $\delta = h/p^2$; Theorem 3 corresponds to $\delta = 0$.

In the case of non-constant $\boldsymbol{b}$, (11) should be supplemented with the term $|\boldsymbol{b}|^2_{W^{1,\infty}(\kappa)}(h^{2\tau_\kappa}/p^{2(k_\kappa-2)})\|u\|^2_{H^{k_\kappa}(\kappa)}$ under the summation sign on the right. When $\bar{a}_\kappa \ge c_0 > 0$ this additional term can be absorbed into the first term on the right; otherwise it degrades the error bound with respect to p. More generally, when streamline-diffusion stabilisation is added to (8), with stabilisation parameter $\delta = (h/p)\min\left(1,\bar{b}h/\bar{a}p^3\right)$, the bound (11) can be, simultaneously, extended to the case of non-constant $\mathbf{b}$ and sharpened to one that is still optimal in h, but now with only $1/2$ a power of p below the optimal rate in the diffusive part and of optimal order in p in the advective part. Specifically, when $\boldsymbol{b}\equiv 0$, we recover the bound $O(h^{\tau-1}/p^{k-3/2})$ of Riviere and Wheeler [15]; on the other hand, if $a=0$, we arrive at the *hp*-optimal error bound $O(h^{\tau-1/2}/p^{k-1/2})$ of [8] in the DG-norm, proved with $\delta = h/p$, which represents the direct generalisation of the optimal h-version bound for the DGFEM (see [9] and [10]) to the *hp*-version. The proof of this is beyond the scope of the present paper and will be delivered in [17]. For further developments regarding these theoretical questions for hyperbolic and nearly-hyperbolic problems and numerical experiments, see [8,17].

References

1. Arnold, D.N.: An interior penalty finite element method with discontinuous elements. SIAM J. Numer. Anal. **19** (1982) 742–760
2. Babuška, I., Suri, M.: The *hp*-Version of the Finite Element Method with quasiuniform meshes. M^2AN Mathematical Modelling and Numerical Analysis **21** (1987) 199–238
3. Baumann, C.: An *hp*-Adaptive Discontinuous Galerkin FEM for Computational Fluid Dynamics. Doctoral Dissertation. TICAM, UT Austin, Texas, 1997
4. Bey, K.S., Oden, J.T.: *hp*-Version discontinuous Galerkin methods for hyperbolic conservation laws. Comput. Methods Appl. Mech. Engrg. **133** (1996) 259–286
5. Biswas, R., Devine, K., Flaherty, J.E.: Parallel adaptive finite element methods for conservation laws. App. Numer. Math. **14** (1994) 255–284
6. Cockburn, B., Hou, S., Shu, C.-W.: TVB Runge-Kutta local projection discontinuous Galerkin finite elements for hyperbolic conservation laws. Math. Comp. **54** (1990) 545–581
7. Cockburn, B., Shu, C.-W.: The local discontinuous Galerkin method for time-dependent reaction-diffusion systems. SIAM J. Numer. Anal. **35** (1998) 2440–2463
8. Houston, P., Schwab, Ch., Süli, E.: Stabilized *hp*-finite element methods for first-order hyperbolic problems. Oxford University Computing Laboratory Technical Report NA-98/14, 1998 (submitted for publication)
9. Johnson, C., Nävert, U., Pitkäranta, J.: Finite element methods for linear hyperbolic problems. Comp. Meth. Appl. Mech. Engrg. **45** (1984) 285–312
10. Johnson, C., Pitkäranta, J.: An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. Math. Comp. **46** (1986) 1–26
11. Lesaint, P., Raviart, P.-A.: On a finite element method for solving the neutron transport equation. In: Mathematical Aspects of Finite Elements in Partial Differential Equations. C.A. deBoor (Ed.), Academic Press, New York (1974) 89–145
12. Nitsche, J.: Über ein Variationsprinzip zur Lösung von Dirichlet Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind. Abh. Math. Sem. Univ. Hamburg **36** (1971) 9-15
13. Oden, J.T., Babuška, I., Baumann, C.: A discontinuous *hp*-FEM for diffusion problems. J. Comp. Phys. **146** (1998) 491–519
14. Reed, W.H., Hill, T.R.: Triangular mesh methods for neutron transport equation. Los Alamos Scientific Laboratory report LA-UR-73-479, Los Alamos, NM, 1973
15. Riviere, B., Wheeler, M.F.: Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. Part I. TICAM Technical Report, University of Texas at Austin, Texas, 1999
16. Schwab, Ch.: *p*- and *hp*-Finite Element Methods. Theory and Applications to Solid and Fluid Mechanics. Oxford University Press (1998)
17. Süli, E., Houston, P., Schwab, Ch.: *hp*-DGFEM for partial differential equations of mixed type. Oxford University Computing Laboratory Technical Report NA-99/05, 1999
18. Wheeler, M.F.: An elliptic collocation finite element method with interior penalties. SIAM J. Numer. Anal. **15** (1978) 152–161

A Discontinuous Galerkin Method Applied to Nonlinear Parabolic Equations

Béatrice Rivière and Mary F. Wheeler

The Center for Subsurface Modeling, TICAM, The University of Texas, Austin TX 78712, USA

Abstract. Semi-discrete and a family of discrete time locally conservative Discontinuous Galerkin procedures are formulated for approximations to nonlinear parabolic equations. For the continuous time approximations *a priori* $L^\infty(L^2)$ and $L^2(H^1)$ estimates are derived and similarly, $l^\infty(L^2)$ and $l^2(H^1)$ for the discrete time schemes. Spatial rates in H^1 and time truncation errors in L^2 are optimal.

1 Introduction

Over the last two decades there has been a collection of papers devoted to the use of approximation spaces with weak continuity for finite element approximations to elliptic and parabolic problems. The motivation for developing these methods was the flexibility afforded by local approximation spaces. These approaches allow meshes which are more general in their construction and degree of nonuniformity both in time and space than is permitted by the more conventional finite element methods. In general numerical methods defined for discontinuous spaces have less numerical diffusion/dispersion and provide more accurate local approximations for problems with rough solutions. Another advantage that has recently become apparent is the application of domain decomposition algorithms for the discrete solution.

Discontinuous Galerkin methods using interior penalties for elliptic and parabolic equations were first introduced by Douglas, Dupont and Wheeler [11],[4] and Arnold [1] in the seventies. These approaches generalize a method by Nitsche [6] for treating Dirichlet boundary condition by the introduction of penalty terms on the boundary of the domain. Applications of these methods to flow in porous media were presented by Douglas, Wheeler, Darlow, Kendall and Ewing in [5],[2]. These methods frequently referred to as interior penalty Galerkin schemes are not locally mass conservative.

A new type of discontinuous Galerkin method for diffusion problems was introduced and analyzed by Oden, Babuška and Baumann [7]. It was shown that the discontinuous Galerkin method was elementwise conservative. Also, *a priori* error estimates were proven for one-dimensional problems and for polynomials of at least order three. Numerical experiments in higher dimension showed the robustness of the method. The authors [8],[9] have derived *a priori* and *a posteriori* error estimates in higher dimensions. In this paper, a discontinuous Galerkin formulation for nonlinear parabolic equations is introduced and analyzed.

This paper consists of three additional sections. In §2 and §3, notation and problem definition and formulation of the discontinuous Galerkin method are described. In §4 and §5, the proofs of the error estimates in the continuous and discrete time setting are respectively given. Conclusions are described in the last section.

2 Model problem

Consider the nonlinear parabolic partial differential equation

$$u_t - \nabla \cdot a(x,u)\nabla u = f(x,u), \quad (x,t) \in \Omega \times (0,T], \tag{1}$$

with the boundary condition

$$a(x,u)\nabla u \cdot \nu = 0, \quad (x,t) \in \partial\Omega \times (0,T], \tag{2}$$

and the initial condition

$$u(x,0) = \psi(x), \quad x \in \Omega, \tag{3}$$

where Ω is a bounded domain in $\mathbb{R}^d$. Since many of the proofs are highly technical, we shall restrict our attention to $d = 1, 2$. We make several assumptions:

- For $(x,p) \in \Omega \times R$, $\exists\ \gamma$ and γ^* s.t. $0 < \gamma \leq a(x,p) \leq \gamma^*$.
- a, f are uniformly Lipschitz continuous with respect to their second variable.
- $u \in C^2(\Omega \times [0,T])$ is a unique solution to (1), (2), and (3), and $u \in L^2([0,T], H^s(\Omega))$, $u_t \in L^2([0,T], H^{s-1}(\Omega))$ for $s \geq 2$.
- ∇u is bounded in $L^\infty(\Omega \times (0,T])$.

(Here for X normed space and n positive integer, $L^n([0,T],X) = \{f : \int_0^T \|f\|_X^n(t)dt < \infty\}$.)

3 Definitions and the Discontinuous Galerkin procedure

Let $\mathcal{E}_h = \{E_1, E_2, \ldots, E_{N_h}\}$ be a subdivision of Ω, where E_j is a triangle or a quadrilateral. Let $h_j = \text{diam}(E_j)$ and $h = \max\{h_j,\ j = 1 \ldots N_h\}$. We denote the edges of the elements by $\{e_1, e_2, \ldots, e_{P_h}, e_{P_h+1}, \ldots, e_{M_h}\}$ where $e_k \subset \Omega, 1 \leq k \leq P_h$,and $e_k \subset \partial\Omega$, $P_h + 1 \leq k \leq M_h$. With each edge e_k , we associate a unit normal vector ν_k. For $k > P_h$, ν_k is taken to be the unit outward vector normal to $\partial\Omega$.
For $s \geq 0$, let

$$H^s(\mathcal{E}_h) = \{v \in L^2(\Omega) : v|_{E_j} \in H^s(E_j), j = 1 \ldots N_h\}.$$

We now define the average and the jump for $\phi \in H^s(\mathcal{E}_h)$, $s > \frac{1}{2}$. Let $1 \le k \le P_h$. For $e_k = \partial E_i \cap \partial E_j$ with ν_k exterior to E_i, set

$$\{\phi\} = \frac{1}{2}(\phi|_{E_i})|_{e_k} + \frac{1}{2}(\phi|_{E_j})|_{e_k}, \quad [\phi] = (\phi|_{E_i})|_{e_k} - (\phi|_{E_j})|_{e_k}.$$

The L^2 inner product is denoted by $(\cdot,\cdot)$. The usual Sobolev norm on $E \subset \mathbb{R}^d$ and for m positive integer, is denoted by $\|\cdot\|_{m,E}$. We define the following broken norms:

$$|||\phi|||_m^2 = \sum_{j=1}^{N_h} \|\phi\|_{m,E_j}^2,$$

$$|||\phi|||_{L^2((\alpha,\beta);H^m)}^2 = \int_\alpha^\beta |||\phi(.,t)|||_m^2 dt, \quad |||\phi|||_{L^\infty((\alpha,\beta);H^m)} = \sup_{t\in(\alpha,\beta)} |||\phi(.,t)|||_m.$$

Let r be a positive integer. The finite element subspace is taken to be

$$\mathcal{D}_r(\mathcal{E}_h) = \prod_{j=1}^{N_h} P_r(E_j),$$

where $P_r(E_j)$ denotes the set of polynomials of (total) degree less than or equal to r on E_j, even if E_j is a quadrilateral. We introduce the interior penalty term

$$J_0^\sigma(\phi,\psi) = \sum_{k=1}^{P_h} \frac{\sigma_k}{|e_k|} \int_{e_k} [\phi][\psi],$$

where $|e_k|$ denotes the length of e_k and σ_k is a real nonnegative number associated to the interior edge e_k.

A proof of the following lemma can be found in [8].

Lemma 1. *Let $u \in H^s(\Omega)$, for $s \ge 2$ and let $r \ge 2$. Let $\bar{a}$ be a positive constant. There is $\hat{u} \in \mathcal{D}_r(\mathcal{E}_h)$ interpolant of p satisfying*

$$\int_{e_k} \{\bar{a}\nabla(\hat{u}-u)\cdot\nu_k\} = 0, \quad \forall k = 1,\dots,P_h \tag{4}$$

$$\|\hat{u}-u\|_{\infty,E_j} \le C\frac{h^\mu}{r^{s-1}}\|u\|_{s,E_j}, \quad \forall E_j, \tag{5}$$

$$\|\nabla(\hat{u}-u)\|_{0,E_j} \le C\frac{h^{\mu-1}}{r^{s-1}}\|u\|_{s,E_j}, \quad \forall E_j, \tag{6}$$

$$\|\nabla^2(\hat{u}-u)\|_{0,E_j} \le C\frac{h^{\mu-2}}{r^{s-2}}\|u\|_{s,E_j}, \quad \forall E_j, \tag{7}$$

$$\|\hat{u}-u\|_{0,E_j} \le C\frac{h^{\mu}}{r^{s-1}}\|u\|_{s,E_j}, \quad \forall E_j, \tag{8}$$

where $\mu = \min(r+1, s)$. Morevoer, for $e_k = \partial E^1 \cap \partial E^2$,

$$\|\nabla\hat{u}\|_{\infty,e_k} \le C\|\nabla u\|_{\infty,E^1\cup E^2} \tag{9}$$

The Discontinuous Galerkin approximation $U(\cdot,t) \in \mathcal{D}_r(\mathcal{E}_h)$ to the solution u of (1), (2), and (3) is defined by

$$(\frac{\partial U}{\partial t}, v) + \sum_{j=1}^{N_h} \int_{E_j} a(U)\nabla U \nabla v - \sum_{k=1}^{P_h} \int_{e_k} \{a(U)\nabla U \cdot \nu_k\}[v]$$

$$+ \sum_{k=1}^{P_h} \int_{e_k} \{a(U)\nabla v \cdot \nu_k\}[U] + J_0^\sigma(U,v) = \int_\Omega f(U)v, \ t > 0, \ v \in \mathcal{D}_r(\mathcal{E}_h), \quad (10)$$

$$U(\cdot,0) = \psi, \quad (11)$$

where we have assumed for simplicity that $\psi \in \mathcal{D}_r(\mathcal{E}_h)$. We note that if $\{v_i\}_{i=1}^M$ is a basis of $\mathcal{D}_r(\mathcal{E}_h)$ and if we write

$$U(x,t) = \sum_{i=1}^{M} \xi_i(t) v_i(x),$$

then (10) and (11) reduces to an initial value problem for the system of nonlinear ordinary differential equations

$$G\xi'(t) = -B(\xi)\xi + F(\xi)$$
$$\xi(0) = b.$$

The matrix G is block diagonal symmetric positive definite. Since a and f are Lipschitz continuous, it follows from the theory of ordinary differential equations that $\xi(t)$ exists and is unique for $t > 0$.

4 Continuous in time *a priori* error estimate

In this section, we demonstrate optimal $L^2(H^1)$ rates of convergence for continuous in time Discontinuous Galerkin approximations of at least quadratic order.

Theorem 2. *Let $s \geq 2$. There exists a constant C^* independent of h and r such that,*

$$\|U-u\|^2_{L^\infty((0,T);L^2)} + \|U-u\|^2_{L^2((0,T);H^1)} \leq C^* \frac{h^{2\mu-2}}{r^{2s-4}} \|u\|^2_{L^2((0,T),H^s)} +$$
$$C^* \frac{h^{2\mu-2}}{r^{2s-4}} \|u_t\|^2_{L^2((0,T),H^{s-1})},$$

where $\mu = \min(r+1, s)$, $r \geq 2$, and $\sigma_k \geq 0$ if $a = a(x)$ and $\sigma_k > 0$ if $a = a(x,u)$.

Proof. It is clear that if u is a solution of (1), (2) and (3), then u satisfies the formulation:

$$(\frac{\partial u}{\partial t}, v) + \int_\Omega a(u)\nabla u \nabla v - \sum_{k=1}^{P_h} \int_{e_k} \{a(u)\nabla u \cdot \nu_k\}[v]$$
$$+ \sum_{k=1}^{P_h} \int_{e_k} \{a(u)\nabla v \cdot \nu_k\}[u] + J_0^\sigma(u, v) = \int_\Omega f(u)v, \quad \forall v \in \mathcal{D}_r(\mathcal{E}_h).$$

We obtain:

$$(\frac{\partial \hat{u}}{\partial t}, v) + \int_\Omega a(U)\nabla \hat{u} \nabla v - \sum_{k=1}^{P_h} \int_{e_k} \{a(U)\nabla \hat{u} \cdot \nu_k\}[v]$$
$$+ \sum_{k=1}^{P_h} \int_{e_k} \{a(U)\nabla v \cdot \nu_k\}[\hat{u}] + J_0^\sigma(\hat{u}, v) = \int_\Omega \frac{\partial(\hat{u} - u)}{\partial t} v + \int_\Omega f(u)v$$
$$+ \int_\Omega a(u)\nabla(\hat{u} - u)\nabla v - \sum_{k=1}^{P_h} \int_{e_k} \{a(u)\nabla(\hat{u} - u) \cdot \nu_k\}[v]$$
$$+ \sum_{k=1}^{P_h} \int_{e_k} \{a(u)\nabla v \cdot \nu_k\}[\hat{u} - u] + J_0^\sigma(\hat{u} - u, v) + \int_\Omega (a(U) - a(u))\nabla \hat{u} \nabla v$$
$$- \sum_{k=1}^{P_h} \int_{e_k} \{(a(U) - a(u))\nabla \hat{u} \cdot \nu_k\}[v] + \sum_{k=1}^{P_h} \int_{e_k} \{(a(U) - a(u))\nabla v \cdot \nu_k\}[\hat{u}]. \quad (12)$$

Subtract (12) from (10), denote $U - \hat{u} = \xi$, $\hat{u} - u = \chi$, and choose $v = \xi$:

$$(\frac{\partial \xi}{\partial t}, \xi) + \sum_{j=1}^{N_h} \int_{E_j} a(U)\nabla \xi \nabla \xi + J_0^\sigma(\xi, \xi) = -\int_\Omega \frac{\partial \chi}{\partial t} \xi + \int_\Omega (f(U) - f(u))\xi$$
$$- \int_\Omega a(u)\nabla \chi \nabla \xi + \sum_{k=1}^{P_h} \int_{e_k} \{a(u)\nabla \chi \cdot \nu_k\}[\xi] - \sum_{k=1}^{P_h} \int_{e_k} \{a(u)\nabla \xi \cdot \nu_k\}[\chi]$$
$$+ J_0^\sigma(\chi, \xi) + T_1 - T_2 + T_3, \quad (13)$$

where

$$T_1 = \sum_{j=1}^{N_h} \int_{E_j} (a(u) - a(U))\nabla \hat{u} \nabla \xi,$$
$$T_2 = \sum_{k=1}^{P_h} \int_{e_k} \{(a(u) - a(U))\nabla \hat{u} \cdot \nu_k\}[\xi],$$
$$T_3 = \sum_{k=1}^{P_h} \int_{e_k} \{(a(u) - a(U))\nabla \xi \cdot \nu_k\}[\hat{u}].$$

We now derive bounds for all the terms on the right-hand side of (13). The constants C_i are generic constants that vary but do not depend on h or r.

$$|\int_\Omega \frac{\partial \chi}{\partial t}\xi| \leq C_1(|||\chi_t|||_0^2 + |||\xi|||_0^2),$$

$$|\int_\Omega (f(U) - f(u))\xi| \leq C_2(|||\chi|||_0^2 + |||\xi|||_0^2),$$

$$|\sum_{j=1}^{N_h} \int_{E_j} a(u)\nabla\chi\nabla\xi| \leq C_3|||\nabla\chi|||_0|||\nabla\xi|||_0,$$

$$\leq \frac{C_3}{\epsilon_1}|||\nabla\chi|||_0^2 + \epsilon_1|||\nabla\xi|||_0^2.$$

To bound the terms involving integrals on the interior edges, we first look at the integral on a given edge e_k, and we assume that $e_k = \partial E^1 \cap \partial E^2$. We denote $E^{12} \equiv E^1 \cup E^2$. Define $\bar{a}$ piecewise constant on each element E_j such that $\bar{a}|_{E_j} = \frac{1}{|E_j|}\int_{E_j} a(u)$.

$$|\int_{e_k} \{a(u)\nabla\chi\cdot\nu_k\}[\xi]| \leq |\int_{e_k} \{\bar{a}\nabla\chi\cdot\nu_k\}[\xi]| + |\int_{e_k} \{(a(u)-\bar{a})\nabla\chi\cdot\nu_k\}[\xi]|.$$

Define the constant c_k associated to each interior edge as follows

$$c_k = \frac{1}{|e_k|}\int_{e_k}[\xi].$$

By Lemma 1, we see that

$$\begin{aligned}|\int_{e_k} \{\bar{a}\nabla\chi\cdot\nu_k\}[\xi]| &= |\int_{e_k} \{\bar{a}\nabla\chi\cdot\nu_k\}([\xi]-c_k)|,\\ &\leq ||\{\bar{a}\nabla\chi\cdot\nu_k\}||_{0,e_k}||[\xi]-c_k||_{0,e_k},\\ &\leq C_4||\nabla\xi||_{0,E^{12}}(||\nabla\chi||_{0,E^{12}} + h||\nabla^2\chi||_{0,E^{12}}).\end{aligned}$$

Now,

$$|\int_{e_k} \{(a(u)-\bar{a})\nabla\chi\cdot\nu_k\}[\xi]| \leq C_4||a(u)-\bar{a}||_{\infty,E^{12}}||\{\nabla\chi\cdot\nu_k\}||_{0,e_k}||[\xi]||_{0,e_k}.$$

But,

$$\begin{aligned}||a(u)-\bar{a}||_{\infty,E^{12}} &\leq C_4 h,\\ ||\{\nabla\chi\cdot\nu_k\}||_{0,e_k} &\leq C_4 h^{-\frac{1}{2}}(||\nabla\chi||_{0,E^{12}} + h||\nabla^2\chi||_{0,E^{12}}),\\ ||[\xi]||_{0,e_k} &\leq C_4 h^{-\frac{1}{2}}||\xi||_{0,E^{12}}.\end{aligned}$$

So,

$$|\int_{e_k} \{(a(u)-\bar{a})\nabla\chi\cdot\nu_k\}[\xi]| \leq C_4||\xi||_{0,E^{12}}(||\nabla\chi||_{0,E^{12}} + h||\nabla^2\chi||_{0,E^{12}}).$$

Summing on k, we have

$$|\sum_{k=1}^{P_h}\int_{e_k}\{a(u)\nabla\chi\cdot\nu_k\}[\xi]| \le \epsilon_2|||\xi|||_0^2+\epsilon_2|||\nabla\xi|||_0^2+C_4(|||\nabla\chi|||_0+h|||\nabla^2\chi|||_0)^2.$$

Similarly, we note that

$$\begin{aligned}|\int_{e_k}\{a(u)\nabla\xi\cdot\nu_k\}[\chi]| &\le C_5||\{a(u)\nabla\xi\cdot\nu_k\}||_{0,e_k}||[\chi]||_{0,e_k},\\ &\le C_5||\nabla\xi||_{0,E^{12}}(h^{-1}||\chi||_{0,E^{12}}+||\nabla\chi||_{0,E^{12}}).\end{aligned}$$

Thus, by summing on k we obtain

$$|\sum_{k=1}^{P_h}\int_{e_k}\{a(u)\nabla\xi\cdot\nu_k\}[\chi]| \le \epsilon_3|||\nabla\xi|||_0^2+\frac{C_5}{\epsilon_3}(h^{-1}|||\chi|||_0+|||\nabla\chi|||_0)^2.$$

For the linear case $a=a(x)$, the following four terms (penalty term and T_1,T_2,T_3) do not appear in (13). For the nonlinear case, we observe that

$$\begin{aligned}|\sum_{k=1}^{P_h}\frac{\sigma_k}{|e_k|}\int_{e_k}[\chi][\xi]| &\le \sum_{k=1}^{P_h}\frac{\sigma_k}{|e_k|}||[\chi]||_{0,e_k}||[\xi]||_{0,e_k},\\ &\le \epsilon_4 J_0^\sigma(\xi,\xi)+\frac{C_6}{\epsilon_4}\sum_{k=1}^{P_h}\frac{1}{|e_k|}||[\chi]||_{0,e_k}^2,\\ &\le \epsilon_4 J_0^\sigma(\xi,\xi)+\frac{C_6}{\epsilon_4}\sum_{k=1}^{P_h}\frac{1}{|e_k|}(h^{-1}||\chi||_{0,E^{12}}^2+h||\nabla\chi||_{0,E^{12}}^2),\\ &\le \epsilon_4 J_0^\sigma(\xi,\xi)+\frac{C_6}{\epsilon_4}(h^{-2}|||\chi|||_0^2+|||\nabla\chi|||_0^2).\end{aligned}$$

In addition, by estimate (9) in Lemma 1,we see that

$$\begin{aligned}|T_1| &\le C_7\sum_{j=1}^{N_h}\int_{E_j}|u-U||\nabla\hat{u}\cdot\nabla\xi|,\\ &\le C_7||\nabla\hat{u}||_\infty|||u-U|||_0|||\nabla\xi|||_0,\\ &\le \frac{C_7}{\epsilon_5}(|||\chi|||_0^2+|||\xi|||_0^2)+\epsilon_5|||\nabla\xi|||_0^2.\end{aligned}$$

Similarly, we observe that

$$\begin{aligned}|\int_{e_k}\{(a(u)-a(U))\nabla\hat{u}\cdot\nu_k\}[\xi]| &\le C_8(||\nabla u||_{\infty,E^{12}})||\{u-U\}||_{0,e_k}||[\xi]||_{0,e_k},\\ &\le \epsilon_6\frac{\sigma_k}{|e_k|}||[\xi]||_{0,e_k}^2+C_8|e_k|h^{-1}||\xi||_{0,E^{12}}^2\\ &\quad +C_8|e_k|(h^{-1}||\chi||_{0,E^{12}}^2+h||\nabla\chi||_{0,E^{12}}^2).\end{aligned}$$

Summing on k, $k = 1, \ldots, P_h$, we have

$$|T_2| \leq \epsilon_6 J_0^\sigma(\xi,\xi) + C_8(|||\chi|||_0^2 + h^2|||\nabla\chi|||_0^2 + |||\xi|||_0^2).$$

Similarly, we have by (8) in Lemma 1

$$\left|\int_{e_k} \{(a(u) - a(U))\nabla\xi \cdot \nu_k\}[\chi]\right| \leq \frac{C_9}{\epsilon_7}||\nabla u||_{\infty,E^{12}}(||\chi||_{0,E^{12}} + h||\chi||_{0,E^{12}}) + \epsilon_7||\nabla\xi||_{0,E^{12}} + \frac{C_9}{\epsilon_7}||\nabla u||_{\infty,E^{12}}||\xi||_{0,E^{12}}.$$

Summing on k, $k = 1, \ldots, P_h$, we have

$$|T_3| \leq \epsilon_7|||\nabla\xi|||_0^2 + C_9(|||\chi|||_0^2 + h^2|||\nabla\chi|||_0^2 + |||\xi|||_0^2).$$

Combining the above bounds for the right-hand side and choosing the ϵ_i small enough, we have

$$\frac{1}{2}\frac{\partial|||\xi|||_0^2}{\partial t} + \frac{\gamma}{2}|||\nabla\xi|||_0^2 + \frac{1}{2}J_0^\sigma(\xi,\xi) \leq (\hat{C}_1 + \frac{\hat{C}_2}{h^2})|||\chi|||_0^2 + \hat{C}_3|||\nabla\chi|||_0^2 + \hat{C}_4h^2|||\nabla^2\chi|||_0^2 + \hat{C}_5|||\xi|||_0^2 + \hat{C}_6|||\chi_t|||_0.$$

where $\hat{C}_1, \hat{C}_2, \hat{C}_3, \hat{C}_4$ and $\hat{C}_5$ are constants independent of h and r.

Now, we integrate with respect to time between 0 and τ and obtain:

$$\begin{aligned}|||\xi|||_0^2(\tau) + \frac{\gamma}{2}\int_0^\tau |||\nabla\xi|||_0^2(t)dt + \frac{1}{2}\int_0^\tau J_0^\sigma(\xi,\xi) \leq |||\xi|||_0^2(0) &+ \hat{C}_5\int_0^\tau |||\xi|||_0^2 \\ &+ (\hat{C}_1 + \frac{\hat{C}_2}{h^2})\int_0^\tau |||\chi|||_0^2 \\ &+ \hat{C}_3\int_0^\tau |||\nabla\chi|||_0^2 \\ &+ \hat{C}_4h^2\int_0^\tau |||\nabla^2\chi|||_0^2 \\ &+ \hat{C}_6\int_0^\tau |||\chi_t|||_0^2.\end{aligned}$$

Using Gronwall's inequality and the approximation results, we obtain

$$|||\xi|||_0^2(\tau) + \frac{\gamma}{2}\int_0^\tau |||\nabla\xi|||_0^2(t)dt + \frac{1}{2}\int_0^\tau J_0^\sigma(\xi,\xi) \leq C\frac{h^{2\mu-2}}{r^{2s-4}}|||u|||_{L^2((0,\tau),H^s)}^2 + C\frac{h^{2\mu-2}}{r^{2s-4}}|||u_t|||_{L^2((0,\tau),H^{s-1})}^2.$$

The result follows by triangle inequality. □

5 Discrete in time Discontinuous Galerkin procedures

Let $\Delta t = T/N$ where N is a positive integer and let $t_j = j\Delta t$. We use the following notation:

$$g_j = g(x, t_j), \quad 0 \le j \le N,$$
$$g_{j,\theta} = \frac{1}{2}(1+\theta)g_{j+1} + \frac{1}{2}(1-\theta)g_j, \quad 0 \le j \le N-1,$$

where $\theta \in [0,1]$. Define the norms:

$$\|g\|_{l^\infty(L^2)} = \max_{j=0,\ldots,N} \|g_j\|_0, \quad \|g\|_{l^2(H^1)} = (\sum_{j=0}^{N-1} \|\nabla g_{j,\theta}\|_0^2)^{\frac{1}{2}}.$$

Consider the following discrete Discontinuous Galerkin procedure: Let $\{U_j\}_{j=0}^N$ be a sequence in $\mathcal{D}_r(\mathcal{E}_h)$ that satisfies:

$$\begin{aligned} &\int_\Omega \frac{U_{j+1}-U_j}{\Delta t} v + \int_\Omega a(U_{j,\theta})\nabla U_{j,\theta}\nabla v \\ &-\sum_{k=1}^{P_h}\int_{e_k}\{a(U_{j,\theta})\nabla U_{j,\theta}\cdot\nu_k\}[v] + \sum_{k=1}^{P_h}\int_{e_k}\{a(U_{j,\theta})\nabla v\cdot\nu_k\}[U_{j,\theta}] \\ &= \int_\Omega f(x, U_{j,\theta})v + J_0^\sigma(U_{j,\theta}, v), \quad t>0, \quad v \in \mathcal{D}_r(\mathcal{E}_h), \end{aligned} \tag{14}$$

$$U_0 = \psi, \tag{15}$$

where $\theta \in [0,1]$. If $\theta = 0$, (14) yields the Crank-Nicolson Discontinuous Galerkin approximation; for $\theta = 1$, (14) is a backward difference Discontinuous Galerkin approximation.

We remark that (14) and (15) have solutions (possibly non unique) if Δt is sufficiently small [3] [10].

We have the following result on the interpolant $\hat{u}$ of u in Lemma 1. In particular, $\frac{\partial \hat{u}}{\partial t}$ is the interpolant of $\frac{\partial u}{\partial t}$.

Lemma 3.

$$\frac{\hat{u}_{j+1}-\hat{u}_j}{\Delta t} = \hat{u}_t(x, t_{j,\theta}) + \Delta t \rho_{j,\theta}, \quad \forall x \in \Omega, \tag{16}$$

where

$$\|\rho_{j,\theta}\|_0 \le C_1 \|u_{tt}\|_{L^\infty((t_j,t_{j+1});H^1)}$$

In the particular case $\theta = 0$, we also have

$$\|\rho_{j,0}\|_0 \le \Delta t C_2 \|u_{ttt}\|_{L^\infty((t_j,t_{j+1});H^1)}$$

C_1 and C_2 are two constants independent of $u, \hat{u}$ and r.

Proof. The Taylor expansions around $t_{j,\theta}$ yield

$$\hat{u}_{j+1} = \hat{u}_{j,\theta} + \frac{1-\theta}{2}\Delta t\hat{u}_t(t_{j,\theta}) + \frac{1}{2}(\frac{1-\theta}{2})^2\Delta t^2\hat{u}_{tt}(t_{j,\theta}) + \frac{1}{6}(\frac{1-\theta}{2})^3\Delta t^3\hat{u}_{ttt}(t^*), \tag{17}$$

$$\hat{u}_j = \hat{u}_{j,\theta} - \frac{1+\theta}{2}\Delta t\hat{u}_t(t_{j,\theta}) + \frac{1}{2}(\frac{1+\theta}{2})^2\Delta t^2\hat{u}_{tt}(t_{j,\theta}) - \frac{1}{6}(\frac{1+\theta}{2})^3\Delta t^3\hat{u}_{ttt}(t^{**}). \tag{18}$$

Subtracting (18) from (17)

$$\hat{u}_{j+1} - \hat{u}_j = \Delta t\hat{u}_t(t_{j,\theta}) + \Delta t\rho_{j,\theta},$$

where

$$\rho_{j,\theta} = \frac{1}{2}((\frac{1-\theta}{2})^2 - (\frac{1+\theta}{2})^2)\hat{u}_{tt}(t_{j,\theta}) + \frac{1}{6}(\frac{1-\theta}{2})^3\Delta t\hat{u}_{ttt}(t^*) + \frac{1}{6}(\frac{1+\theta}{2})^3\Delta t\hat{u}_{ttt}(t^{**}).$$

Clearly, for $\theta \in (0,1]$, we have

$$\|\rho_{j,\theta}\|_0 \leq C(\theta)\|\hat{u}_{tt}\|_{L^\infty((t_j,t_{j+1});L^2)}.$$

Given t, we have

$$\begin{aligned}\|\hat{u}_{tt}\|_0 &\leq \|\hat{u}_{tt} - u_{tt}\|_0 + \|u_{tt}\|_0 \\ &\leq Ch\|u_{tt}\|_1 + \|u_{tt}\|_0.\end{aligned}$$

Thus, for $\theta \in (0,1]$, we have

$$\|\rho_{j,\theta}\|_0 \leq C(h+1)\|u_{tt}\|_{L^\infty((t_j,t_{j+1});H^1)}.$$

For $\theta = 0$, we have in a similar fashion

$$\begin{aligned}\|\rho_{j,\theta}\|_0 &\leq C\Delta t\|\hat{u}_{ttt}\|_{L^\infty((t_j,t_{j+1});H^1)} \\ &\leq C\Delta t(h+1)\|u_{ttt}\|_{L^\infty((t_j,t_{j+1});H^1)}.\end{aligned}$$

□

Theorem 4. *Assume*

- $u_{tt} \in L_\infty([0,T];H^1(\Omega))$;
- *For* $\theta = 0$, $u_{ttt} \in L_\infty([0,T];H^1(\Omega))$;

Let $U_j, 0 \leq j \leq N$, *be defined by (14) and (15) for* $\theta \in (0,1]$. *Then, if* Δt *is sufficiently small, there exist* C^* *and* $\hat{C}$ *independent of* h *and* r *such that for*

$\mu = \min(r+1, s)$ *and* $\sigma_k \geq 0$ *if* $a = a(x)$ *and* $\sigma_k > 0$ *if* $a = a(x,u)$,

$$|||U-u|||^2_{l^\infty(L^2)} + \Delta t\gamma |||U-u|||^2_{l^2(H^1)} \leq C^* \frac{h^{2\mu-2}}{r^{2s-4}} \Delta t \sum_{j=0}^{N} |||u_j|||^2_{H^s} + \hat{C}\Delta t^2 \sum_{j=0}^{N-1} \Delta t |||u_{tt}|||^2_{L^2((t_j,t_{j+1});H^1)}.$$

For $\theta = 0$, *we have*

$$|||U-u|||^2_{l^\infty(L^2)} + \Delta t\gamma |||U-u|||^2_{l^2(H^1)} \leq C^* \frac{h^{2\mu-2}}{r^{2s-4}} \Delta t \sum_{j=0}^{N} |||u_j|||^2_{H^s} + \hat{C}\Delta t^4 \sum_{j=0}^{N-1} \Delta t |||u_{ttt}|||^2_{L^\infty((t_j,t_{j+1});H^1)}.$$

Proof. We see that for $t = t_{j,\theta}, 0 \leq j \leq N-1$ and $v \in \mathcal{D}_r(\mathcal{E}_h)$,

$$\begin{aligned}
&(\frac{\hat{u}_{j+1} - \hat{u}j}{\Delta t}, v) + \int_\Omega a(U_{j,\theta})\nabla \hat{u}_{j,\theta}\nabla v - \sum_{k=1}^{P_h} \int_{e_k} \{a(U_{j,\theta})\nabla \hat{u}_{j,\theta}\cdot \nu_k\}[v] \\
&+ \sum_{k=1}^{P_h} \int_{e_k} \{a(U_{j,\theta})\nabla v \cdot \nu_k\}[\hat{u}_{j,\theta}] = \int_\Omega f(x, u_{j,\theta})v + \int_\Omega \Delta t \rho_{j,\theta} v \\
&\quad + \int_\Omega a(u_{j,\theta})\nabla(\hat{u}_{j,\theta} - u_{j,\theta})\nabla v + J_0^\sigma(\hat{u}_{j,\theta} - u_{j,\theta}, v) \\
&\quad - \sum_{k=1}^{P_h} \int_{e_k} \{a(u_{j,\theta})\nabla(\hat{u}_{j,\theta} - u_{j,\theta})\cdot \nu_k\}[v] \\
&\quad + \sum_{k=1}^{P_h} \int_{e_k} \{a(u_{j,\theta})\nabla v \cdot \nu_k\}[\hat{u}_{j,\theta} - u_{j,\theta}] \\
&\quad + \int_\Omega (a(U_{j,\theta}) - a(u_{j,\theta}))\nabla \hat{u}_{j,\theta}\nabla v \\
&\quad - \sum_{k=1}^{P_h} \int_{e_k} \{(a(U_{j,\theta}) - a(u_{j,\theta}))\nabla \hat{u}_{j,\theta}\cdot \nu_k\}[v] \\
&\quad + \sum_{k=1}^{P_h} \int_{e_k} \{(a(U_{j,\theta}) - a(u_{j,\theta}))\nabla v \cdot \nu_k\}[\hat{u}_{j,\theta}].
\end{aligned} \tag{19}$$

Subtracting (19) from (14), denoting $\xi_{j,\theta} = U_{j,\theta} - \hat{u}_{j,\theta}, \chi_{j,\theta} = \hat{u}_{j,\theta} - u_{j,\theta}$ and choosing $v = \xi_{j,\theta}$:

$$
\begin{aligned}
(\frac{\xi_{j+1} - \xi j}{\Delta t}, \xi_{j,\theta}) + \int_\Omega a(U_{j,\theta}) \nabla \xi_{j,\theta} \cdot \nabla \xi_{j,\theta} + J_0^\sigma(\xi_{j,\theta}, \xi_{j,\theta}) = \\
- \int_\Omega a(u_{j,\theta}) \nabla \chi_{j,\theta} \nabla \xi_{j,\theta} + \int_\Omega (f(x, U_{j,\theta}) - f(x, u_{j,\theta})) \xi_{j,\theta} \\
+ \sum_{k=1}^{P_h} \int_{e_k} \{a(u_{j,\theta}) \nabla \chi_{j,\theta} \cdot \nu_k\} [\xi_{j,\theta}] - \sum_{k=1}^{P_h} \int_{e_k} \{a(u_{j,\theta}) \nabla \xi_{j,\theta} \cdot \nu_k\} [\chi_{j,\theta}] \\
- \int_\Omega \Delta t \rho_{j,\theta} \xi_{j,\theta} + J_0^\sigma(\chi_{j,\theta}, \xi_{j,\theta}) + \int_\Omega (a(u_{j,\theta}) - a(U_{j,\theta})) \nabla \hat{u}_{j,\theta} \nabla \xi_{j,\theta} \\
- \sum_{k=1}^{P_h} \int_{e_k} \{(a(u_{j,\theta}) - a(U_{j,\theta})) \nabla \hat{u}_{j,\theta} \cdot \nu_k\} [\xi_{j,\theta}] \\
+ \sum_{k=1}^{P_h} \int_{e_k} \{(a(u_{j,\theta}) - a(U_{j,\theta})) \nabla \xi_{j,\theta} \cdot \nu_k\} [\hat{u}_{j,\theta}].
\end{aligned}
$$

It is easy to show that we have:

$$
\frac{1}{2\Delta t} (|||\xi_{j+1}|||_0^2 - |||\xi_j|||_0^2) \leq (\frac{\xi_{j+1} - \xi_j}{\Delta t}, \xi_{j,\theta}).
$$

By using similar arguments as in the time-continuous case, we have

$$
\begin{aligned}
\frac{1}{2\Delta t} (|||\xi_{j+1}|||_0^2 - |||\xi_j|||_0^2) + \frac{\gamma}{2} |||\nabla \xi_{j,\theta}|||_0^2 \leq (C_1 + \frac{C_2}{h^2}) |||\chi_{j,\theta}|||_0^2 + C_3 |||\nabla \chi_{j,\theta}|||_0^2 \\
+ C_4 h^2 |||\nabla^2 \chi_{j,\theta}|||_0^2 + C_5 |||\xi_{j,\theta}|||_0^2 \\
+ C_6 |||\chi_{t_{j,\theta}}|||_0^2 + C_7 \Delta t^2 |||\rho_{j,\theta}|||_0^2.
\end{aligned}
$$

After some manipulation, we get

$$
\begin{aligned}
\frac{1}{2\Delta t} (|||\xi_{j+1}|||_0^2 - |||\xi_j|||_0^2) + \frac{\gamma}{2} |||\nabla \xi_{j,\theta}|||_0^2 \leq (C_1 + \frac{C_2}{h^2}) (|||\chi_{j+1}|||_0^2 + |||\chi|||_0^2) \\
+ C_3 (|||\nabla \chi_{j+1}|||_0^2 + |||\nabla \chi_j|||_0^2) \\
+ C_4 h^2 (|||\nabla^2 \chi_{j+1}|||_0^2 + |||\nabla^2 \chi_j|||_0^2) \\
+ C_5 (|||\xi_{j+1}|||_0^2 + |||\xi_j|||_0^2) + C_7 \Delta t^2 |||\rho_{j,\theta}|||_0^2 \\
+ C_6 (|||(\chi_t)_{j+1}|||_0^2 + |||(\chi_t)_j|||_0^2).
\end{aligned}
$$

Multiplying by $2\Delta t$ and then summing for $j = 0, \ldots, N-1$, we obtain

$$
\begin{aligned}
|||\xi_N|||_0^2 - |||\xi_0|||_0^2 + \Delta t \gamma \sum_{j=0}^{N-1} |||\nabla \xi_{j,\theta}|||_0^2 \leq C \Delta t \sum_{j=0}^{N} [(C_1 + \frac{C_2}{h^2}) |||\chi_j|||_0^2 + C_6 |||(\chi_t)_j|||_0^2 \\
+ C_3 |||\nabla \chi_j|||_0^2 + C_4 h^2 |||\nabla^2 \chi_j|||_0^2] \\
+ C_5 \Delta t \sum_{j=0}^{N} |||\xi_j|||_0^2 + C_7 \Delta t^3 \sum_{j=0}^{N} |||\rho_{j,\theta}|||_0^2.
\end{aligned}
$$

If Δt is sufficiently small we obtain, by Gronwall's lemma,

$$\begin{aligned}\|\xi_N\|_0^2 + \Delta t\gamma \sum_{j=0}^{N-1} \|\nabla \xi_{j,\theta}\|_0^2 \leq C\|\xi_0\|_0^2 + C_7 \Delta t^3 \sum_{j=0}^{N-1} \|\rho_{j,\theta}\|_0^2 \\ + C\Delta t \sum_{j=0}^{N} [(C_1 + \frac{C_2}{h^2})\|\chi_j\|_0^2 + C_6\|(\chi_t)_j\|_0^2 \\ + C_3\|\nabla \chi_j\|_0^2 + C_4 h^2 \|\nabla^2 \chi_j\|_0^2].\end{aligned}$$

Using approximation properties and the choice of the initial condition, we get for $\theta \in (0, 1]$:

$$\begin{aligned}\|\xi_N\|_0^2 + \Delta t\gamma \sum_{j=0}^{N-1} \|\nabla \xi_{j,\theta}\|_0^2 \leq C\frac{h^{2\mu-2}}{r^{2s-4}} \sum_{j=0}^{N} \Delta t(\|u_j\|_{H^s}^2 + \|(u_t)_j\|_{H^{s-1}}^2) \\ + \hat{C}\Delta t^2 \sum_{j=0}^{N-1} \Delta t \|u_{tt}\|_{L^\infty((t_j, t_{j+1}); H^1)}^2.\end{aligned}$$

For $\theta = 0$, we have

$$\begin{aligned}\|\xi_N\|_0^2 + 2\Delta t\gamma \sum_{j=0}^{N-1} \|\nabla \xi_{j,\theta}\|_0^2 \leq C\frac{h^{2\mu-2}}{r^{2s-4}} \sum_{j=0}^{N} \Delta t(\|u_j\|_{H^s}^2 + \|(u_t)_j\|_{H^{s-1}}^2) \\ + \hat{C}\Delta t^4 \sum_{j=0}^{N-1} \Delta t \|u_{ttt}\|_{L^\infty((t_j, t_{j+1}); H^1)}^2.\end{aligned}$$

□

6 Conclusion

A continuous time and a family of discrete time Discontinuous Galerkin procedures have been formulated for nonlinear parabolic problems. Optimal rates of convergence in $L^2(H^1)$ for the continuous time method and $l_2(H^1)$ were derived. As far as the authors are aware, these are the first optimal H^1 estimates established for the DG method for parabolic problems. These estimates can also be extended to treat the addition of a constraint on the integral average of the jump on edges. Computational results are under development.

References

1. Arnold D.N.: An interior penalty finite element method with discontinuous elements. SIAM J. Numer. Anal. **19** (1982) 724–760
2. Darlow B., Ewing R., Wheeler M.F.: Mixed finite element methods for miscible displacement problems in porous media. Soc. Pet. Eng. report SPE 10500 (1982)

3. Douglas J., Dupont T.: Galerkin methods for parabolic equations. SIAM J. Numer. Anal. **7** (1970) 575–626
4. Douglas J., Dupont T.: Interior penalty procedures for elliptic and parabolic Galerkin methods. Lecture Notes in Physics **58** (1976) 207–216
5. Douglas J., Wheeler M.F., Darlow B.L., Kendall R.P.: Self-adaptive finite element simulation of miscible displacement in porous media. Computer Methods in Applied Mechanics and engineering. **47** (1984) 131–159
6. Nitsche J.A.: Über ein Variationspringzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind. Abh. Math. Sem. Univ. Hamburg. **36** (1971) 9–15
7. Oden J.T., Babuška I., Baumann C.E.: A discontinuous hp finite element method for diffusion problems. J. Compu. Phys. **146** (1998) 491–519
8. Rivière B., Wheeler M.F, V. Girault: Part I. Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. TICAM Report. **99-09** (1999)
9. Rivière B., Wheeler M.F., Baumann C.E.: Part II. Discontinuous Galerkin method applied to a single phase flow in porous media. TICAM Report. **99-10** (1999)
10. Wheeler M.F.: A priori L_2 error estimates for Galerkin approximations to parabolic partial differential equations. SIAM J. Numer. Anal. **10** (1973) 723–759
11. Wheeler M.F.: An elliptic collocation-finite element method with interior penalties. SIAM J. Numer. Anal. **15** (1978) 152–161

Part III

Contributed Papers

Parallel Iterative Discontinuous Galerkin Finite-Element Methods

Dan Aharoni * and Amnon Barak

Institute of Computer Science
The Hebrew University of Jerusalem
Jerusalem 91904, Israel

Abstract. We compare an iterative asynchronous parallel algorithm for the solution of partial differential equations, with a synchronous algorithm, in terms of termination detection schemes and performance. Both algorithms are based on discontinuous Galerkin finite-element methods, in which the local elements provide a natural decomposition of the problem into computationally-independent sets. We demonstrate the superiority of the asynchronous algorithm over the synchronous one in terms of the overall execution time. Our goal is to persuade parallel developers that it is worthwhile to implement the more complex asynchronous algorithm.

1 Introduction

Controlling the flow of information between parallel tasks is a major problem in distributed applications. In the case of finite-element numerical methods for the solution of partial differential equations, it has a significant effect on the efficiency. Since these methods were originally developed for serial, single-processor computers, substantial effort has been made to convert them into efficient distributed algorithms, managed by synchronous and asynchronous techniques.

We present a synchronous parallel algorithm, and compare it with the new asynchronous algorithm introduced in [1]. Both algorithms are based on a combination of the following three methods:

1. The discontinuous Galerkin method. This method is used as a divide-and-conquer strategy, where the original problem is divided into independent tasks, which can easily be adapted to parallel computing.
2. Domain Decomposition (DD) methods, enabling coarse grain parallelism.
3. Parallel iterative synchronous and asynchronous techniques, originally developed for data communication networks, e.g. broadcasting. These algorithms implement efficiently the above-mentioned two numerical methods into parallel computation, mainly for controlling the information flow between the parallel tasks and for termination detection [10] [12].

* Corresponding author E-mail: danaha@cs.huji.ac.il

2 The Numerical Methods

The combination of the DD and the discontinuous methods has many advantages. They are fully parallelizable and efficient in load balancing, since the calculation procedure follows an element-by-element iteration in each sub-domain, thus allowing the movement of elements from one sub-domain to another during computation. This is especially useful in complex adaptive h-p meshes, in which the work load is constantly changing [9]. In this case, adaptation is easily implemented, since no continuity is *a priori* required between the element boundaries, which, in turn, makes it unnecessary to use transition elements (nonconforming meshes). Moreover, when using the discontinuous element-by-element method, it is not necessary to assemble a global matrix, and to reassemble it for every adaptive change in the mesh (particularly in nonlinear problems), thus making the calculation flexible and robust. [7]. We note that this method can be implemented on elliptic (e.g. our test model), parabolic and hyperbolic problems [5] with minimal changes to the algorithm.

Of course, the discontinuous element-by-element method has some drawbacks. The first one is that every "point" is calculated from many directions, which may be an advantage in a discontinuous problem, but not necessarily so in a regular, continuous one. Another drawback is having characteristics similar to relaxation methods [11], where the convergence (in elliptic problems) is much slower than other, more popular, iterative methods such as gradient methods (PCG, GMRES etc.) and the powerful Multi-grid methods [8]. Since our main goal is to compare the asynchronous algorithm with the synchronous one, and not compare convergence schemes, we did not implement alternative solvers. Improving the convergence of the element-by-element method may be accomplished by using it as a preconditioner in gradient methods or as one calculation phase in the Multi-grid methods.

3 Synchronous and asynchronous methods

In any iterative parallel algorithm, it is necessary to coordinate, to some extent, the activities of the different processors. In our case, the coordination is the communication of the boundary conditions between the sub-domains. When the processor, solving the sub-domain, depends on the results from its neighbouring processors in previous phases, we refer to the algorithm as *synchronous*. An *asynchronous* algorithm requires only that the update is carried out with knowledge of the results of some past update of the neighbouring processes, not necessarily the most recent update. The processes are not required to wait for predetermined data to become available. They keep computing, and trying to solve the problem, with whatever data they have at that point in time. There is, however, a possibility that iterations performed on the basis of outdated information will not be effective and may even be

counterproductive. The main difficultly with asynchronous computation is to devise an efficient and safe termination-detection scheme, where the global termination is detected based on the local termination conditions at the processes. We use a scheme that is based on Dijkstra's termination detection algorithm [6], originally developed for data communication networks, e.g., broadcasting.

When comparing the synchronous and asynchronous schemes, we find that the overall execution time is substantially reduced in the asynchronous implementation. This may be accounted for by several reasons: In the case of a synchronous algorithm, the timing of operations at each processor is completely determined and is enforced by the synchronizer. The efficiency of the calculation is affected by the idle time spent waiting for processes with differing work loads to catch up with one another. However, in the asynchronous algorithm, the time and order in which the processes are computing, receiving, and transmitting messages vary widely, since the processes are not required to wait for predetermined data to become available. Although this causes the communication requirements to exceed those of the synchronous algorithm, it allows much more flexibility in the use of information, causing the efficient computation of the asynchronous scheme.

4 Parallel implementation and termination detection

The implementation of the parallel strategy is as follows: A process is spawned for computing each sub-domain. A communication phase follows, where each process sends (and receives) the interface sub-domain boundary values to its nearest neighbours. In the asynchronous scheme, all processes are equal, forming a one-level hierarchy. This equality eliminates the bottlenecks encountered in the synchronous *master-slave* type models, where the scalability is bound by the special role of the master process.

The termination detection scheme is based on Dijkstra's termination detection algorithm [6], where a process may be in one of two states: *active*, or *inactive*, and each message is acknowledged with an *ack* message by the receiver.

In the *active* state, the process is in its computation and communication phase. In the *inactive* state, the process can only receive messages. When a message is received from one of the neighbours, the process becomes *active*. Until the process enters the *inactive* state, the message is called the *critical message*, and the sending process is called the *parent*.

An *active* process becomes *inactive* when the local termination condition is satisfied, the process has transmitted prior to that time an *ack* for each message it had received, except for the critical message, and had received an *ack* for each message that it had transmitted.

Initially one of the processes, named the *initiator*, is *active*, while all the other processes are *inactive*. Since in our implementation all processes are

equal, forming a one-level hierarchy, the *initiator* is arbitrarily chosen to be the first process spawned. Global termination occurs as soon as the *initiator* becomes *inactive*.

Fig. 1 depicts the pseudo-code of each of the parallel processes in the asynchronous scheme.

`work` is the iterative procedure, performing the calculation phase. It starts with a zero uniform initial guess, and ends when either the error is smaller than ϵ, or the number of iterations has reached a maximum number. The error is defined as: $error = max|u^k - u^{k-1}|$ (infinity norm), where u^k corresponds to the solution at the k-th iteration.

`send_domain_bc()`: Sends boundary conditions to all neighbouring domains.

`blocking_receive()`: Receives messages from the neighbouring domains.

```
ACTIVE = ( me == initiator ) ; /* Initially, only the initiator is ACTIVE */
do{
    if ( ACTIVE ){
        do{
            converged = work() ; /* Performs the calculation phase */
            if ( no. of processes > 1 ){
                send_domain_bc() ;
                blocking_receive() ;
            }
        } while ( not converged or ACK_counter != 0 );
        send ACK with CRITICAL message to PARENT
        ACTIVE = FALSE ;
    } /* end of ACTIVE state */
    else{  /* INACTIVE state */
        inactive_blocking_receive() ;
        ACTIVE = TRUE ;
    } /* end of INACTIVE state */
} while  ( not TERMINATE ) ;
```

Fig. 1. *Pseudo-code of the asynchronous scheme.*

For the synchronous termination detection scheme we choose a straightforward global synchronizer, which is accomplished by using a *master-slave paradigm*. The master controls the global phase propagation by waiting until local termination condition is satisfied in *all* the slave processes and the communication with all neighbours has ended, before going on to the next phase. Global termination is detected by the master, and a notification is sent to all slaves.

5 Performance

This section compares the performance of the synchronous parallel algorithm with the asynchronous algorithm. Computations were done on a 2D elliptic test problem with 10 000 elements. We used a simple-structured mesh topology with linear elements of variable size, although our implementation allowed a variable spectral element order.

5.1 The MOSIX cluster computing system

The implementation and numerical experiments of the presented algorithms were performed on the MOSIX [2] system, which is an enhancement of LINUX for cluster computing [3],[4]. The MOSIX system is designed to respond to variations in the resource usage among the PC's (nodes), by migrating processes from one node to another, preemptively and transparently, for load-balancing and memory sharing. Our system consists of almost 100 Pentium 200-400MHz nodes that are connected by Ethernet and the Myrinet LANs, creating a powerful multi-user time-sharing parallel environment.

5.2 The model problem

Consider a scalar, first order mixed elliptic equation system, with mixed Dirichlet and Neumann type boundary conditions,

$$\begin{aligned} -\nabla \cdot (p) &= f \quad \text{in } \Omega, \\ p &= a\nabla u \quad \text{in } \Omega, \end{aligned} \tag{1}$$

$$u = \bar{u} \text{ on } \partial\Omega_u, \quad \text{and} \quad p = \bar{p} \text{ on } \partial\Omega_p.$$

Partitioning Ω into a finite number of regularly-shaped elements Ω^e, multiplying Eq. (1) by appropriate test functions, and integrating over these subintervals, yields the following discontinuous Galerkin formulation of Eq. (1):

$$\int_{\Omega^e} [(p-a\nabla u)\delta p+(-\nabla\cdot p-f)\delta u\,]\, d\Omega^e+\oint_{\partial\Omega_u^e}(u-\bar{u})\delta p\, ds+\oint_{\partial\Omega_p^e}(p-\bar{p})\delta u\, ds = 0. \tag{2}$$

The boundary integrals are used as jump-conditions for the vectors u and p, thus applying the discontinuities on the inter-element boundaries. The vectors $\bar{u}$ and $\bar{p}$ are either the nodal values on the interface of the neighbouring elements, or the global boundary conditions for elements on the global boundary. Substituting base functions into Eq. (2) leads to a set of linear equations that are solved in each element, to obtain the nodal values of u and p.

5.3 Experimental results

We solved Eq. (1) with the following boundary conditions and parameters:

$$u(0,y) = p(x,1) = p(x,0) = 0.0\ ; p(1,y) = 1.0\ ; f(x) = 1.0 + 1.0\ x\ , a = 1.0 \quad (3)$$

Let the true error be defined as the difference between the analytic and the numerical solutions of the test problem. The executions were performed on a 10 000 (100×100) element problem. In all cases, the maximal displacement (u) true error was 0.0039, and the maximal flux (p) true error was 0.0033. All the executions were performed on a 25 (identical) node MOSIX system.

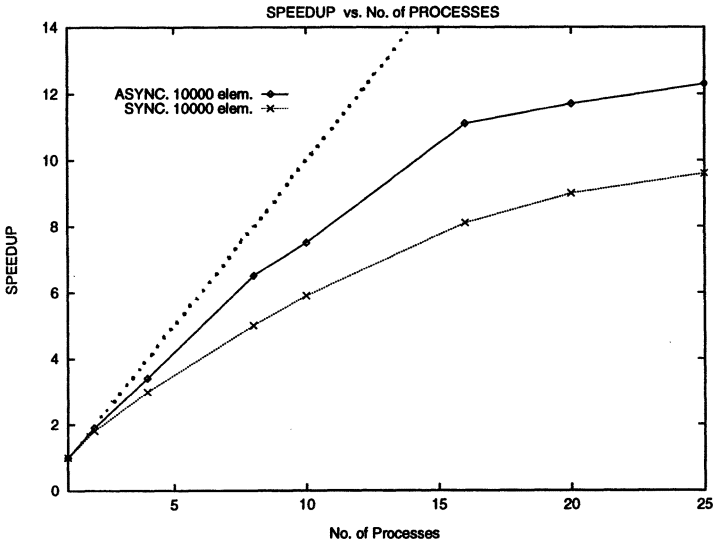


Fig. 2. *Speedup vs. number of processes*

Scalability is measured using the speedup parameter, defined as the ratio between the measured time of one process and the measured time of a number of processes. The speedups obtained for the executions of the asynchronous and synchronous parallel algorithms are shown in Fig. 2. It can be seen from the figure that the speedup for a small number of processes is almost linear, and that it decreases with the increase of the number of processes for both algorithms. This is attributed to the decrease in the granularity of each of the sub-domains. At approximately 25 processes, the speedup is reduced to a point where the execution time is of the same order of magnitude as the total overhead. Overhead consists of communication overhead, plus the time required for initializing the problem and performing the garbage collection at the end of the calculation. For larger problems, where the granularity is

sufficiently coarse, this break-point occurs at a higher number of processes, and for smaller problems break-point occurs at a lower number of processes.

Fig. 2 also shows that the asynchronous algorithm is more efficient than the synchronous one for all number of processes, as predicted in section 3.

6 Conclusions

We have presented a parallel implementation of the discontinuous finite-element method, using asynchronous and synchronous parallel methods. Our numerical experiments were performed on the MOSIX cluster computing system, and used a 2D elliptic partial-differential equation as a model-problem. Results demonstrate the superiority of the asynchronous algorithm over the synchronous one in terms of the overall execution time. Results also show that the algorithms appear to be a powerful tool for achieving high parallel efficiencies. They scale up well and yield good speedups, for coarse granularity cases.

Acknowledgment

This work was supported by grants from the Israeli Ministry of Defense and the Ministry of Science.

References

1. Aharoni D., Barak A., *Asynchronous Parallel Discontinuous Finite Element Method*, Proc. Int. Conference on High-Performance Computing and Networking (HPCN Europe '98), 347-357, Amsterdam, April 1998.
2. Barak A., La'adan O., *Performance of the MOSIX Parallel System for a Cluster of PC's*, Proc. Int. Conference on High-Performance Computing and Networking (HPCN Europe '97), 624-635, Vienna, April 1997.
3. *MOSIX Cluster Computing for Linux*, The Hebrew University of Jerusalem, http://www.mosix.cs.huji.ac.il, 1999.
4. Barak A., La'adan O., Shiloh A., *Scalable Cluster Computing with MOSIX for LINUX*, Proc. 5-th Annual Linux Expo, 95-100, May 1999.
5. Bar-Yoseph, P., Elata D., *An Efficient L2 Galerkin Finite Element Method for Multi-Dimensional Nonlinear Hyperbolic System*, Int. J. Numer. Methods Eng., 29, 1229-1245, 1990.
6. Bertsekas, D.P., Tsitsiklis J.N., *Parallel and Distributed Computation, Numerical Methods*, Prentice-Hall, 1989.
7. Layton, W.J., Maubach, J.M., Rabier, P. J., *Robustness of an Elementwise Parallel Finite Element Method for Convection-Diffusion Problems*, SIAM J. Sci. Comput., Vol 19, No. 6, pp 1870-1891, Nov. 1998.
8. McCormick S, *Multigrid Methods*, Volume 3 of the SIAM Frontiers Series, SIAM, Philadelphia, 1987.

9. Oden J.T., Abani P., Yusheng F., *Domain Decomposition for Adaptive hp Finite Element Methods*, Proc. Seventh Int. Conf. on Domain Decomposition, 180, 1994.
10. Savari S.A., Bertsekas D.P., *Finite termination of asynchronous iterative algorithms*, Parallel Computing 22, 1, 39-56, 1996.
11. Smith B., Bjørstad P., Gropp W., *Domain Decomposition Parallel Multilevel Methods for Elliptic PDE*, Cambridge University Press, 1996.
12. Üresin A., Dubois M., *Asynchronous Iterative Algorithms: Models and Convergence, Advances in Parallel Algorithms*, Kronsjö L., Shumsheruddin D., Ed, John Wiley & Sons, 302-342, 1992.

A Discontinuous Projection Algorithm for Hamilton Jacobi Equations *

Steeve Augoula** and Rémi Abgrall***

Mathématiques Appliquées de Bordeaux, Université de Bordeaux I
351 Cours de la Libération, 33 405 Talence Cedex, France

Abstract. We present a class of numerical schemes for the numerical integration of first order Hamilton Jacobi equations. The method can be considered as Discontinuous Galerkin scheme, the viscosity solution is directly adapted into the numerical scheme, contrary to other authors.

1 Introduction

We are interested in computing the numerical solution of Cauchy problem for the first order Hamilton Jacobi equation :

$$u_t + H(x, \nabla u) = 0,\ u(x,0) = u_0(x),\ x \in \mathbb{R}^2,\ u\ : \mathbb{R}^2 \times \mathbb{R}^+ \to \mathbb{R} \quad (1)$$

In (1), the Hamiltonian H is uniformly continuous on $\mathbb{R}^2$. The solution u is considered in the viscosity sense [5,7]. The initial condition u_0 lies in $BUC(\mathbb{R}^2)$, the set of bounded, uniformly continuous functions defined in $\mathbb{R}^2$. The viscosity solution u is in $BUC(\mathbb{R}^2 \times [0,T[)$, the set of bounded, uniformly continuous functions defined in $\mathbb{R}^2 \times [0,T[$ for any $T > 0$.

The discretization of (1) has already been considered when the computational domain is discretized by a Cartesian mesh [6] and a triangular mesh [3] but in both cases, the schemes were first order accurate. High order in space and time schemes are also described in an ENO way [1,9] but on Cartesian meshes. In [8] a Discontinous Galerkin scheme is developed. It uses a formal analogy between some conservation laws and Hamilton Jacobi equations.

In many applications (Geometrical Optics, Shape deformations, Mesh generation), it is natural to approximate (1) on unstructured meshes. Many times, the question of accuray is important.

In the present paper, we propose a technique between DGM and ENO methods. We are looking for an approximation of the solution of (1) where :

1. u_0 is a piecewise polynomial $\mathbb{P}^k$, continuous function,
2. u^{n+1} is a $\mathbb{P}^k$ function computed from u^n by first, solving the Cauchy problem (1) with a $\mathbb{P}^k$ interpolant of u^n as initial condition and second, projecting the exact solution onto $\mathbb{P}^k$.

* Research was supported in part by SNPE and CNES
** augoula@math.u-bordeaux.fr
*** abgrall@math.u-bordeaux.fr

The paper is organized as follows. In section 2, we give some properties of the exact evolution operator of (1) which allows us to construct a numerical Hamiltonian for a convex or concave H. In section 3, we build an approximation of the viscosity solution of one dimensional Cauchy problem related to (1). A natural extension to the two dimensional case for triangular meshes and $\mathbb{P}^2$ Lagrange interpolant follows in section 4. Numerical examples are provided in section 5 to illustrate the method. In particular we will observe stability, monotonicity and the expected accuracy is also reached.

2 Preparation

2.1 The Hopf formulas

For a convex (resp. concave) function ψ, one defines its Legendre transform, $\psi^*(p) = \sup_{y\in\mathbb{R}^2}\{p.y - \psi(y)\}$, resp. $\psi^*(p) = \inf_{y\in\mathbb{R}^2}\{-p.y - \psi(y)\}$, $\forall p \in \mathbb{R}^2$.

An analytical expression of the solution of (1) is given in [4] for the particular case where u_0 is uniformly continuous and :

- for convex Hamiltonian,

$$u(x,t) = \inf_{y\in\mathbb{R}^2}\left\{u_0(y) + tH^*(\frac{x-y}{t})\right\} \text{ for all } x \in \mathbb{R}^2 \qquad (2)$$

- for a concave Hamiltonian, $u(x,t) = \sup_{y\in\mathbb{R}^2}\left\{u_0(y) + tH^*(\frac{y-x}{t})\right\}$ for all $x \in \mathbb{R}^2$

Equation (2) gives an analytical formula for the evolution operator associated to (1). It has the following properties. Recall that for any $a \in \mathbb{R}$, $a^+ = \max(a,0)$ and $a^- = \min(a,0)$, $|a| = a^+ - a^-$.

Proposition 1. (Crandall-Lions)
Let u_0 and v_0 be two uniformly continuous, uniformly Lipschitz function on $\mathbb{R}^2$. We denote by u and v the solutions of (1) with the initial conditions u_0 and v_0. For any time $t > 0$, the evolution operator $S(t) : u_0 \mapsto u(.,t)$ has the following properties :

1. $\|(S(t)u_0 - S(t)v_0)^+\|_\infty \leq \|(u_0 - v_0)^+\|_\infty$,
2. $\|S(t)u_0 - S(t)v_0\|_\infty \leq \|u_0 - v_0\|_\infty$,
3. *if L' a Lipschitz constant for u_0, then it is a Lipschitz constant for $S(t)u_0$,*
4. $\inf_{y\in\mathbb{R}^2}\{u_0(y) - tH(0)\} \leq S(t)u_0(x) \leq \sup_{y\in\mathbb{R}^2}\{u_0(y) - tH(0)\}$ *for any* $x \in \mathbb{R}^2$

2.2 The Cone of Dependence

To evaluate expressions like (2) it is interesting to reduce the minimization domain each time it is possible. We recall another property of the exact evolution operator S defined above due to Crandall-Lions in [5].

Proposition 2.
Let $u_0, v_0 \in BUC(\mathbb{R}^2)$. Let $u, v \in BUC(\mathbb{R}^2 \times [0,T[)$ be solutions of (1) on $Q_T = \mathbb{R}^2 \times [0,T[$ with the initial conditions u_0 and v_0. Let L a Lipschitz constant for $H(x,p)$ in p. Let R a given positive real such that $u_0 \equiv v_0$ on $\bar{B}(O,R)$[1]*. Then, for all $t < \frac{R}{L}$, $u(.,t) \equiv v(.,t)$ on $\bar{B}(O, R - Lt)$.*

We see that $u(x,t)$ depends on u_0 in $\bar{B}(O,R)$ only if $R - Lt > 0$ i.e $t < \frac{R}{L}$. Then one can write the expression (2) as follows : if $u_0 \in BUC(\mathbb{R}^2)$ then for all $x_0 \in \mathbb{R}^2$, $u(x,t) = \inf_{y \in \bar{B}(x_0,R)} \{u_0(y) + tH^*(\frac{x-y}{t})\}$, $\forall x \in \bar{B}(x_0, R - Lt)$, $\forall t < \frac{R}{L}$.
The main advantage of this formulation is that we can construct a numerical scheme based on the exact evolution operator S as it has been shown in [3] for the specifical case of piecewise linear and continuous data. The next section aims at constructing a high order scheme for the one dimensional Cauchy problem.

3 Construction of the Numerical Scheme in $\mathbb{R}$

We consider a mesh $(x_{j+\frac{1}{2}})_{j \in \mathbb{Z}}$ and set : $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$, $x_j = \frac{1}{2}(x_{j-\frac{1}{2}} + x_{j+\frac{1}{2}})$, $h_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$.
In all the numerical experiments, the mesh is uniform.

We also consider the space of continuous piecewise polynomials $V_h^k = \{v \in C^0(\mathbb{R}) : v_{|I_j} \in \mathbb{P}^k(I_j), \forall j \in \mathbb{Z}\}$ where $\mathbb{P}^k$ stands for the set of polynomials of degree at most k.
For any $u_0 \in V_h^k$, we have for any $x \in [z - h + Lt, z + h - Lt]$

$$u(x,t) = \inf_{y \in [z-h, z+h]} \left\{ u_0(y) + tH^*(\frac{x-y}{t}) \right\} \tag{3}$$

In general, (3) does not define an element of V_h^k, hense we have to project u onto V_h^k. What is the projector ?

To simplify the text, we assume $k = 2$. In each interval I_j, a polynomial of degree 2 is uniquely determined by its values at $x_{j-\frac{1}{2}}$, $x_{j+\frac{1}{2}}$ and x_j. Since we need to enforce the continuity requirement, we use the Lagrange projector : we only need to apply (3) for $x = x_{j-\frac{1}{2}}$, $x = x_{j+\frac{1}{2}}$ and $x = x_j$. There are two different cases :

[1] $\bar{B}(O,R')$ is the closed ball of center O and radius $R' > 0$.

1. $x = x_{j-\frac{1}{2}}$, $x = x_{j+\frac{1}{2}}$ then the minimisation is carried over two intervals, namely $[x_{j-\frac{3}{2}}, x_{j-\frac{1}{2}}]$ and I_j if for example $x = x_{j-\frac{1}{2}}$ and under the CFL condition $\Delta t \leq \frac{h_j}{2L}$.
2. $x = x_j$ the minimisation is carried out only on I_j under the same CFL condition.

Since $k = 2$ the minimisation can be done exactly. The scheme is formally second order accurate. In the numerical examples, no limitation on the solution (to avoid the creation of over or under shoots) has been applied.

4 Extension to the Two Dimensional Case

Here, we present an extension of above construction for the Cauchy problem (1) and the $\mathbb{P}^2$ Lagrange interpolation. We first begin to give some notations. One consider a triangulation of $\mathbb{R}^2$ named $\mathcal{T}$. The triangles are denoted by T, their vertices are i_1, i_2, i_3 and mid edges i_4, i_5, i_6 as it is shown in Fig. 1. For each mesh point M_i we consider the n_i angular sectors $\Omega_1^i, \cdots, \Omega_{n_i}^i$ meeting at point M_i. We also define the cells $\mathcal{C}_{M_i}$ made of the triangles associated to the n_i angular sectors (see Fig. 1).

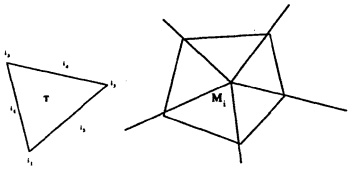


Fig. 1.

Construction of the Stencil

For T a triangle of $\mathcal{T}$, we have to evaluate u_{i_k} using the following stencil :
$\mathcal{S}_{i_k} \equiv \mathcal{C}_{i_k}$ for $k = 1, \cdots, 3$
or $\mathcal{S}_{i_k} \equiv T \cup T'$ for $k = 4, \cdots, 6$ for two adjacent triangles containing i_k.
Thanks to Proposition 2, the solution $\tilde{u}^{n+1}$ of (1) with initial data u^n, on T can be computed, if Δt is small enough, only with the knowledge of u^n in a neighborhood of T. Since we project the exact solution on V_h^2, we need only to know the value of $\tilde{u}^{n+1}$ at the vertices and mid points of T. Namely :

$$\tilde{u}_{i_k}^{n+1} = \begin{cases} \min_{l \in \{1,\cdots,n_{i_k}\}} \inf_{y \in \Omega_l^{i_k}} \left\{ u^n(y) + \Delta t H^* \left(\frac{x-y}{\Delta t} \right) \right\}, \ \forall\, k \in \{1, \cdots, 3\} \\ \min_{l=1,2} \inf_{y \in \Omega_l^{i_k}} \left\{ u^n(y) + \Delta t H^* \left(\frac{x-y}{\Delta t} \right) \right\}, \ \forall\, k \in \{4, \cdots, 6\} \end{cases} \tag{4}$$

and the formula is exact provided that $\Delta t \leq \frac{h}{2L}$, $h = \min(d(i_k, \partial \mathcal{C}_{i_k})_{k=1,2,3}$, $d(i_k, \partial T_i)_{k=4,5,6})$ (see Fig. 2) with $d(.,.)$ the distance function.

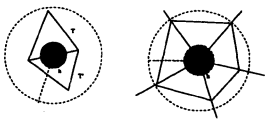


Fig. 2. Cone of dependence.

5 Numerical Examples

Example 1. One dimensional Burger's equation :

$$\begin{cases} u_t + \frac{(u_x+1)^2}{2} = 0 \quad -1 < x < 1 \\ u(x,0) = -\cos(\pi x) \end{cases} \tag{5}$$

with periodic conditions.

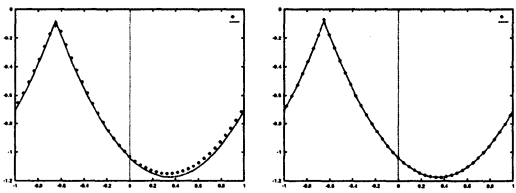


Fig. 3. $N = 42$ to the left P^1 and to the right P^2.

The interest of this test is that the solution develops an unsteady discontinuous derivative which is not aligned on the mesh. This is a stability and monotony test. In Fig. 3, we show the sharp corner-like numerical solution with 42 elements obtained with $\mathbb{P}^k$ for $k = 1,\ 2$ at $t = 0.35$. Here the solid line is the reference solution.

Example 2. Eikonal equation :

$$\begin{cases} \|\nabla u\| = 1 \quad x \in \Omega \\ u(x,0) = 0 \end{cases} \tag{6}$$

where Ω represents a square with different holes (see Fig. 4). The solution of (6) is computed as the steady solution of :

$$\begin{cases} u_t + \|\nabla u\| - 1 = 0 \quad x \in \Omega,\ t > 0 \\ u(x,t) = 0 \quad x \in \partial\Omega,\ t > 0 \\ u(x,0) = 0 \quad x \in \Omega \end{cases} \tag{7}$$

We enforce numerically the boundary condition by (see [2]) $u^{n+1}(x) = \max(\tilde{u}^{n+1}(x), 0)$ with x a node or a mid point and $u^{n+1}(x)$ the predicted value of the interior scheme. We show two different tests for this last example. The meshes are shown in Fig. 4 and the solutions are shown in Fig. 5.

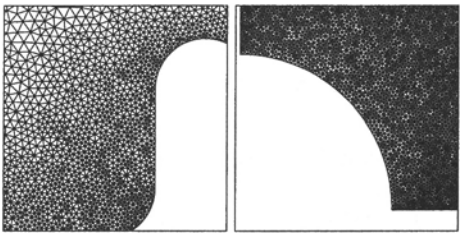


Fig. 4. Zoom meshes.

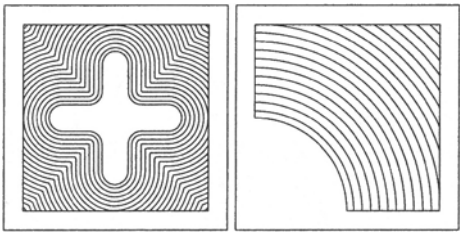


Fig. 5. isolines of u.

Comments. Our 1D and 2D solutions are clearly monotone. No under- or overshoots exist. This (pleasant) phenomena is not yet completly understood. The scheme is second order accurate, see Table 1 for the test case of Fig. 5, right where the boundary solution allows a C^∞ solution. The meshes are not obtained by successive refinement, but are independent one of the other.

	$\mathbb{P}^1$				$\mathbb{P}^2$			
hmax	L^1 error	Order	L^∞ error	Order	L^1 error	Order	L^∞ error	Order
1.09 E-1	0.11 E-1	-	0.75 E-2	-	0.35 E-2	-	0.15 E-2	-
5.25 E-2	0.44 E-2	1.25	0.31 E-2	1.21	0.92 E-3	1.83	0.41 E-3	1.77
2.94 E-2	0.22 E-2	1.19	0.16 E-2	1.14	0.26 E-3	2.18	0.13 E-3	1.98
1.88 E-2	0.15 E-2	0.86	0.99 E-3	1.07	0.12 E-3	1.73	0.58 E-4	1.80

Table 1. Accuracy for the boundary C^∞ solution. *hmax* is the largest edge of the mesh.

6 Conclusion

We have presented the construction of a high order scheme for first order Hamilton-Jacobi equations. After some numerical illustrations, we note that the new scheme is stable and numerically converges to the viscosity solution. For the moment, we are not able to prove these results. Nevertheless, under some conditions on the Lipschitz constant of the exact solution we are able to prove a stability result for L^∞ norm. A topic of current research is the simplification of this scheme, we also want to extend it to non convex/concave Hamiltonian.

References

1. A. Harten, B. Engquist, S. Osher and S. Chakravarthy. Uniformly High Order Accurate Essentially Non-oscillatory Scheme,III. *J. Comput. Phys.*, **71**, (1987).
2. R. Abgrall. Numerical Discretization of Boundary Conditions for Hamilton-Jacobi Equations. *SIAM J. Numer. Anal., submitted.*
3. R. Abgrall. Numerical Discretization of First-Order Hamilton-Jacobi Equations on Triangular Meshes. *Comm. Pure Appl. Math.*, pages 1339–1373, (1996).
4. M. Bardi and L.C. Evans. On Hopf's Formulas for Solutions of Hamilton-Jacobi Equations. *Nonlinear Analysis, Methods and Applications*, **8**(11):1373–1381, (1984).
5. M.G. Crandall and P.L. Lions. Viscosity Solution of Hamilton-Jacobi Equations. *Transaction of the American Mathematical Society*, **277**(1), May (1983).
6. M.G. Crandall and P.L. Lions. Two approximations of solutions of Hamilton-Jacobi equations. *Math. Comp.*, **43**:1–49, May (1984).
7. M.G. Crandall L.C. Evans and P.L. Lions. Some Properties of Viscosity Solutions of Hamilton-Jacobi equations. *Transaction of the American Mathematical Society*, **282**(2), April (1984).
8. C. Hu and C. W. Shu. A Discontinuous Galerkin scheme for Hamilton-Jacobi Equations. *SIAM J. Sci. Comp.*, to appear.
9. S. Osher and C.-W. Shu. High-Order Essentially Non-oscillatory schemes for Hamilton-Jacobi equations. *SIAM J. Numer. Anal.*, **28**(4), (1991).

Successes and Failures of Discontinuous Galerkin Methods in Viscoelastic Fluid Analysis

Arjen C.B. Bogaerds, Wilco M.H. Verbeeten, and Frank P.T. Baaijens

Eindhoven University of Technology
Faculty of Mechanical Engineering
Materials Technology (http://www.mate.tue.nl/)
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Abstract. To date, the more successful numerical methods in viscoelastic fluid dynamics are based upon the so called Discrete Elastic Viscous Stress Splitting (DEVSS) algorithm [6] together with a suitable form of upwinding of the hyperbolic part of the constitutive equation. An elegant way to perform upwinding on the viscoelastic stress tensor can be found in Discontinuous Galerkin techniques [4]. In particular the recently developed DEVSS/DG version [1], has proven to be successful in analyzing viscoelastic fluid flow problems in both smooth and non-smooth geometries. A particularly attractive feature of DG-based methods is that they allow for an efficient resolution of flow problems with multiple relaxation times, as was demonstrated in Baaijens *et al.* [1] which has recently been extended to three dimensional flows [2].

However, one of the key issues in simulations of viscoelastic flows remains the assessment of temporal stability of the computational method. Especially, increasing elasticity beyond critical values of the Weissenberg number can give rise to numerical instabilities in flows that are otherwise mathematically stable.

1 Introduction

In recent years a lot of research has been performed on the development of numerical tools for viscoelastic flow simulations. The nature of the governing equations, i.e. constitutive equations and conservation of mass and momentum, requires that special attention has to be paid to the numerical solution algorithm. For instance, a major problem that needs to be resolved is the loss of numerical stability of standard Galerkin solution procedures with increasing Weissenberg numbers. To overcome this problem, several stabilizing techniques have been proposed such as application of Petrov-Galerkin weighting or discontinuous methods, as is common for finite difference schemes.

Another problem is associated with the rheological character of the viscoelastic fluid itself. Besides a proper choice of the constitutive relation, accurate flow analysis of both polymer solutions and melts compels the use of multiple relaxation times. Using mixed finite element methods this results in a very large number of degrees of freedom and therefore solution efficiency,

both in terms of CPU time and memory requirement becomes an important issue.

2 Problem definition

Here, only incompressible, isothermal and inertia-less flows are considered. In the absence of body forces, these flows can be described by a reduced equation for conservation of momentum (1) and conservation of mass (2):

$$-\nabla\cdot\boldsymbol{\sigma} = \mathbf{0}\,, \tag{1}$$

$$\nabla\cdot\boldsymbol{u} = 0\,, \tag{2}$$

with $\boldsymbol{u}$ the velocity field and $\boldsymbol{\sigma}$ the Cauchy stress tensor. In general, the stress tensor is divided into 3 separate contributions following:

$$\boldsymbol{\sigma} = -p\boldsymbol{I} + 2\eta_s\mathbf{D} + \boldsymbol{\tau}\,, \tag{3}$$

with p an isostatic pressure, $\mathbf{D} = \left(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^T\right)/2$ the symmetric rate of deformation, η_s a solvent viscosity and a viscoelastic mode $\boldsymbol{\tau}$. When the Newtonian contribution is omitted, as is the case for most realistic viscoelastic fluids, the Cauchy stress is fully determined by the pressure and the extra stress which, in turn, can be described by a differential type constitutive equation like, for instance the, Upper Convected Maxwell (UCM) model:

$$\lambda\overset{\nabla}{\boldsymbol{\tau}} + \boldsymbol{\tau} = 2\eta\mathbf{D}\,. \tag{4}$$

The upper convected time derivative ($\overset{\nabla}{\boldsymbol{\tau}}$) is defined as:

$$\overset{\nabla}{\boldsymbol{\tau}} = \frac{\partial\boldsymbol{\tau}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{\tau} - \mathbf{L}\cdot\boldsymbol{\tau} - \boldsymbol{\tau}\cdot\mathbf{L}^T\,,$$

with $\mathbf{L} = \nabla\boldsymbol{u}^T$ the velocity gradient and λ, η relaxation time and viscosity for this viscoelastic mode. Since the UCM model fails to describe viscoelastic phenomena (i.e. shear thinning, extensional thickening and second normal stress difference) needed to perform realistic simulations, it is often necessary to replace equation 4 by a nonlinear constitutive equation like the Giesekus model or one of the more flexible recently developed Feta models of Schoonen *et al.* [10] and Peters *et al.* [9].

Due to its mathematical stability, the inertia-less plane Couette flow of an UCM fluid provides an extremely useful tool for the determination of the numerical stability of a computational method. Figure 1 shows a schematic representation of this shear flow where the lower plate is fixed and the parallel upper plate moves with constant velocity. Linear stability of the governing equations (1),(2) and (4) about the steady state solution was first shown by Gorodtsov & Leonov [5]. Superimposing spatially periodic disturbances of the

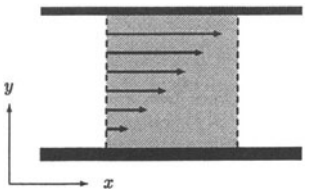


Fig. 1. Plane Couette flow, the upper plate moves with constant velocity.

form $\varepsilon = \varepsilon(y)\exp(\alpha i x + \sigma t)$ to the steady solution results in an eigenvalue problem for σ. Introducing a dimensionless Weissenberg number as $\mathrm{We} = \lambda\dot{\gamma}$ with $\dot{\gamma}$ equal to the constant shear rate, Gorodtsov & Leonov [5] showed that the largest real part of the eigenspectrum behaves as:

$$\mathrm{real}(\sigma) \approx -\frac{1}{2\mathrm{We}},$$

for $\mathrm{We} \gg 1$. Thus, neutral stability is approached for increasing elasticity.

Linearization of the finite element equations about the steady base flow results in a linear differential equation in the superimposed disturbances:

$$\mathsf{M}\,\dot{\varepsilon} + \mathsf{K}\,\varepsilon = 0, \tag{5}$$

with M, K the mass matrix and the Jacobian evaluated about the steady solution and ε the imposed disturbance. Temporal stability of the computational methods is obtained from direct time integration of (5). Hence, the following problem is solved:

$$\left(\mathsf{M} + \Delta t\,\mathsf{K}\right)\varepsilon^{n+1} = \mathsf{M}\,\varepsilon^{n} \tag{6}$$

for every time step. Notice that, eventually, the disturbances should evolve in time at the most unstable decay rate resulting from the Gorodtsov & Leonov [5] analysis.

3 Computational methods

The mixed finite element methods considered in this work are based upon the DEVSS scheme of Guénette & Fortin [6]. Adding extra stabilization to the momentum equation and omitting the boundary integral leaves the modified Stokes relations:

$$\left(\mathbf{D}_v\,,\, 2\beta\eta\left(\mathbf{D} - \bar{\mathbf{D}}\right) + \boldsymbol{\tau}\right) - \left(\nabla\cdot\boldsymbol{v}\,,\, p\right) = 0\,, \tag{7}$$

$$\left(q\,,\, \nabla\cdot\boldsymbol{u}\right) = 0\,, \tag{8}$$

with $(.,.)$ the L_2-inner product on the computational domain (Ω) and $\mathbf{D}_v = (\nabla \boldsymbol{v} + \nabla \boldsymbol{v}^T)/2$. Also, the original 3 field formulation $(\boldsymbol{u}, \boldsymbol{\tau}, p)$ has been transformed into a 4 field formulation $(\boldsymbol{u}, \boldsymbol{\tau}, p, \bar{\mathbf{D}})$ by introducing a discrete approximation for the rate of deformation tensor $(\bar{\mathbf{D}})$. For the UCM case where the Newtonian contribution is absent, this brings back the ellipticity into the momentum equation. As a consequence of this approach an extra projection to solve for $\bar{\mathbf{D}}$ is required, hence:

$$\left(\mathbf{Q}\,,\,\bar{\mathbf{D}} - \mathbf{D}\right) = 0\,. \tag{9}$$

The parameter β in (7) can be varied in order to give optimal results. Furthermore, since the standard Galerkin method gives poor results for high Weissenberg numbers, some form of upwinding is performed on the hyperbolic constitutive equation.

With our DG-method, spatial discretization is formed from projections onto continuous subspaces for all variables except for the extra stress modes. Also, upwinding by discontinuity on the transport of extra stress is performed as was first introduced by Lesaint & Raviart [7]. Defining a jump operator over the element boundaries $[\boldsymbol{\tau}]$, the DEVSS/DG method of Baaijens *et al.* [1] yields the weighted UCM equation:

$$\left(\mathbf{S}\,,\,\lambda\left[\frac{\partial \boldsymbol{\tau}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{\tau} - \mathbf{L}\cdot\boldsymbol{\tau} - \boldsymbol{\tau}\cdot\mathbf{L}^T\right] + \boldsymbol{\tau} - 2\eta\mathbf{D}\right) - \sum_{e=1}^{\text{Elements}} \int_{\Gamma^e_{\text{Inflow}}} \mathbf{S} : \lambda\, \boldsymbol{u}\cdot\boldsymbol{n}\,[\boldsymbol{\tau}]\, d\Gamma = 0\,, \tag{10}$$

with $\boldsymbol{n}$ the unit vector pointing outward normal on the element boundary Γ^e. Using overall Euler implicit time integration and exploiting the fact that the external stress in the jump operator is taken explicitly allows for static condensation of the element stresses and hence results in a drastic reduction of global degrees of freedom [1].

Alternatively, projecting all variables onto continuous subspaces, we consider the DEVSS-G/SUPG scheme in analogy with Szady *et al.* [12]. For reasons of compatibility, instead of discretizing rate of deformation $(\bar{\mathbf{D}})$, a discrete velocity gradient $(\bar{\mathbf{G}})$ is considered in (7) and (9). Application of consistent streamline upwinding to the UCM equation yields a projection of the constitutive equation:

$$\left(\mathbf{S} + \frac{h}{|\boldsymbol{u}|}\,\boldsymbol{u}\cdot\nabla\mathbf{S}\,,\,\lambda\left[\frac{\partial \boldsymbol{\tau}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{\tau} - \bar{\mathbf{G}}\cdot\boldsymbol{\tau} - \boldsymbol{\tau}\cdot\bar{\mathbf{G}}^T\right] + \boldsymbol{\tau} - \eta\left[\bar{\mathbf{G}} + \bar{\mathbf{G}}^T\right]\right) = 0\,, \tag{11}$$

with h some characteristic element size.

A third alternative is based on the ideas of Oden *et al.* [8] and involves the forming of spatial discretization from discontinuous subspaces for all variables

$\boldsymbol{u}, p, \bar{\mathbf{D}}$ and $\boldsymbol{\tau}$. Consider a (non-conforming) mesh Ω with non-overlapping elements Ω^e, a set of inter-element edges S and a Dirichlet boundary Γ_D. Boundary conditions of the type $\boldsymbol{u} = \hat{\boldsymbol{u}}$ are imposed on Γ_D. Other types of boundary conditions can be included but are omitted here. Since all variables are considered to be discontinuous across S, a weak formulation of the momentum equation could be defined as:

$$\sum_{\Omega^e \in \Omega} \left\{ \left(\mathbf{D}_v \, , \, 2\beta\eta\left(\mathbf{D} - \bar{\mathbf{D}}\right) + \boldsymbol{\tau} \right)^e - \left(\nabla \cdot \boldsymbol{v} \, , \, p \right)^e \right\}$$
$$+ \int_S \boldsymbol{n} \cdot \left\langle \mathcal{P}^* \right\rangle \cdot [\boldsymbol{u}] - [\boldsymbol{v}] \cdot \left\langle \mathcal{P}^{**} \right\rangle \cdot \boldsymbol{n} \, d\Gamma + \int_{\Gamma_D} \boldsymbol{n} \cdot \mathcal{P}^* \cdot \boldsymbol{u} - \boldsymbol{v} \cdot \mathcal{P}^{**} \cdot \boldsymbol{n} \, d\Gamma$$
$$= \int_{\Gamma_D} \boldsymbol{n} \cdot \mathcal{P}^* \cdot \hat{\boldsymbol{u}} \, d\Gamma \tag{12}$$

with $\langle \cdot \rangle$ an averaging operator on S, $\boldsymbol{n}$ (one of) the unit normal vector(s), $\mathcal{P}^{**} = 2\beta\eta\left(\mathbf{D} - \bar{\mathbf{D}}\right) + \boldsymbol{\tau} - p\mathbf{I}$ and $\mathcal{P}^* = 2\beta\eta\left(\mathbf{D}_v - \mathbf{Q}\right) + \mathbf{S} - q\mathbf{I}$ which renders (12) to a skew-symmetric form. Partial integration of the convective term in (4) yields a weighted constitutive equation:

$$\sum_{\Omega^e \in \Omega} \left\{ \left(\mathbf{S} \, , \, \lambda \{ \frac{\partial \boldsymbol{\tau}}{\partial t} - \mathbf{L} \cdot \boldsymbol{\tau} - \boldsymbol{\tau} \cdot \mathbf{L}^T \} + \boldsymbol{\tau} - 2\eta\mathbf{D} \right)^e - \right.$$
$$\left. \left(\boldsymbol{u} \cdot \nabla \mathbf{S} \, , \, \boldsymbol{\tau} \right)^e + \int_{\Gamma^e} \mathbf{S} : \lambda \mathcal{L}^{\pm} \, d\Gamma \right\}, \tag{13}$$

with $\mathcal{L}^{\pm}$ a consistent approximation of the flux across the element boundaries [3].

A choice remains to be made about the order of the interpolation polynomials of the different variables with respect to each other. As is known from solving Stokes flow problems, velocity and pressure interpolation cannot be chosen independently and has to satisfy the Babuska-Brezzi condition. Likewise, interpolation of velocity and extra stress has to satisfy a similar compatibility condition. We report calculations using low order finite elements. Hence, using linear interpolation functions for viscoelastic stress, pressure and $\bar{\mathbf{D}}$ (or $\bar{\mathbf{G}}$) and quadratic interpolation for velocity unknowns.

4 Results

A comparison is presented between numerical and experimental results for a steady 3D complex flow through a cross-slot device (figure 2) using the DEVSS/DG method. For the cross-slot flow, due to its non-homogeneous nature (material near the center will experience a much higher strain rate than near the in- or outlet), the behavior of the constitutive models can be evaluated for complex flows. The aspect-ratio of both main axes of the rectangular cross section of the channel has been chosen close to unity and thus,

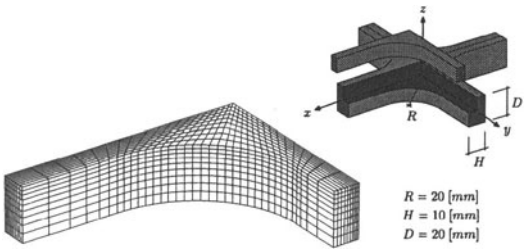


Fig. 2. FE mesh of cross-slot device ($D/H = 2$), inflow along y-axis, outflow along x-axis, #elements=2835, #nodes=25631, #DOF(u,p)=71988, #DOF(τ) = 4 × #elements × 48 = 544320.

a fully 3D flow field is obtained. A polymeric solution has been characterized with the Giesekus model and the Feta-PTT model using four discrete modes. Results of 3D viscoelastic calculations are shown in figure 3 together with point-wise Flow Induced Birefringence (FIB) measurements. Calculated data can be compared with the measured optical signal (M) following an integration of the first normal stress difference $N_1 = \tau_{xx} - \tau_{yy}$ along the depth of the flow:

$$M = -k_0 C \int N_1 \, dz \,, \tag{14}$$

with k_0 the initial propagation number and C the stress optical coefficient.

Stability results for the planar Couette flow are obtained by evaluating the L_2-norm of the disturbance resulting from equation 6. Figure 4 shows results obtained with the fully discontinuous method and the DEVSS/DG method which obviously prove to be unstable for this problem at a certain limiting Weissenberg number. Also results for the DEVSS-G/SUPG methods are shown in figure 5. It can be seen from these graphs that the fully continuous method proves to be stable even for high Weissenberg numbers. However, another factor that seems to influence the behavior of the DEVSS-

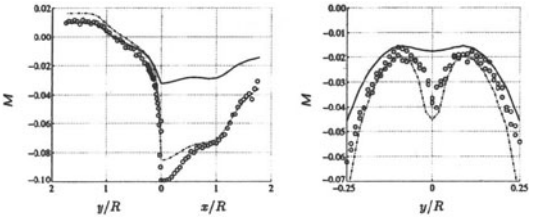


Fig. 3. Calculated and measured optical signal (M) along positive y-axis towards the stagnation point and along positive x-axis (*left*) and at outflow cross section $x/R = 1.5$ (*right*), —— Giesekus, – · – *Feta*-PTT.

Fig. 4. Temporal stability results for the fully discontinuous method (*left*) and the DEVSS/DG method (*right*).

G/SUPG method is a proper choice of the parameter β in (7). Sun *et al.* [11] proposed an adaptively scaled β by some norm of the viscoelastic stresses over the viscous stresses. From figure 5 it can be seen that one has to be very careful when finite element equations are constructed with such a choice for β because of the apparent loss of temporal stability of the scheme.

5 Discussion

From the previous section it follows that the DEVSS/DG method is highly efficient for the simulation of realistic 3D viscoelastic flows. Careful analysis of the DG-methods show that they are only stable up to a limiting value of the Weissenberg number in the plane Couette flow benchmark. However, given the smoothness of the plane Couette flow, the performance of these methods for non-smooth benchmark problems (e.g. the 4-to-1 contraction or the stick-slip problem) is still unresolved.

In order to rule out instabilities that are a result of the temporal discretization, time integration is performed using an Euler implicit method. Hence, with the exception of the semi implicit/explicit DEVSS/DG method of Baaijens *et al.* [1], an unconditionally stable time marching scheme should be obtained. However, numerical experiments with a fully implicit time integration variant of this method yielded identical stability results.

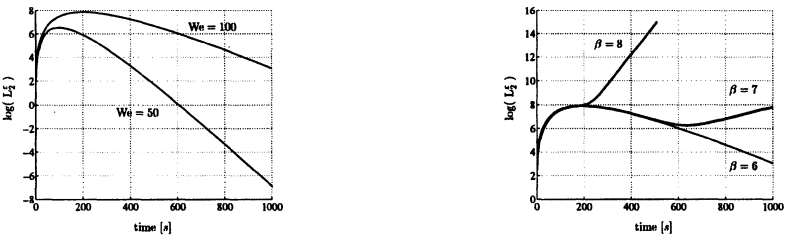


Fig. 5. Temporal stability results for the DEVSS-G/SUPG method, $\beta = 1$ (*left*) and $\beta > 1$, We $= 100$ (*right*).

Bibliography

[1] Baaijens, F.P.T., Selen, S.H.A., Baaijens, H.P.W., Peters, G.W.M., & Meijer, H.E.H. 1997. Viscolelastic flow past a confined cylinder of a low density polyethylene melt. *Journal of non-newtonian fluid mechanics*, **68**, 173–203.

[2] Bogaerds, A.C.B., Verbeeten, W.M.H., Peters, G.W.M., & Baaijens, F.P.T. 1999. 3D viscoelastic analysis of a polymer solution in a complex flow. *Computer methods in applied mechanics and engineering.* Accepted for publication.

[3] Cockburn, B. 1999. Discontinuous galerkin methods for convection-dominated problems. *Pages 69–224 of:* Barth, T.J., & Deconinck, H. (eds), *High-order methods for computaional physics.* Lecture Notes in Computational Science and Engineering, no. 9. Springer-Verlag.

[4] Fortin, M., & Fortin, A. 1989. A new approach for the FEM simulation of viscoelastic flows. *Journal of non-newtonian fluid mechanics*, **32**, 295–310.

[5] Gorodtsov, V.A., & Leonov, A.I. 1967. On a linear instability of a plane parallel couette flow of viscoelastic fluid. *Journal of applied mathematics and mechanics*, **31**, 310–319.

[6] Guénette, R., & Fortin, M. 1995. A new mixed finite element method for computing viscoelastic flows. *Journal of non-newtonian fluid mechanics*, **60**, 27–52.

[7] Lesaint, P., & Raviart, P.A. 1974. *On a finite element method for solving the neutron transport equation.* Academic Press.

[8] Oden, J.T., Babuska, I., & Baumann, C.E. 1998. A discontinuous hp finite element method for diffusion problems. *Journal of computational physics*, **146**, 491–519.

[9] Peters, G.W.M., Schoonen, J.F.M., Baaijens, F.P.T., & Meijer, H.E.H. 1999. On the performance of enhanced constitutive models for polymer melts in a cross-slot flow. *Journal of non-newtonian fluid mechanics*, **82**, 387–427.

[10] Schoonen, J.F.M., Swartjes, F.H.M., Peters, G.W.M., Baaijens, F.P.T., & Meijer, H.E.H. 1998. A 3D numerical/experimental study on a stagnation flow of a polyisobuthylene solution. *Journal of non-newtonian fluid mechanics*, **79**, 529–562.

[11] Sun, J., Phan-Thien, N., & Tanner, R.I. 1996. An adaptive viscoelastic stress splitting scheme and its applications: AVSS/SI and AVSS/SUPG. *Journal of non-newtonian fluid mechanics*, **65**, 75–91.

[12] Szady, M.J., Salomon, T.R., Liu, A.W., Bornside, D.E., Armstrong, R.C., & Brown, R.A. 1995. A new mixed finite element method for viscoelastic flows governed by differential constitutive equations. *Journal of non-newtonian fluid mechanics*, **59**, 215–243.

High Order Current Basis Functions for Electromagnetic Scattering of Curved Surfaces

Wei Cai[1]

Department of Mathematics
University of North Carolina at Charlotte
Charlotte, NC 28223

Abstract. We construct high order current vector basis functions on an arbitrary curved surface. The objective is to construct vector basis functions which consist of high order polynomials of the surface parameterization variables on curved triangles and have continuous normal components. Explicit formulation of high order current basis functions is provided.

1 Introduction

Integral equation formulation of electromagnetic scattering of conductive surfaces is a very popular approach for many applications including the parametric extraction for IC interconnects and computer packaging simulations [1], and antenna calculations. The main advantage of the integral formulation is its flexibility in handling very complex geometry of the scatter surface and the automatic enforcement of Sommerfeld exterior decaying conditions by the construction of proper Green's functions.

To represent the current vector field over conductor's surfaces, in many cases it is important to have a vector basis with continuity in its normal component across interfaces among adjacent elements. The RWG basis function is the most used first order basis function for engineering applications [2]. In this paper, we will extend such a basis to higher order with the required continuity across element interfaces. We consider an approach which could be applied to more general surfaces which can be subdivided as a union of curved triangles. Higher order current basis functions have been attempted in [3], but no systematic ways are presented to derive the basis functions so higher accuracy could be insured.

2 Current Basis Functions

In applying a Galerkin procedure to form MoM matrix [4] for the integral equation formulation of the electromagnetic scattering [5], normal continuity of the test current basis function $\vec{\mathbf{J}}$ is needed across the common interface of

triangular patches. In this section, we will present such current basis functions, the normal continuity of the current basis functions is the key property of the popular RWG basis functions, which issues no accumulation of charges across the element interfaces. In the following, we will give the formulation of higher order extension of the first order RWG basis over arbitrary curved patches. The details of the derivation of those basis functions can be found in [6] [7].

Let S be a curved triangle surface in 3-D and S is parameterized by $\vec{x}=\vec{x}\ (u_1,u_2),(u_1,u_2) \in T_0$. Here T_0 is a standard reference triangle in Figure 1.

Tangential vectors: $\partial_i\ \vec{x}\ \ i=1,2$ are defined as

$$\partial_i\ \vec{x}=\frac{\partial\ \vec{x}}{\partial u_i} \quad i=1,2. \tag{1}$$

Metric Tensor: The distance between two points on S parameterized by (u_1,u_2) and (u_1+du_1,u_2+du_2) is given by

$$(ds)^2=g_{\mu\nu}\,(u)\,du_\mu du_\nu \tag{2}$$

where repeated indices imply summation

$$g_{\mu\nu}=\frac{\partial\ \vec{x}}{\partial u_\mu}\cdot\frac{\partial\ \vec{x}}{\partial u_\nu} \quad 1\le\mu,\nu\le 2 \tag{3}$$

and $\{g_{\mu\nu}\}$ is defined as the covariant tensor [8]. The determinant of $\{g_{\mu\nu}\}$ is denoted by

$$g=\det\{g_{\mu\nu}\}=g_{11}g_{22}-g_{12}^2=\left\|\partial_1\ \vec{x}\ \times\partial_2\ \vec{x}\right\|^2. \tag{4}$$

Surface Element: The oriented differential surface element is given by

$$d\ \vec{S}=\partial_1\ \vec{x}\ \times\partial_2\ \vec{x}\ du_1du_2 \tag{5}$$

$$\left|d\ \vec{S}\right|=\left\|\partial_1\ \vec{x}\ \times\partial_2\ \vec{x}\right\|du_1du_2=\sqrt{g}du_1du_2 \tag{6}$$

where

$$\sqrt{g}=\left\|\partial_1\ \vec{x}\ \times\partial_2\ \vec{x}\right\|=\sqrt{g_{11}g_{22}-g_{12}^2}. \tag{7}$$

2.1 Hierarchical Polynomial Basis over Triangle T_0

Let T_0 be the unit reference triangle with vertices $a=(1,0),b=(0,0),c=(0,1)$ in Figure 1, we group polynomials into three modes: vertex modes, edge modes and internal modes [9].

– Vertex modes:

$$\begin{aligned} g_a(u_1,u_2) &= u_1 \\ g_b(u_1,u_2) &= 1-u_1-u_2 \\ g_c(u_1,u_2) &= u_2. \end{aligned} \tag{8}$$

Each vertex mode will take value 1 at one vertex and zero at other two vertices.

– Edge modes: for $2 \le l \le M$

$$\begin{aligned} g_l^{ab}(u_1,u_2) &= g_a g_b p_{l-2}(g_b-g_a) \\ g_l^{bc}(u_1,u_2) &= g_b g_c p_{l-2}(g_c-g_b) \\ g_l^{ca}(u_1,u_2) &= g_c g_a p_{l-2}(g_a-g_c). \end{aligned} \tag{9}$$

where $p_l(\xi), \xi \in [-1,1]$ is l-th order Lengendre polynomial.

Each of the edge mode is only nonzero along one edge of the triangle T_0.

– Internal Modes: $0 \le k+l \le M-3$

$$g_{l,k}^{int}(u_1,u_2) = g_a g_b g_c p_k(2g_c-1)p_l(g_b-g_a). \tag{10}$$

And each of the internal mode will vanish over all edges of T_0.

2.2 Triangular and Triangular Patches Matching

Consider two curved triangular patches T^+ and T^- with a common interface AC with length ℓ in Figure 1 . Let T^+ and T^- be parameterized, respectively, by

$$\boldsymbol{x} = \boldsymbol{x}^+(u_1,u_2) : T_0 \to T^+ \tag{11}$$

$$\boldsymbol{x} = \boldsymbol{x}^-(u_1,u_2) : T_0 \to T^-. \tag{12}$$

We assume that the interface AC in both T^+ and T^- is parameterized by $u_1+u_2=1$ and is labeled as side e_2^+ in T^+ and side e_2^- in T^-.

The high order basis function with continuous normal components can be written as [6]

$$\vec{f}(\boldsymbol{x}) = \begin{cases} \frac{l}{\sqrt{g^+}}(P_1^+(u_1,u_2)\,\partial_1\vec{x} + P_2^+(u_1,u_2)\,\partial_2\vec{x}) \\ \qquad \text{if} \quad \boldsymbol{x} = \boldsymbol{x}^+(u_1,u_2) \in T^+ \\ \frac{l}{\sqrt{g^-}}(P_1^-(u_1,u_2)\,\partial_1\vec{x} + P_2^-(u_1,u_2)\,\partial_2\vec{x}) \\ \qquad \text{if} \quad \boldsymbol{x} = \boldsymbol{x}^-(u_1,u_2) \in T^- \end{cases}$$

where

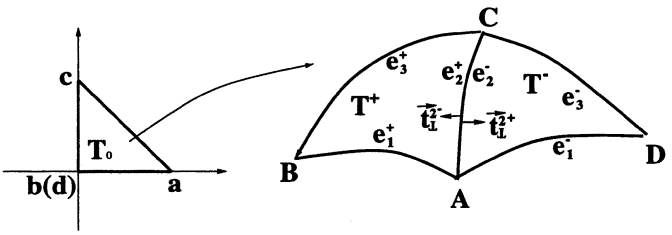


Fig. 1. (Left) Reference triangle T_0, (Right) Triangle-Triangle Patch

$$P_1^+(u_1,u_2) = I_n^a g_A(u_1,u_2) + \sum_{m=2}^{M} \frac{I_n^{(m)} - \widetilde{I}_t^{(m)}}{2} g_m^{e_2^+}(u_1,u_2) + \sum_{(l,m)\in L_\Delta} c_{lm}^1 g_{lm}^{int}$$

$$P_2^+(u_1,u_2) = I_n^c g_C(u_1,u_2) + \sum_{m=2}^{M} \frac{I_n^{(m)} + \widetilde{I}_t^{(m)}}{2} g_m^{e_2^+}(u_1,u_2) + \sum_{(l,m)\in L_\Delta} c_{lm}^2 g_{lm}^{int} \tag{13}$$

and coefficients P_1^-, P_2^- are given as

$$P_1^-(u_1,u_2) = -I_n^a g_A(u_1,u_2) + \sum_{m=2}^{M} \frac{-I_n^{(m)} - \widetilde{I}_t^{(m)}}{2} g_m^{e_2^-}(u_1,u_2) + \sum_{(l,m)\in L_\Delta} d_{lm}^1 g_{lm}^{int}$$

$$P_2^-(u_1,u_2) = -I_n^c g_C(u_1,u_2) + \sum_{m=2}^{M} \frac{-I_n^{(m)} + \widetilde{I}_t^{(m)}}{2} g_m^{e_2^-}(u_1,u_2) + \sum_{(l,m)\in L_\Delta} d_{lm}^2 g_{lm}^{int}. \tag{14}$$

with

$$\mathcal{L}_\Delta = \{(l,m),\ 0 \le l+m \le M-3\}. \tag{15}$$

Unknowns for each edge AC are

$$I_n^a, I_n^c, I_n^{(m)}, \widetilde{I}_t^{(m)}, \widehat{I}_t^{(m)}, \quad 2 \le m \le M \tag{16}$$

and interior unknowns for each triangular patch are

$$c_{lm}^1, c_{lm}^2 \quad (l,m) \in L_\Delta. \tag{17}$$

– RWG Basis

If we assume that the normal component of the current basis function remains constant, we have

$$\vec{f}\ (\boldsymbol{x}) = I_n \begin{cases} \frac{l}{\sqrt{g^+}}(g_A(u_1,u_2)\partial_1\ \vec{x}\ + g_C(u_1,u_2)\partial_2\ \vec{x}\ ) \\ \qquad \text{if} \qquad \boldsymbol{x} = \boldsymbol{x}^+(u_1,u_2) \in T^+ \\ \frac{l}{\sqrt{g^-}}(-\ g_A(u_1,u_2)\partial_1\ \vec{x}\ - g_C(u_1,u_2)\partial_2\ \vec{x}\ ) \\ \qquad \text{if} \qquad \boldsymbol{x} = \boldsymbol{x}^-(u_1,u_2) \in T^- \end{cases} . \tag{18}$$

and for flat triangle patches, we have in T^+

$$\boldsymbol{x} = \boldsymbol{x}^+(u_1,u_2) = g_A(u_1,u_2)\boldsymbol{x}_A + g_B(u_1,u_2)\boldsymbol{x}_B + g_C(u_1,u_2)\boldsymbol{x}_C \tag{19}$$

$$\begin{aligned} \partial_1\ \vec{x} &= \boldsymbol{x}_A - \boldsymbol{x}_B \\ \partial_2\ \vec{x} &= \boldsymbol{x}_C - \boldsymbol{x}_B, \end{aligned} \tag{20}$$

and in T^-

$$\boldsymbol{x} = \boldsymbol{x}^-(u_1,u_2) = g_A(u_1,u_2)\boldsymbol{x}_A + g_D(u_1,u_2)\boldsymbol{x}_D + g_C(u_1,u_2)\boldsymbol{x}_C \tag{21}$$

where $g_D(u_1,u_2) = g_B(u_1,u_2)$,

$$\begin{aligned} \partial_1\ \vec{x} &= \boldsymbol{x}_A - \boldsymbol{x}_D \\ \partial_2\ \vec{x} &= \boldsymbol{x}_C - \boldsymbol{x}_D. \end{aligned} \tag{22}$$

Thus, we have the RWG basis function [2]

$$\vec{f}\ (\boldsymbol{x}) = I_n \begin{cases} \frac{l}{2A^+}(\ \boldsymbol{x} - \boldsymbol{x}_B) & \text{if} \quad \boldsymbol{x} = \boldsymbol{x}^+(u_1,u_2) \in T^+ \\ -\frac{l}{2A^-}(\ \boldsymbol{x} - \boldsymbol{x}_D) & \text{if} \quad \boldsymbol{x} = \boldsymbol{x}^-(u_1,u_2) \in T^- \end{cases} . \tag{23}$$

where A^+ and A^- are the areas of triangles $T^+$and T^-, respectively.

The unknown for each edge AC is just I_n.

3 Conclusion

In the paper, we have presented the construction of higher order polynomial basis for current vector field on arbitrary curved surfaces. The vector basis functions are constructed for curved surfaces made of triangular patches and the current flow continuously along the normal directions of common

interfaces between triangle/triangle patches. For current basis functions of triangle/quadrilateral patches and applications, please refer to [6][7].

ACKNOWLEDGMENT

This work is supported by a DARPA grant through AFOSR grant number F49620-96-1-0341.

References

1. K. Nabors and J. White, Fastcap: A multipole accelerated 3-D capacitance extraction program, *IEEE Trans. Computer-Aided Design of Integrated Circuits Syst.*, **vol. 10**, pp. 1447-1459, Nov. 1991.
2. S. M. Rao, D. R. Wilton, and A. W. Glisson, Electromagnetic scattering by surfaces of arbitrary shape, *IEEE Trans. on Antenna and Propagation*, **Vol. AP-30**(3), pp. 409-418, May 1982.
3. Stephan Wandzura, Electric current basis functions for curved surfaces, *Electromagnetics* **12**:77-91, 1992.
4. R. F. Harrington, *Field Computation by Moment Methods.* New York: Macmillian, 1968.
5. W.C. Chew, *Waves and Fields in Inhomogeneous Media*, second edition, New York: IEEE Press, 1995.
6. Wei Cai, "High Order Current Basis functions for Electromagnetic Scattering of Curved Surfaces," to appear in *Journal of Scientific Computation*, 1999.
7. High order mixed RWG basis functions for electromagnetic scattering and applications, submitted to *IEEE Trans. on Microwave Theory and Techniques*, 1999.
8. Erwin Kreyszig, Differential Geometry, Dover Publication, Inc., New York, 1991.
9. Barna Szabo and Ivo Babuska, Finite Element Analysis, John Wiley and Sons, Inc., New York, 1991.

An Adaptive Discontinuous Galerkin Model for Coupled Viscoplastic Crack Growth and Chemical Transport

F. L. Carranza[1] * and R. B. Haber[2]

[1] Hibbit, Karlsson and Sorenson, Inc., 1080 Main Street, Pawtucket, RI 02860
[2] Department of Theoretical & Applied Mechanics, University of Illinois at Urbana-Champaign, Urbana IL 61801, USA

Abstract. This paper presents an adaptive finite element model for oxidation-driven fracture that uses space-time elements to track continuous crack-tip motion. The model incorporates viscoplastic material behavior, stress-enhanced diffusive transport of reactive chemical species and a cohesive interface fracture criterion. We discuss the weak formulation of the coupled system, including stabilized discontinuous Galerkin formulations for the chemical diffusion and the material evolution equations.

1 Introduction

Environmental embrittlement associated with the presence of oxygen controls crack growth in high-temperature applications of nickel-base superalloys in gas turbine engines. Experimental observations indicate that brittle, intergranular decohesion governs creep crack growth in polycrystalline specimens exposed to oxidizing environments at typical operating temperatures. The specific chemical mechanisms associated with environmental embrittlement may include long-range surface diffusion at the grain boundary, segregation to the grain boundary of various alloying species and impurities found in the bulk, as well as a variety of chemical reactions involving environmental agents introduced through the crack tip.

There is coupling between the chemical and mechanical response because the local stress state can enhance intergranular diffusion. The situation is further complicated because superalloys exhibit significant viscoplastic response at moderate to high temperatures. A local zone of strong inelastic deformation can develop to shield the crack tip from the far-field loading. Thus, a computational model for this form of fracture must address coupled chemical-viscoplastic response and the geometric evolution associated with crack propagation.

This paper reviews an adaptive finite element model in which a space-time grid tracks continuous crack-tip motion [2][3]. This significantly improves the temporal coherence of the discrete solution, and direct steady-state solutions

* Supported in part by NASA Grant No. NGT-70374.

in the frame of the moving crack tip can be obtained. The evolution equations for the viscoplastic state variables present a system of nonlinear first-order hyperbolic equations in the moving frame. These are approximated by a stabilized space-time-discontinuous Galerkin method. The advection-diffusion equations describing chemical transport are approximated by a stabilized time-discontinuous model.

2 Space-Time Representation of Crack Growth

Crack growth changes the topology of a body and fundamentally alters the geometry of the undeformed configuration. Thus, the formulation and solution of boundary value problems defined on variable domains is intrinsic to crack growth modeling. A space-time domain can be configured to track the evolving geometry of a growing crack for both transient and steady-state problems. Finite element discretization of the space-time domain ensures a continuous representation of crack advance. This careful treatment is critical to consistent resolution of the crack-tip fields, where convective effects help determine the qualitative features of the physical response.

Let Ω denote the space-time analysis domain for the mechanical problem with $\Gamma = \partial\Omega$, and let the vector $\mathbf{n}$ be the spatial component of the outward space-time normal to Γ. We assign a unit magnitude to $\mathbf{n}$, except where it vanishes. We also partition Γ into disjoint regions, Γ_u, Γ_t and Γ_{t^*}, with prescribed displacements, surface tractions and cohesive tractions, respectively. The inflow boundary of Ω is $\Gamma_- = \{\mathbf{x} \in \Gamma : \mathbf{n} \cdot \dot{\mathbf{a}} > 0\}$, where $\dot{\mathbf{a}}$ is the crack-tip velocity vector. The diffusion problem is defined on a second space-time domain $\hat{\Omega}$, that represents the uncleaved grain boundary ahead of an intergranular crack. The boundary $\hat{\Gamma} = \partial\hat{\Omega} = \overline{\hat{\Gamma}_c \cup \hat{\Gamma}_g}$, in which $\hat{\Gamma}_c$ and $\hat{\Gamma}_g$ are disjoint regions with prescribed solute concentration and flux. The vector $\hat{\mathbf{n}}$ is the spatial component of the outward space-time normal to $\hat{\Gamma}$.

3 Problem Formulation

3.1 Elastic-Viscoplastic Response

We neglect thermomechanical effects and employ a Norton–Soderberg power-law creep model without hardening.

$$\sigma = \mathsf{C} : (\varepsilon - \varepsilon_{\mathrm{p}}) \text{ on } \Omega \tag{1}$$

$$\varepsilon = \frac{1}{2}\left[\nabla\mathbf{u} + (\nabla\mathbf{u})^t\right] \text{ on } \Omega \tag{2}$$

$$\dot{\varepsilon}_{\mathrm{p}} = B\left(\frac{\sigma_{\mathrm{e}}}{\sigma_0}\right)^{m-1}\left(\frac{\sigma'}{\sigma_0}\right) \text{ on } \Omega \tag{3}$$

$$\varepsilon_{\mathrm{p}} = \bar{\varepsilon}_{\mathrm{p}} \text{ on } \Gamma_- \tag{4}$$

Here $\mathbf{u}$, σ, ε, ε_{p} and C are the displacement, the stress tensor, the total strain tensor, the inelastic strain tensor and the elasticity tensor, respectively. In the evolution equation (3), σ' is the deviatoric stress tensor, σ_{e} is the von Mises effective stress, and B, m and σ_0 are material parameters. A superposed tilde denotes a prescribed quantity. Equation 4 includes the initial condition for plastic strain, as well as inflow conditions for the moving control volume.

Under quasi-static assumptions, the equilibrium boundary value problem takes the form

$$\nabla \cdot \sigma + \mathbf{b} = \mathbf{0} \text{ on } \Omega \tag{5}$$

$$\mathbf{t} = \sigma \mathbf{n} \text{ on } \Gamma \tag{6}$$

$$\mathbf{u} = \tilde{\mathbf{u}} \text{ on } \Gamma_u \tag{7}$$

$$\mathbf{t} = \tilde{\mathbf{t}} \text{ on } \Gamma_t \tag{8}$$

$$\mathbf{t} = \mathbf{t}^* \text{ on } \Gamma_t^* \tag{9}$$

where $\mathbf{b}$ is a body force, $\mathbf{t}$ is a surface traction and $\mathbf{t}^*$ is a deformation-dependent traction determined by the cohesive fracture model. Equation 7 includes both initial and Dirichlet conditions for the displacement solution.

3.2 Cohesive Fracture Model

We use a cohesive interface model to provide a criterion for crack growth. Specifically, we adapt the work of Xu and Needleman [4] to the space-time moving domain model. We assume that the body contains a number of internal surfaces, called cohesive interfaces, along which fracture can potentially occur. The cohesive properties are determined by a characteristic length for the fracture process δ_n and a cohesive strength $\sigma_{\max}$. Let Δ_n be the normal component of the displacement jump across a cohesive interface, with the opening mode assigned a positive value by convention. Under pure mode-I conditions, the traction-separation law given in [4] simplify to

$$t_n^* = \sigma_{\max} e^{(1-\Delta_n/\delta_n)} \frac{\Delta_n}{\delta_n}. \tag{10}$$

The moving cohesive interface model introduces an explicit location, designated the *pseudo crack tip*, which separates the crack's free surface from the active cohesive surface. The usual traction-free boundary conditions hold on the free crack surface, while the traction-separation law applies on the cohesive surface ahead of the pseudo crack tip.

3.3 Stress-Assisted Diffusion

The mass transport of a solute species can be modeled as

$$\dot{c} + \nabla \cdot \mathbf{h} + r = 0 \text{ on } \hat{\Omega} \tag{11}$$

$$\mathbf{h} = -\mathbf{D}\nabla c + \mathbf{M}c\nabla\varepsilon_{\mathrm{D}} \text{ on } \hat{\Omega} \tag{12}$$

$$g = \mathbf{h} \cdot \hat{\mathbf{n}} \text{ on } \hat{\Gamma} \tag{13}$$

$$c = \tilde{c} \text{ on } \hat{\Gamma}_c \tag{14}$$

$$g = \tilde{g} \text{ on } \hat{\Gamma}_g \tag{15}$$

where c is the concentration of the solute, $\mathbf{h}$ is the mass flux vector, g is the surface flux and r is a source term which describes, for example, the rate at which the solute is generated due to chemical reactions. Equation 7 includes the initial conditions for solute concentration. The constitutive law for the mass flux (12) is an enhanced form of Fick's law, which has been modified to include the effects of dilatational strains.

Equation (11) becomes an advection-diffusion equation when the problem is formulated in a moving domain attached to a running crack tip. In this case, boundary layers appear in the concentration solution, and these play an important role in the physics of oxygen-embrittlement [2].

4 Finite Element Formulation

The coupled model described above has been implemented in an adaptive space–time finite element code. The details of the implementation are described in [3]; here we only summarize the code's key features. Our finite element method is based on a variational formulation that weakly enforces the evolution equation (3), the equilibrium equation (5) and the diffusion equation (11) on a space–time analysis domain. Independent space–time interpolations are introduced for the displacements $\boldsymbol{u}$, the inelastic strain $\boldsymbol{\varepsilon}_p$ and the concentration c. A Galerkin formulation approximates the equilibrium problem in which the displacements are modeled as continuous in both space and time. A stabilized discontinuous Galerkin (DG) formulation for the evolution equation (3) employs piecewise continuous interpolants for the inelastic strains that admit jumps across element boundaries. The diffusion problem is modeled by a second discontinuous Galerkin method that requires the concentrations to be continuous in space but admits jumps across temporal interfaces. The code includes residual-based error indicators and offers adaptive grid refinement to control the error in the solutions of the equilibrium and evolution problems. The DG approximation of the evolution ptoblem requires stabilization to treat the strong gradients in the crack-tip field.

5 Iterative Solution Strategy

A preconditioned CGS scheme solves the coupled chemical-mechanical problem for the displacements, the viscoplastic state variables and the chemical

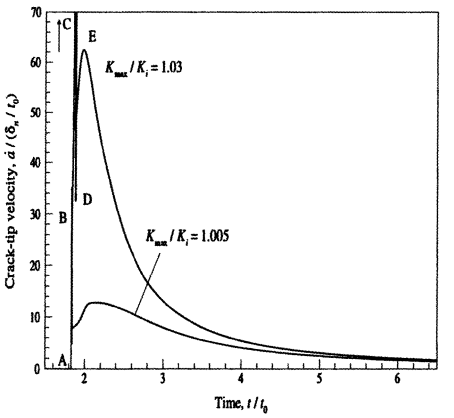


Fig. 1. History of crack-tip velocity for two levels of sustained overload.

concentration. The low-cost preconditioner incorporates a direct element-by-element solution of the uncoupled equations for the viscoplastic state variables. An h-adaptive analysis procedure uses a special a posteriori error indicator[1] to ensure an accurate solution of the hyperbolic subproblem.

6 Numerical Results

Figure 1 shows histories of the crack-tip velocity for a crack growing along a brittle grain boundary in a viscoplastic solid. Histories for two levels of the overload parameter $K_{\max}/K_i$ are shown, in which K_i is the stress intensity factor for initiation. Figure 2 shows the effective plastic strain field in the vicinity of the crack tip at $t/t_0 = 1.982$. Figure 3 shows the boundary layer ahead of a running crack tip in the steady advection-diffusion solution for oxygen concentration. We obtained solutions to the coupled boundary-value problem at two distinct crack velocities. The physically stable solution matches the velocity-dependent width of the boundary layer to the size of the active process zone in the cohesive model [2].

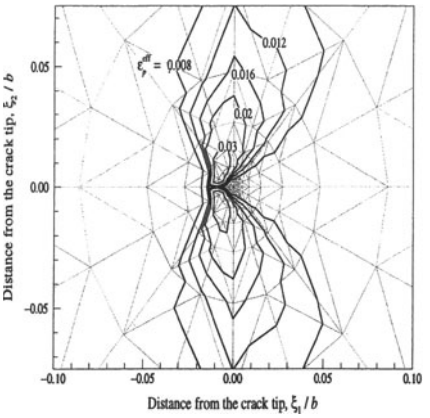


Fig. 2. Distribution of effective plastic strain in the vicinity of a running crack.

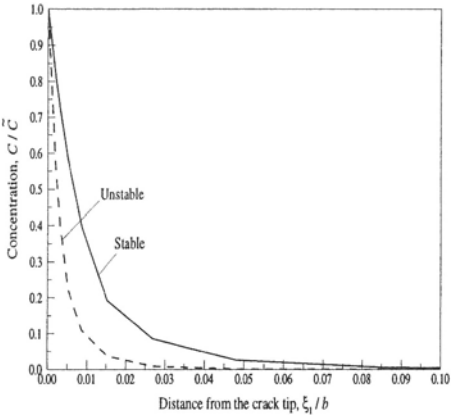


Fig. 3. Oxygen concentration profiles on the grain boundary ahead of a running crack tip.

References

1. Bey, K. S. and Oden, J. T.: *hp*-version discontinuous Galerkin methods for hyperbolic conservation laws, Comput. Methods Appl. Mech. Engng. 133 (1996) 259-286.
2. Carranza, F. L. and Haber, R. B.: A numerical study of intergranular fracture and oxygen embrittlement in an elastic-viscoplastic solid, J. Mech. Phys. Solids 47 (1999) 27-58.
3. Carranza, F. L., Fang, B. and Haber, R. B.: An adaptive space-time finite element model for oxidation-driven fracture, Comput. Methods Appl. Mech. Engng. 157 (1998) 399-423.
4. Xu, X.-P. and Needleman, A.: Numerical simulations of fast crack growth in brittle solids, J. Mech. Phys. Solids, 42, (1994) 1397-1434.

An Optimal Estimate for the Local Discontinuous Galerkin Method

Paul Castillo *

Scientific Computation,
University of Minnesota,
Minneapolis, MN 55455.

Abstract. L^2 error estimates for the Local Discontinuous Galerkin (LDG) method have been theoretically proven for linear convection diffusion problems and periodic boundary conditions. It has been proven that when polynomials of degree k are used, the LDG method has a suboptimal order of convergence k. However, numerical experiments show that under a suitable choice of the numerical flux, higher order of convergence can be achieved. In this paper, we consider Dirichlet boundary conditions and we show that the LDG method has an optimal order of convergence $k+1$.

1 Introduction

Over the last decade, the Runge Kutta Discontinuous Galerkin (RKDG) method introduced and analyzed by Cockburn and Shu, see [CS98,CSK], has become a practical method for solving hyperbolic systems. The RKDG method can provide high order of approximation and has a high degree of parallelism. Originally, it was designed for solving nonlinear hyperbolic systems. Recently, Cockburn and Shu, [CS98], have extended the RKDG method to time-dependent convection diffusion systems. The extension proposed in [CS98], the Local Discontinuous Galerkin (LDG) method, can also provide high order accuracy and, because of its local nature, is highly parallelizable. In [CS98], it was shown that, for linear convection diffusion problems and periodic boundary conditions, the method is at least of order k when polynomials of degree k are used. In this paper, we show that, under a suitable choice of the numerical flux, an optimal order $k+1$ can be achieved even for Dirichlet boundary conditions.

We now describe the general formulation of the LDG method, applied to the following Cauchy problem

$$\partial_t u + \partial_x \left(f(u) - a(u)\partial_x u \right) = 0 \quad u(0,x) = u_0. \tag{1}$$

* Partially supported by National Science Foundation grant DMS-9805617 and by the University of Minnesota Supercomputer Institute

First, the equation is rewritten in conservation form. We obtain the following system

$$\begin{aligned} \partial_t u + \partial_x \left(f(u) - \sqrt{a(u)} q \right) &= 0, \quad \text{in } (0,T) \times \Omega, \\ q - \partial_x g(u) &= 0, \quad \text{in } (0,T) \times \Omega, \\ u(0,x) &= u_0. \quad \text{on } \Omega. \end{aligned} \tag{2}$$

where $q = \sqrt{a(u)}\partial_x u$ and $g(u) = \int^u \sqrt{a(s)} ds$. Then, the system is discretized by adapting the ideas developed in the RKDG method. Let $\{x_{j+\frac{1}{2}}\}_{j=0}^N$ be a partition of the domain Ω, and $I_j = (x_{j-1/2},\ x_{j+1/2})$. At each time $t \in [0,\ T]$, the solution $w = (u, q)$ of the system (2) is approximated by a function $w_h = (u_h, q_h)$ of the finite element space $V_h = V_h^k \times V_h^k$ of piecewise discontinuous functions

$$V_h^k = \{v \in L^1(\Omega) | v_{|I_j} \in P^k(I_j)\}.$$

The restriction of w_h to an interval I_j satisfies, for any test function (v_u, v_q) in V_h, the following weak relations

$$\begin{aligned} \int_{I_j} \partial_t u_h v_u &= \int_{I_j} h_u(w_h) \partial_x v_u - \hat{h}_u(w_h) v_u \Big|_{x_{j-1/2}}^{x_{j+1/2}}, \\ \int_{I_j} q_h v_q &= \int_{I_j} h_q(w_h) \partial_x v_q - \hat{h}_q(w_h) v_q \Big|_{x_{j-1/2}}^{x_{j+1/2}} \end{aligned}$$

where $h(u,q) = (f(u) - \sqrt{a(u)}q, -g(u))$ and $\hat{h}$ is a numerical flux which incorporates the boundary conditions. In our error analysis, we consider the linear convection diffusion equation

$$\partial_t u + \partial_x (cu - a\partial_x u) = 0,\ u(0,x) = u_0,\ u(t,\alpha) = \delta_\alpha(t),\ u(t,\beta) = \delta_\beta(t) \tag{3}$$

where c is an arbitrary, non zero, constant, $a > 0$ and $\Omega = [\alpha,\ \beta]$. For this problem the flux h is given by

$$h(u,q) = \left(cu - \sqrt{a}q, -\sqrt{a}u\right).$$

We consider the following choice of the numerical flux at the interior grid points :

$$\hat{h}\left(w^-, w^+\right) = \begin{cases} \left(cu^- - \sqrt{a}q^+, -\sqrt{a}u^-\right), & \text{if } c > 0, \\ \left(cu^+ - \sqrt{a}q^-, -\sqrt{a}u^+\right), & \text{if } c < 0, \end{cases} \tag{4}$$

With this particular choice an optimal order of convergence can be obtained for the LDG method as described in the following section.

2 Error analysis for linear convection diffusion

In this section, we present an error estimate for the LDG method applied to the linear model problem (3). Our error estimate is based on the technique presented in [CS98]. However, in this paper, we not only obtain an optimal error estimate, but we also consider general Dirichlet boundary conditions.

Theorem 1. *(Error estimate). Let $e(t,x)$ be the error of the approximation obtained by using the LDG method to solve the linear model problem, the quantity defined by,*

$$|||e|||^2 = \| e_u(T) \|^2_{L^2(\Omega)} + \int_0^T \left\{ \| e_q \|^2_{L^2(\Omega)} + |c||e_u(\alpha^+)|^2 + \frac{|c|}{2}|e_u(\beta^-)|^2 \right\}$$

satisfies

$$|||e||| = O\left((\Delta x)^{k+1}\right),$$

provided $u_o \in H^{k+3}$.

Proof. Without loss of generality, we assume a positive convection coefficient, $c > 0$ in (3). The solution w_h obtained from the LDG method satisfies the weak formulation

$$\mathcal{B}_h(w_h, v) = \mathcal{L}_h(v), \quad \forall v \in V_h, \tag{5}$$

where the bilinear form $\mathcal{B}_h$ and the linear operator $\mathcal{L}_h$ are defined as follows

$$\begin{aligned}
\mathcal{B}_h(w,v) = &\int_0^T \int_\Omega \partial_t w_u v_u + \int_0^T \int_\Omega w_q v_q - \int_0^T \sum_1^N \int_{I_j} h(w)^T \partial_x v \\
&- \int_0^T \sum_1^{N-1} \hat{h}(w)^T [v]_{j+1/2} + \int_0^T \left(c w_u(\beta^-) - \sqrt{a} w_q(\beta^-)\right) v_u(\beta^-) \\
&+ \int_0^T \sqrt{a} w_q(\alpha^+) v_u(\alpha^+),
\end{aligned}$$

$$\mathcal{L}_h(v) = \int_0^T \left\{ c\delta_\alpha v_u(\alpha^+) - \sqrt{a}\delta_\alpha v_q(\alpha^+) + \sqrt{a}\delta_\beta v_q(\beta^-) \right\}.$$

Let w be the exact solution of (3). Since w and w_h satisfy (5), we obtain for the error $e = w - w_h$

$$\mathcal{B}_h(e, v) = 0, \quad \forall v \in V_h.$$

Let $\Pi = (\Pi_u, \Pi_q)$ be an arbitrary projection onto $V_h = V^k \times V^k$ and set $v = \Pi(e)$, $p = (p_u, p_q) = \Pi(w) - w$. ¿From the above error equation we get

$$\mathcal{B}_h(v, v) = \mathcal{B}_h(p, v).$$

An error estimate can now be obtained from the previous equation by choosing an appropriate projection Π and by providing an upper bound for the right-hand side. We proceed as follows: First, observe that since the coefficient c is positive, the terms involving the numerical fluxes can be simplified as follows

$$\hat{h}(p)^T[v]_{j+1/2} = \left(cp_u^-[v_u] - \sqrt{a}p_q^+[v_u] - \sqrt{a}p_u^-[v_q]\right)_{j+1/2}.$$

By simply selecting an appropriate projection Π, we can set the above quantity to zero. For the u component, we consider a polynomial that interpolates the function at the right endpoint. For the q component, we consider the left endpoint. We also have $p_u(\beta^-) = p_q(\alpha^+) = 0$. Hence, we obtain the following simplified expression for $\mathcal{B}_h(p,v)$:

$$\int_0^T \int_\Omega \partial_t p_u v_u + p_q v_q - \int_0^T \sum_1^N \int_{I_j} h(p)^T \partial_x v - \int_0^T \sqrt{a} p_q(\beta^-) v_u(\beta^-)$$

Note that in the pure hyperbolic case, $a = 0$, we recover the expression of $\mathcal{B}_h(p,v)$ obtained in [CS98]. An optimal order can be obtained by considering interpolation polynomials at the Gauss-Radau quadrature points. Following the ideas of LeSaint and Raviart [LR74], we define $\Pi_u(f)$ and $\Pi_q(f)$ as the interpolation polynomial on the Gauss-Radau quadrature points, with the appropriate fixed point. From the integration properties of these polynomials and the Bramble-Hilbert lemma, it can be shown that if $f \in H^{k+2}(I_j)$ and $v \in P^k(I_j)$ we have the following estimates

$$\int_{I_j} |(\Pi(f) - f)\, v| \leq C_1(\Delta x)^{k+1} |f|_{H^{k+1}(I_j)} \parallel v \parallel_{L^2(I_j)},$$

$$\int_{I_j} |(\Pi(f) - f)\, \partial_x v| \leq C_2(\Delta x)^{k+1} |f|_{H^{k+2}(I_j)} \parallel v \parallel_{L^2(I_j)}.$$

Applying again the Bramble-Hilbert lemma, we can derive an L^∞ estimate for the error at the boundary endpoint

$$\left|(\Pi(f) - f)\,(\beta^-)\right| \leq C(\Delta x)^{k+1} \parallel f \parallel_{W_\infty^{k+1}(I_j)}.$$

Using these estimates and the regularity of the initial conditions, we obtain the following estimate

$$|\mathcal{B}_h(p,v)| \leq \gamma_1(a,c) C_1(u_0)(\Delta x)^{k+1} \int_0^T \parallel \Pi_u(e) \parallel_{L^2(\Omega)} \tag{6a}$$

$$+ \gamma_2(a) C_2(u_0)(\Delta x)^{k+1} \int_0^T \parallel \Pi_q(e) \parallel_{L^2(\Omega)} \tag{6b}$$

$$+ \gamma_3(a) C_3(u_0)(\Delta x)^{k+1} \int_0^T |\Pi_u(e)(\beta^-)|, \tag{6c}$$

where the constants $\gamma_1(a,c), \gamma_2(a)$ and $\gamma_3(a)$ are such that, $\gamma_1(0,.)$, $\gamma_1(.,0)$, $\gamma_2(0)$ and $\gamma_3(0)$ do not vanish. Applying Young's inequality to the second, (6b), and third, (6c), term we get

$$\begin{aligned}|\mathcal{B}_h(p,v)| \leq \rho_1(a,c)C_1(u_0)(\Delta x)^{k+1}\int_0^T \| \Pi_u(e) \|_{L^2(\Omega)} \\ + \frac{1}{2}\int_0^T \| \Pi_q(e) \|^2_{L^2(\Omega)} + \rho_2(a)\int_0^T \{C_2(u_0)(\Delta x)^{k+1}\}^2 \\ + \frac{c}{4}\int_0^T |\Pi_u(e(\beta^-))|^2 + \rho_3(a,c)\int_0^T \{(\Delta x)^{k+1}C_3(u_0)\}^2. \quad (7)\end{aligned}$$

Since

$$\begin{aligned}\mathcal{B}_h(v,v) = \frac{1}{2}\| \Pi_u(e(T)) \|^2_{L^2(\Omega)} - \frac{1}{2}\| \Pi_u(e(0)) \|^2_{L^2(\Omega)} + \int_0^T \| \Pi_q(e) \|^2_{L^2(\Omega)} \\ + \frac{c}{2}\int_0^T \{|\Pi_u(e(\alpha^+))|^2 + |\Pi_u(e(\beta^-))|^2\} + \frac{c}{2}\int_0^T \sum_{j=1}^{N-1}[\Pi_u(e)]^2 \quad (8)\end{aligned}$$

combining (7) and (8), we obtain

$$\begin{aligned}|||\Pi(e)|||^2 \leq \| \Pi_u(e(0)) \|^2_{L^2(\Omega)} + C_1(\Delta x)^{2(k+1)} \\ + C_2(\Delta x)^{k+1}\int_0^T \| \Pi_u(e) \|_{L^2(\Omega)},\end{aligned}$$

where C_1 and C_2 depend solely on T and u_0. By application of Gronwall's lemma, we finally get

$$|||\Pi(e)||| = O\left((\Delta x)^{k+1}\right).$$

From this bound, the error estimate follows. □

3 Numerical experiments

We present here some numerical results, for the LDG method applied to the linear convection diffusion problem (3). The domain is the one dimensional interval $[-1, 1]$ and the initial conditions are $u_o(x) = \sin(x)$. We compute the solution on the time interval $[0, 0.1]$. A sufficiently small fixed time step has been used, such that the contribution to the error is dominated by the spatial error. We present the computed order of convergence for polynomials of degree 1 to 4. In Table 1, we show the L^2-error of the $u-$component at the end time and the computed order for a convection diffusion equation with a small diffusion coefficient $a = 0.01$. In Table 2, the diffusion coefficient a is increased to 1.0. We have also included the L^2 error of u_x at the end time, asserting our claim that an order of $k+1$ is also obtained for the $q-$component.

Table 1. $c = 1.0,\ a = 0.01$

k	$N = 10$	$N = 20$		$N = 40$	
	error	error	order	error	order
1	1.3593e-03	3.5642e-04	1.9313	8.9377e-05	1.9956
2	3.7061e-05	4.4837e-06	3.0471	5.5999e-07	3.0012
3	2.8634e-07	1.6651e-08	4.1040	1.0444e-09	3.9949
4	4.5401e-09	1.3242e-10	5.0996	4.1731e-12	4.9878

Table 2. $c = 1.0,\ a = 1.0$

k		$N = 10$	$N = 20$		$N = 40$	
		error	error	order	error	order
1	u	1.2161e-03	3.1400e-04	1.9534	7.9697e-05	1.9782
	u_x	1.9568e-03	5.0633e-04	1.9503	1.2871e-04	1.9759
2	u	3.9617e-05	4.6240e-06	3.0989	5.4412e-07	3.0872
	u_x	1.9065e-05	2.4675e-06	2.9498	3.1107e-07	2.9878
3	u	2.3411e-07	1.4748e-08	3.9886	9.3217e-10	3.9838
	u_x	3.7879e-07	2.3641e-08	4.0020	1.5160e-09	3.9630
4	u	4.8128e-09	1.3674e-10	5.1374	4.2538e-12	5.0065
	u_x	2.3370e-09	7.7297e-11	4.9181	5.1103e-12	3.9189

4 Conclusion

In this paper we have presented a new error estimate for the Local Discontinuous Galerkin method applied to linear convection diffusion problem with Dirichlet boundary conditions. We have, theoretically and numerically, demonstrated an optimal order of convergence of $k + 1$, when using a suitable numerical flux. Extension to other types of boundary conditions such as Neumann boundary conditions and to higher dimensional problems will be treated in a forthcoming paper.

References

[CS98] Cockburn, B., Shu, C.W.: The Local Discontinuous Galerkin Finite Element Method for convection diffusion systems. SIAM J. Numer. Anal. **35**, (1998) 2440–2463

[CSK] Cockburn, B., Shu, C.W., Karniadakis, G.E.: An overview of the development of DG methods. This volume, Springer Verlag, Berlin, Germany,

[LR74] LeSaint, P., Raviart, P.A.: On a finite element method for solving the neutron transport equation. In: Mathematical aspects of finite elements in partial differential equations. Academic Press, C.A de Boor (Ed), Academic Press, New York (1974), 89-145

Post-Processing of Galerkin Methods for Hyperbolic Problems

Bernardo Cockburn[1], Mitchell Luskin[1], Chi-Wang Shu[2], and Endre Süli[3]

[1] School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455, USA
[2] Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912, USA
[3] Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, U.K.

Abstract. It is well known that the discontinuous Galerkin (DG) method for scalar linear conservation laws has an order of convergence of $k + 1/2$ when polynomials of degree k are used and the exact solution is sufficiently smooth. In this paper, we show that a suitable post-processing of the DG approximate solution is of order $2k+1$ in $\mathrm{L}^2(\Omega_0)$ where Ω_0 is a domain on which the exact solution is smooth enough. The post-processing is a convolution with a kernel whose support has measure of order one if the meshes are arbitrary; if the meshes are translation invariant, the support of the kernel is a cube whose edges are of size of order Δx only. The post-processing has to be performed only once, at the final time level.

1 Introduction

In this paper, we consider finite element methods for linear scalar conservation laws and show how to exploit their *inherent oscillatory nature* to enhance the quality of their approximation. The enhancement is achieved by means of a simple post-processing of the approximate solution at the very end of the computation.

To illustrate this idea, let us consider the following problem:

$$\begin{aligned} &u_t + u_x = 0, && \text{in } (0,1) \times (0,T), \\ &u(x,0) = \sin(2\,\pi\,x) && \text{for } x \in (0,1), \end{aligned}$$

with periodic boundary condition, and let us compute an approximation u_h to its solution u by using the discontinuous Galerkin (DG) method with piecewise polynomials of degree one. We also consider the post-processed approximation $u_h^\star = K_h^{4,2} \star u_h$, where the convolution kernel $K_h^{4,2}$ is defined by

$$K_h^{4,2}(x) = -\frac{1}{12}\psi^{(2)}(x-1) + \frac{7}{6}\psi^{(2)}(x) - \frac{1}{12}\psi^{(2)}(x+1);$$

here $\psi^{(2)}$ is the B-spline obtained by convolving the characteristic function of $[-1/2, 1/2]$ with itself once. In Fig. 1, we display, for $T = 0.1$ and $h = 1/10$

and 1/20, the errors $x \mapsto u(T,x) - u_h(T,x)$ and $x \mapsto u(T,x) - u_h^\star(T,x)$. Notice the oscillatory nature of the error $x \mapsto u(T,x) - u_h(T,x)$ typical of finite element methods and the super-convergence at the Gauss-Radau points, a fact conjectured in 1994 by Biswas, Devine, and Flaherty [4] and proven recently by Adjerid, Flaherty and Krivodono [2]; see also the work done by Adjerid, Aiffa, and Flaherty in [1]. Notice also that the oscillations are totally absent from the error $x \mapsto u(T,x) - u_h^\star(T,x)$. This shows that convolving the approximate solution u_h with the kernel $K_h^{4,2}$ results in the *removal* of oscillations around the exact solution. Moreover, the result of such a filtering is a new approximation $u_h^\star$ that converges faster to u than u_h. Indeed, in Fig. 2, we display the function $x \mapsto \log(\,|\,u(T,x) - u_h^\star(T,x)\,|\,)$, for $h = 1/10, 1/20, 1/40$ and $1/80$. Notice also that each time h is halved, the maximum error is divided by a factor of eight, at least. This indicates that the post-processed approximation has the order of convergence of three, at least; the approximate solution u_h has an order of convergence of two only.

In Figs. 3 and 4, we repeat the experiment using polynomials of degree two. In Fig. 3, we see again the oscillatory nature of the approximation, the super-convergence at the three Gauss-Radau points, and how the oscillations are removed after the convolution. This time, the convolution kernel is defined by

$$K_h^{6,3}(x) = \frac{37}{1920}\psi^{(3)}(x-2) - \frac{97}{480}\psi^{(3)}(x-1) - \frac{437}{320}\psi^{(3)}(x) - \frac{97}{480}\psi^{(3)}(x+1) + \frac{37}{1920}\psi^{(3)}(x+2),$$

where $\psi^{(3)}$ is the B-spline obtained by convolving the characteristic function of the interval $[-1/2, 1/2]$ two times. In Fig. 4, we see that each time h is halved, the maximum error is divided by a factor clearly bigger than thirty two. This shows that the error in the post-processed approximation is of order at least five.

In connection with this fact, we must point out that in 1996 Lowrie [11] found analytical and numerical evidence that when polynomials of degree k are used, a 'component of the error' of the DG method converges with an order of $2k+1$; this is in striking contrast with the fact that the order of convergence of the DG approximation is $k+1/2$ (in the one-dimensional case and for special meshes in several space dimensions, the order is $k+1$). In this paper, we give this indication a firm mathematical basis. Moreover, we actually show how to compute such a super-convergent approximation $u_h^\star$ by a simple post-processing technique which is independent of the equation and of the numerical method.

This paper is organized as follows. In section 2, we present a brief overview of the development of the ideas which form the basis of this paper. In section 3, we state and discuss our main results and in section 4, we end with some concluding remarks. The results shown in this paper are a particular case

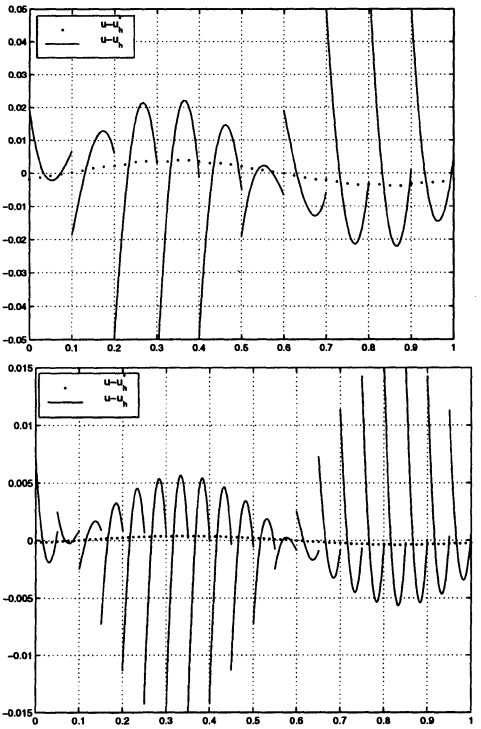


Fig. 1. The errors $u - u_h$ (solid line) and $u - u_h^\star$ (dots) for $h = 1/10$ (top) and $h = 1/20$ (bottom). The function u is the smooth exact solution, u_h is the approximation given by the DG method with polynomials of degree one, and $u_h^\star = K_h^{4,2} \star u_h$.

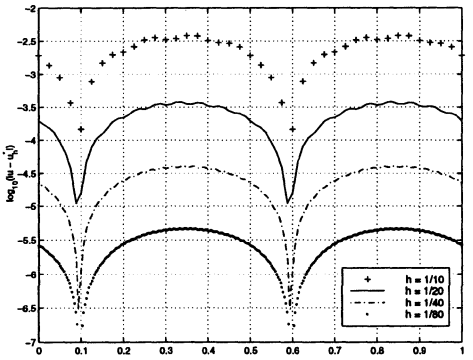


Fig. 2. The error $\log(|\, u - u_h^\star \,|)$ for $h = 1/10$ (top), $h = 1/20$, $h = 1/40$, and $h = 1/80$ (bottom). Notice that each time h is halved, the maximum error decreases by a factor not smaller than 8. The order of convergence is therefore not less than 3.

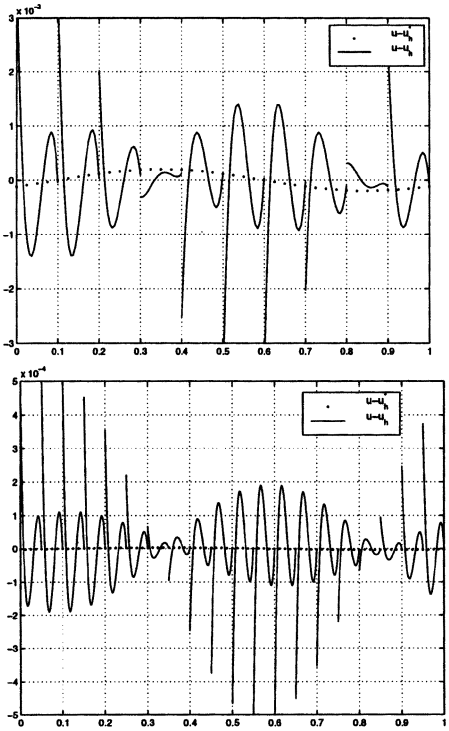


Fig. 3. The errors $u - u_h$ (solid line) and $u - u_h^\star$ (dots) for $h = 1/10$ (top) and $h = 1/20$ (bottom). The function u is the smooth exact solution, u_h is the approximation given by the DG method with polynomials of degree two, and $u_h^\star = K_h^{6,3} \star u_h$.

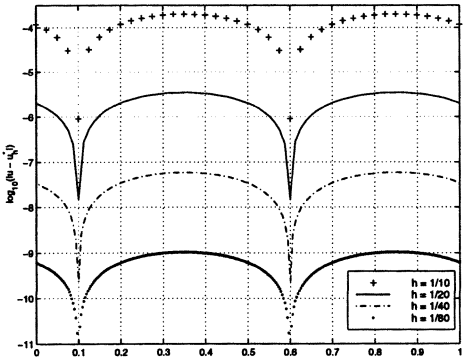


Fig. 4. The error $\log(|\, u - u_h^\star \,|)$ for $h = 1/10$ (top), $h = 1/20$, $h = 1/40$, and $h = 1/80$ (bottom). Notice that each time h is halved, the maximum error decreases by a factor not smaller than 32. The order of convergence is therefore not less than 5.

of those contained in [6] where general hyperbolic systems and several finite element methods are considered.

2 A brief overview

To introduce the ideas on which our work is based, we briefly review the development of post-processing techniques devised to improve the quality of the approximation. The reader interested in a complete material, should consult the book by Wahlbin [16] on super-convergence in Galerkin finite element methods.

Finite difference and spectral methods for hyperbolic problems. In 1977, Majda and Osher [13] considered formally high-order accurate dissipative difference schemes for solving hyperbolic problems. They considered a one-dimensional model problem of a two-by-two hyperbolic system whose characteristics are parallel to $x = \pm t$ with initial condition a step function whose discontinuity sits at the origin. They showed that on the region between the characteristics issuing from the origin $|x/t| < 1 - \delta^2$, the order of convergence is independent of the numerical scheme. They pointed out that in 1962 Fedorenko [8] and in 1969 Apelkrans [3] displayed numerical evidence that the order of convergence had to be one. However, they showed that by a simple pre-processing of the initial data, the rate of convergence was two. Moreover, they found that they could actually recover the full formal order of accuracy of the scheme on the region $|x/t| < 1 - \delta^2$ provided they pre-processed the initial data in a suitable way. This seems to have been the first time that the idea of pre-processing the initial data to enhance the order of convergence was used in the framework of hyperbolic equations. In 1986, Johnson and Pitkäranta [10] employed a similar idea when analyzing the DG for linear hyperbolic problems.

In 1978, Mock and Lax [14] showed that for a difference scheme of formal order of accuracy μ for linear hyperbolic systems, the *moments* of the exact solution converge with order μ provided, again, that the initial data is suitably pre-processed. This results holds even if the exact solution contains discontinuities. They also showed how to post-process the approximate solution by a simple convolution to enhance its accuracy on the regions of smoothness of the exact solution: If the solution is locally smooth enough, they could obtain almost the full order of convergence μ provided the support of the post-processing kernel was of order almost one. This seems to have been the first instance that the ingredients of (i) pre-processing the initial data, (ii) getting error estimates for the moments, and (iii) post-processing the approximation, appear clearly delineated.

Later, in 1985, Gottlieb and Tadmor [9], motivated by the work of Mock and Lax [14], found a spectrally accurate post-processing kernel for spectral methods; see also the 1978 paper by Majda, McDonough and Osher [12].

Again, the full spectral accuracy could be recovered by using a convolution; the measure of the support of the kernel had to be of order almost one.

Finite element methods for elliptic problems. Quite independently of the developments described above, in 1977, Bramble and Schatz [5] considered linear elliptic problems and demonstrated how to post-process the finite element solution by means of a simple convolution to enhance the quality of the approximation. They showed that the order of convergence could be doubled if the exact solution was locally smooth. It is important to point out that, just like Mock and Lax, Bramble and Schatz proved (i) a negative-order norm error estimate (an error estimate for the moments in Mock and Lax's terminology) and then showed (ii) how to use it to enhance the approximation by a convolution. However, unlike Mock and Lax's convolution kernel, their kernel is contained in a cube of size of order h only for locally translation invariant meshes; a fact of tremendous computational advantage.

Also in 1977, Thomée [15] extended Bramble and Schatz's results [5] to include super-convergence results of the derivatives and gave an elegant proof of their approximation results by using Fourier analysis.

An application of the Bramble and Schatz technique to the simulation of miscible displacement was devised and analyzed by Douglas [7]. See other applications in the book of Wahlbin [16].

The main ideas. In this paper, we apply the ideas of Bramble and Schatz [5] to hyperbolic problems which, as we saw, are closely related to the ideas of Mock and Lax [14].

Since the negative-order norms capture the oscillatory nature of a function, we gather approximation results that tell us how to use negative-order norm estimates for the error between an arbitrary function u and an arbitrary approximation u_h to bound the L^2_{loc}-error between u and a post-processing of u_h. These results do not depend on the partial differential equation or on the numerical scheme. Next, we obtain negative-order norm estimates for the error between the exact solution of a hyperbolic problem and its finite element approximation u_h. Finally, we simply combine the above results to obtain the desired estimates.

In what follows, we carry out this program for the DG method applied to the scalar linear conservation law.

3 The results

Approximation results. In this section, we collect and discuss two results that relate negative-order estimates of the difference between u and an arbitrary approximation u_h with L^2-error estimates of the difference between u and a post-processed u_h.

In what follows, we denote the d-dimensional unit cube $[-1,1]^d$ by I. We denote by $\| u \|_{0,\Omega}$ the standard L^2-norm in Ω of u. For any natural number ℓ, we set

$$\| u \|_{\ell,\Omega} = \left\{ \sum_{|\alpha| \le \ell} \| D^\alpha u \|_{0,\Omega}^2 \right\}^{1/2}, \quad \| u \|_{-\ell,\Omega} = \sup_{\phi \in C_0^\infty(\Omega)} \int_\Omega u\,\phi\,dx / \| \phi \|_{\ell,\Omega}.$$

Note that for $\Omega = [-1,1]$ and $u_N(x) = \sin(2\pi N x)$, a simple computation gives $\| u_N \|_{-\ell,\Omega} = 1/(2\pi N)^\ell$. That is, a negative-order norm can sense that u_N hits 0 in a very regular pattern. This is why negative-order norms are used to detect the oscillatory nature of a function.

The first result is a standard approximation result.

Theorem 1. *Let ν and ℓ be two given natural numbers. Let K^ν be a function in $H^\nu(R^d)$ with support contained in I, such that, for all polynomials v of degree $\nu - 1$,*

$$K^\nu \star v = v.$$

Set $K_\epsilon^\nu(x) = K^\nu(x/\epsilon)/\epsilon^d$. Now, let u_h be a function in $L^2(\Omega_1)$ and let u be a function in $H^\nu(\Omega_1)$. Then, for any set Ω_0 such that $\Omega_0 + \epsilon_{opt} I \subset \Omega_1$, there is a constant $C_{d,\nu}$ that depends solely on d and ν such that

$$\| u - K_{\epsilon_{opt}}^\nu \star u_h \|_{0,\Omega_0} \le C_{d,\nu} \| u \|_{\nu,\Omega_1}^\theta \| u - u_h \|_{-\ell,\Omega_1}^{1-\theta}, \tag{1}$$

where

$$\theta = \frac{d + 2\ell}{2\nu + d + 2\ell},$$

and

$$\epsilon_{opt}^{d/2+\ell} = \left(\frac{d+2\ell}{2\nu} \right)^\theta \| u \|_{\nu,\Omega_1}^\theta \| u - u_h \|_{-\ell,\Omega_1}^{-\theta}.$$

From the above result, it is clear that we want to make θ as close to zero as possible since $\| u - K_{\epsilon_{opt}}^\nu \star u_h \|_{0,\Omega_0}$ would then have almost the *same* rate of convergence as $\| u - u_h \|_{-\ell,\Omega_1}$. However, to achieve this, the value ϵ_{opt} would have to be a quantity of order almost one. This means that the support of the convolution kernel $K_{\epsilon_{opt}}^\nu$ contains a cube whose edge size is of order almost one.

It is possible, however, to overcome this difficulty and obtain a convolution kernel whose support is contained in a cube with edges of the order of the diameters of the elements of the mesh. To state the theorem containing this result, we need to introduce some notation. For each multi-index $\alpha = (\alpha_1, \cdots, \alpha_d)$, let us set

$$\partial_h^\alpha := \partial_{h,1}^{\alpha_1} \cdots \partial_{h,d}^{\alpha_d}, \quad \text{where} \quad \partial_{h,j} v(x) = \frac{1}{h}\left(v(x + \frac{1}{2} h\, e_j) - v(x - \frac{1}{2} h\, e_j) \right).$$

We can now formulate a second approximation result.

Theorem 2 (Bramble and Schatz[5]). *Let ν and ℓ be two natural numbers. Let $K^{\nu,\ell}$ be a function in $H^\ell(R^d)$ with compact support such that*

$$K^{\nu,\ell} \quad \text{is a linear combination of B-splines,} \tag{2}$$
$$K^{\nu,\ell} \star v = v,$$

for all polynomials v of degree $\nu - 1$. Set $K_h^{\nu,\ell}(x) = K^{\nu,\ell}(x/h)/h^d$. Now, let u_h be a function in $L^2(\Omega_1)$ and let u be a function in $H^\nu(\Omega_1)$. Let Ω_0 be such that $\Omega_0 +$ support of $K_h^{\nu,\ell} \subset \Omega_1$. Then there is a constant $C_{d,\nu,\ell}$ that depends solely on d, ν, and ℓ such that

$$\| u - K_h^{\nu,\ell} \star u_h \|_{0,\Omega_0} \leq C_{d,\nu,\ell} \left\{ h^\nu \| u \|_{\nu,\Omega_1} + \sum_{|\alpha| \leq \ell} \| \partial_h^\alpha (u - u_h) \|_{-\ell,\Omega_1} \right\}. \tag{3}$$

Comparing the new error estimate (3) with the previous estimate (1), we can see that the price we must pay for working with a convolution kernel with a small support is that we now have to estimate, not the negative-order norm of $u - u_h$, but the negative-order norm of *divided differences* of $u - u_h$, $\partial_h^\alpha (u - u_h)$. This is possible thanks to the property (2), since derivatives of smooth B-splines are divided finite differences of less smooth B-splines.

The construction of the convolution kernels can be found in [5], [15], or in [16].

Negative-order norm error estimates. We now give our negative-order norm error estimates for finite element methods for hyperbolic equations. For the sake of simplicity, we consider the following model problem:

$$u_t + \nabla \cdot (a\, u) = 0, \qquad \text{in } R^d \times (0,T), \tag{4}$$
$$u(x,0) = u_0(x) \qquad \text{for } x \in R^d, \tag{5}$$

and we assume the initial data u_0 to be smooth. We consider the approximate solution u_h determined by the DG method with polynomials of degree k.

Theorem 3. *Let u be the exact solution of problem (4) and (5), and let u_h be the approximation given by the DG method with polynomials of degree k. Then, we have*

$$\| u(T) - u_h(T) \|_{-k-1,\Omega_0} \leq C \| u_0 \|_{k+1,R^d} h^{2k+1},$$

where C depends on T, k and the regularity of the mesh.

The error estimates. Now we only have to put together the results obtained in the above sections. Thus, by combining Theorem 1 and Theorem 3 with $\nu \geq 2k+1$ and $\ell = k+1$, we get

$$\| u(T) - K_{\epsilon_{opt}}^\nu \star u_h(T) \|_{0,\Omega_0} \leq C \| u_0 \|_{\nu,R^d} h^{(2k+1)(1-\theta)}.$$

Notice that if $\theta < 1/2$, the order of convergence of $\| u(T) - K^{\nu}_{\epsilon_{opt}} \star u_h(T) \|_{0,\Omega_0}$ is higher that the order of convergence of $\| u(T) - u_h(T) \|_{0,\Omega_0}$, $k+1/2$. However, since in our case the number of elements that the support of the convolution kernel contains is of the order of $(\epsilon_{opt}/h)^d \geq C\, h^{d(2\theta-1)}$, performing the post-processing might be a computationally expensive undertaking.

This situation can be significantly improved if the mesh is uniform. Indeed, a simple application of Theorem 3 gives that

$$\begin{aligned} \| \partial_h^{\alpha}(u(T) - u_h(T)) \|_{-k-1,\Omega_0} &\leq C \| \partial_h^{\alpha} u_0 \|_{k+1,R^d} h^{2k+1} \\ &\leq C' \| u_0 \|_{k+1+|\alpha|,R^d} h^{2k+1}. \end{aligned}$$

Inserting this estimate in Theorem 2, with $\nu = 2k+1$ and $\ell = k+1$, we get

$$\| u(T) - K_h^{2k+1,k+1} \star u_h(T) \|_{0,\Omega_0} \leq C \| u_0 \|_{2k+2,R^d} h^{2k+1}.$$

Notice that now the support of the kernel contains a number of elements proportional to k^d, which is significantly smaller than the number of elements needed for non-uniform meshes.

The simple but illustrative results displayed in the introduction were obtained with this local post-processing technique.

4 Conclusions

In this paper, we have shown, for a simple model problem and the DG method, that a simple post-processing of the approximate solution of finite element of time-dependent linear hyperbolic problems, can dramatically enhance the quality of the approximation. The case of *locally smooth* exact solutions and *locally uniform* triangulations will be treated in [6] where general hyperbolic systems and other finite element methods will be considered.

Acknowledgements

The authors would like to thank Paul Castillo for providing the figures appearing in this paper and to Fernando Reitich for fruitful discussions. The first author was supported in part by NSF Grant DMS-9807491 and by the University of Minnesota Supercomputing Institute. The second author was supported in part by NSF Grant DMS 95-05077, by AFOSR Grant F49620-98-1-0433, by ARO Grant DAAG55-98-1-0335, by the Institute for Mathematics and its Applications, and by the Minnesota Supercomputing Institute. The third author was supported in part by ARO Grant DAAG55-97-1-0318, NSF Grant DMS-9804985, NASA Langley Grant NAG-1-2070 and AFOSR Grant F49620-99-1-0077. The fourth author is grateful to the IMA at the University of Minnesota and the University of Minnesota Supercomputing Institute for their generous support.

References

1. S. Adjerid, M. Aiffa, and J.E. Flaherty. High-order finite element methods for singularly-perturbed elliptic and parabolic problems. *SIAM J. Appl. Math.*, 55:520–543, 1995.
2. S. Adjerid, J.E. Flaherty, and L. Krivodonova. Superconvergence and a posteriori error estimation for continuous and discontinuous Galerkin methods applied to singularly perturbed parabolic and hyperbolic problems. in preparation.
3. M.Y.T. Apelkrans. Some properties of difference schemes for hyperbolic equations with discontinuities and a new method with almost quadratic convergence. Technical Report 15A, Uppsala University, Dept. of Computer Science, 1969.
4. R. Biswas, K.D. Devine, and J. Flaherty. Parallel, adaptive finite element methods for conservation laws. *Appl. Numer. Math.*, 14:255–283, 1994.
5. J.H. Bramble and A.H. Schatz. Higher order local accuracy by averaging in the finite element method. *Math. Comp.*, 31:94–111, 1977.
6. B. Cockburn, M. Luskin, C.-W. Shu, and E. Süli. Enhanced accuracy by post-processing for finite element methods for hyperbolic equations. in preparation.
7. J. Douglas, Jr. Superconvergence in the pressure in the simulation of miscible displacement. *SIAM J. Numer. Anal.*, 22:962–969, 1985.
8. R.P. Fedorenko. The application of high-accuracy difference schemes to the numerical solution of hyperbolic equations. *Zh. Vychisl. Mat. i Mat. Fiz.*, 2:1122–1128, 1962. (in Russian).
9. D. Gottlieb and E. Tadmor. Recovering pointwise values of discontinuous data within spectral accuracy. In *Proceedings of U.S.-Israel Workshop*, volume 6 of *Progress in Scientific Computing*, pages 357–375. Birkhäuser Boston Inc., 1985.
10. C. Johnson and J. Pitkäranta. An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comp.*, 46:1–26, 1986.
11. R.B. Lowrie. *Compact higher-order numerical methods for hyperbolic conservation laws.* PhD thesis, University of Michigan, 1996.
12. A. Majda, J. McDonough, and S. Osher. The Fourier method for nonsmooth initial data. *Math. Comp.*, 32:1041–1081, 1978.
13. A. Majda and S. Osher. Propagation of error into regions of smoothness for accurate difference approximate solutions to hyperbolic equations. *Comm. Pure and Appl. Math*, 30:671–705, 1977.
14. M.S. Mock and P.D. Lax. The computation of discontinuous solutions of linear hyperbolic equations. *Comm. Pure and Appl. Math*, 31:423–430, 1978.
15. V. Thomée. High order local approximations to derivatives in the finite element method. *Math. Comp.*, 31:652–660, 1977.
16. L.B. Wahlbin. *Superconvergence in Galerkin Finite Element Methods*, volume 1605 of *Lecture Notes in Mathematics*. Springer Verlag, 1995.

Introduction to Discontinuous Wavelets

Nicholas Coult

Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, MN 55455

Abstract. Wavelets provide a tool for efficient representation of functions. This efficient representation has proven useful in the numerical solution of non-linear evolution equations. In this paper we provide a brief review of the use of wavelets for efficient representation of functions, and in particular we describe the piecewise-discontinuous basis of wavelets proposed by Alpert. We review the useful properties this basis has for the solution of PDE's, and introduce an illustrative approach to the representation of boundary conditions. We also discuss the extension to higher dimensional problems.

1 Introduction

Wavelets have enjoyed success as a tool for efficient representation of functions and operators (see e.g. [5] and [3] and references therein). More recently, wavelet-based algorithms for the solution of non-linear evolution equations have been developed [4], [2]. In our view, the basic idea is to use the efficient and adaptive representation the wavelet basis offers at all stages in the computation of the solution. By using this "under the hood" replacement, we can construct fast adaptive numerical solvers for these equations. In particular, suppose we are solving a non-linear evolution equation whose approximate wavelet solution at time t is given by $u(t)$. If this solution is relatively smooth except for a finite number of singularities, then we require relatively few coefficients in the wavelet basis to represent it accurately; we call such an expansion "sparse". If we are able to compute the updated solution $u(t + \Delta t)$ in such a way that we maintain this sparse representation at all times during the intermediate computations, then we know by virtue of the small number of coefficients that our method is efficient. Typically this is achieved by maintaining an equivalently sparse representation for all the of operators involved, see [3] and [2].

We do not concern ourselves in this paper with the particular equations to be solved or the methods used to solve them. Rather, we aim to provide an introduction to the basic concepts of sparse wavelet expansions for functions, and the advantageous properties of a particular wavelet system proposed by Alpert in [1]. This system has the usual vanishing moment property of other wavelet systems, which is the crucial characteristic for sparse representations. Additionally, this system permits

1. an interpolating basis, so that the conversion from point-values of a function to its coefficients in the basis can be done using a diagonal transformation
2. an adaptive algorithm for applying non-linear operators to functions
3. convenient, accurate representation of boundary conditions
4. extensions to higher dimensions using e.g. rectangles or triangles

Items 1, 2, and 3 are covered in detail in [2]. We provide a review of the Alpert system, efficient representation of functions, and items 1 and 2 in Sections 2, 3, and 4. Finally, in Section 5 we present an illustrative boundary condition construction, and in Section 6 we discuss the extension of this basis to higher dimensional problems.

2 Multiwavelets and multiresolution analysis

We consider a multiresolution analysis (MRA) to be an infinite sequence of nested spaces $\{\mathbf{V}_j\}$ such that $\mathbf{V}_j \subset \mathbf{V}_{j+1}$. Each space $\mathbf{V}_j$ resolves a particular scale of a larger function space, e.g. $L^2(\mathbf{R})$. The wavelet spaces provide a description of the purely fine-scale information which is lost in the transition from $\mathbf{V}_{j+1}$ to $\mathbf{V}_j$ (i.e. the orthogonal complement of $\mathbf{V}_j$ in $\mathbf{V}_{j+1}$), and the wavelets themselves are simply an orthonormal basis for this complementary space. In this section we present the concept of a multiresolution analysis using Alpert's piecewise-discontinuous multiwavelet construction [1]. (For more on wavelets and MRA's, see e.g. [5]).

First, let us consider an L^2-orthonormal basis of m-th degree piecewise polynomials given by $\{\phi_0, \ldots, \phi_m\}$. We choose these functions so that they are polynomials on $[0,1]$ and zero elsewhere. The Legendre polynomials, translated and dilated to the unit interval, are a good example and are used in [1].

Then, we define the translates and dyadic dilates of these functions by

$$\phi_{l,k}^j(x) = 2^{j/2}\phi_l(2^j x - k). \tag{1}$$

These functions are called the *scaling functions*. We define the space $\mathbf{V}_j$ as

$$\mathbf{V}_j = \mathrm{Span}\left\{\phi_{l,k}^j(x)\right\}_{\substack{l=0,\ldots,m \\ k=0,\ldots 2^j-1}} \tag{2}$$

Note that $\dim(\mathbf{V}_j) = 2^j(m+1)$, and the supports of $\phi_{l,k}^j(x)$ and $\phi_{l',k'}^j(x)$ are disjoint if and only if $k \neq k'$.

We also see that (due to (1)), if $f(x) \in \mathbf{V}_j$, then $f(2x) \in \mathbf{V}_{j+1}$. Thus, the space $\mathbf{V}_j$ may be thought of as a resolution of $L^2[0,1]$, and as j increases, this resolution becomes finer. The collection of spaces $\{\mathbf{V}_j\}_{j=0,\ldots,\infty}$ is called a *multiresolution analysis.*

Now we define the orthogonal complement of $\mathbf{V}_j$ in $\mathbf{V}_{j+1}$,

$$\mathbf{W}_j \oplus \mathbf{V}_j = \mathbf{V}_{j+1}. \tag{3}$$

We view $\mathbf{W}_j$ as the "detail" space, because it contains the information which is lost in the transition from $\mathbf{V}_{j+1}$ to the next coarser scale $\mathbf{V}_j$. In the context of the MRA, this space is called the wavelet space, and it can easily be shown that

$$L^2[0,1] = \mathbf{V}_0 \oplus \left(\bigoplus_{j=0}^{\infty} \mathbf{W}_j \right). \tag{4}$$

We also see that we may construct an orthonormal basis for $\mathbf{W}_j$ via

$$\psi_{l,k}^j(x) = 2^{j/2}\psi_l(2^j x - k), \tag{5}$$

where $\{\psi_l(x)\}_{l=0,\ldots,m}$ is an appropriately chosen orthonormal set. We do not explicitly list these functions here, see [1] for such constructions.

The relation (5) is very similar to the usual such relation for wavelets, except that instead of the wavelets being defined as translations and dyadic dilations of a single function ψ, (5) uses the set of functions $\{\psi_l\}_{l=0,\ldots,m}$. Hence, we use the term *multiwavelets* for such a construction. We note that the piecewise-polynomial multiwavelets and scaling functions defined above are *discontinuous*. This quality distinguishes the above construction from most (if not all) other non-trivial wavelet systems.

Because of the property (4), we may represent any function $f \in L^2[0,1]$ in terms of its wavelet coefficients (plus its coefficients on the coarsest scale $\mathbf{V}_0$) as

$$f(x) = \sum_{l=0}^{m} s_{l,0}^0 \phi_{l,0}^0(x) + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \sum_{l=0}^{m} d_{l,k}^j \psi_{l,k}^j(x), \tag{6}$$

where $s_{l,k}^j = (f, \phi_{l,k}^j)$ and $d_{l,k}^j = (f, \psi_{l,k}^j)$.

We may use this basis for a Galerkin solution of a PDE. If we use the scaling function basis on some scale $\mathbf{V}_j$, then this discretization corresponds to a typical discontinuous Galerkin approximation. Alternatively we can use the wavelet basis; the problem is the same as in the previous case, but we are simply using a different basis. The advantage to this is that functions may be more efficiently represented in the wavelet basis, as shown in the next section.

3 Efficient Representation of Functions

One of the chief advantages of the wavelet system of coordinates is that many functions have a sparse expansion in the wavelet basis. By "sparse expansion," we mean that for a given small threshold parameter, many of the coefficients in the wavelet expansion are smaller in magnitude than this threshold. We may therefore discard most of the wavelet coefficients and store only the remaining coefficients, resulting in a significant savings in storage space. Since the wavelet basis is orthonormal, a small truncation in the wavelet coefficients

perturbs the function we are representing by only a small amount in the L^2 norm.

To see that the wavelet expansion has this property, we appeal to the vanishing moments property of the wavelet basis. By construction, the wavelet functions $\psi^j_{l,k}(x)$ have $m+1$ vanishing moments; that is

$$\int_0^1 \psi^j_{l,k}(x) x^i \, dx = 0 \quad \text{for } i = 0, \ldots, m. \tag{7}$$

For any function $f \in C^{m+1}$, we see via Taylor's Theorem and (7) that

$$|d^j_{l,k}| = \left| \int_0^1 f(x) \psi^j_{l,k}(x) \, dx \right| \leq C_{f,m} \, 2^{-j(m+1)}, \tag{8}$$

where the constant $C_{f,m}$ depends on m and $f^{(m+1)}(x)$.

In practice, sparsity is achieved as follows. We see from (8) that if f is smooth in some region, then the wavelet coefficients which correspond to that spatial region are small in magnitude. Thus, if we truncate to zero those wavelet coefficients which are smaller than our desired accuracy ϵ, and store only the remaining non-zero coefficients, we may achieve a significant savings in storage. For functions with a finite number of singularities, one can easily show that the number of coefficients can be reduced from N to $\mathcal{O}(\log N)$, where N is the number of points in the fine-scale representation. Since the wavelet transform is orthogonal, this perturbation does not change the function being represented by more than ϵ in the relative $L^2[0,1]$ norm.

The key to fast numerical algorithms using this representation is that numerical operations such as differentation, integration, sums, products, etc. are then applied to functions represented in this sparse (or compressed) form. A much more detailed description in the context of scalar wavelets may be found in [3].

Remark. We note that since the support of $\psi^j_{l,k}$ exactly covers the supports of $\phi^{j+1}_{l,2k}$ and $\phi^{j+1}_{l,2k+1}$, we may choose to store the coefficients $s^j_{l,k}$ instead of $d^j_{l,k}$ in our compressed representation. The storage requirements are identical. This means that the sparse multiwavelet expansions corresponds to a recursive dyadic partition of the interval. The advantage to this representation will be described in the next section.

4 Interpolating basis and non-linear operators

In this section we describe an interpolating orthonormal basis, which allows for a diagonal transformation between values of a function and its coefficients in the basis. This leads to an adaptive algorithm for computing non-linear functions of functions in compressed form.

In [1], the functions $\phi_k(x)$ are chosen to be the Legendre polynomials $\pi_k(x)$ (appropriately scaled, dilated and translated so that they form an orthonormal basis on $[0,1]$). The disadvantage of this construction is that the quadratures which compute the expansion of a function f in this basis require $(m+1)^2$ multiply-add operations (not including evaluation of f).

Instead (as is also done in [2]) we construct an interpolating basis. Let $\{r_k\}_{k=0,\ldots,m}$ denote the roots of $\pi_{m+1}(x)$. Then, we define

$$\phi_k(x) = c_k \prod_{\substack{l=0,\ldots,m \\ l \neq k}} (x - r_l) \tag{9}$$

where c_k is a constant chosen so that $\|\phi_k(x)\|_{L^2[0,1]} = 1$. Thus, computing the coefficient $f_k = (f, \phi_k)$ requires a single-point quadrature obtained by evaluating $f(r_k)$, and the expansion of f in the basis defined by $\phi_k(x)$ requires only $m+1$ multiplications and no additions (not including the evaluation of f at $x = r_0, r_1, \ldots, r_m$). This interpolating property also proves useful in the application of non-linear operators to functions represented in this basis, as we describe below.

An adaptive algorithm for computing a pointwise non-linear function of a function in the multiwavelet basis is presented in [2]. We review this algorithm here. For simplicity we consider the function $g(u) = u^2$. (We note that $uv = \frac{1}{4}((u+v)^2 - (u-v)^2)$, so the following analysis applies to ordinary products as well. By Taylor series expansions, we can extend it to a wide class of non-linear operators). Clearly it is the case that if $u \in \mathbf{V}_j$, and $u^2 \in \mathbf{V}_j$, then $(u^2, \phi_k) = d_k (u, \phi_k)^2$, where d_k is some constant; that is, we can find the coefficients of the square of u simply by squaring its coefficients. This is due to the interpolating property of the scaling functions.

Furthermore, if we have a function u in compressed form, we recall that this representation is equivalent to the scaling function coefficients of u on some dyadic partition of the unit interval (see the remark in Section 3). For each subinterval I_k in this partition, it is clear that $u|_{I_k} \in \mathbf{V}_{j_k}$ for some j_k, where j_k depends on k. However, we see that in general it is not the case that $u|_{I_k} \in \mathbf{V}_{j_k} \Rightarrow u^2 \in \mathbf{V}_{j_k}$. Instead, we achieve this relationship only approximately. We can easily see that for any ϵ, there exists $j'_k \geq 0$ and $e_{j'_k} \in L^2[0,1]$ such that

$$u|_{I_k} \in \mathbf{V}_{j_k + j'_k} \Rightarrow (u^2 + e_{j'_k})|_{I_k} \in \mathbf{V}_{j_k + j'_k} \text{ and } \|e_{j'_k}\|_{L^2[0,1]} \leq \epsilon. \tag{10}$$

In practice, we usually require only $j'_k = 1$ (see [2]), which means oversampling by a factor of two in order to compute the square of u.

5 Boundary conditions

A construction which permits boundary conditions is relatively simple. The usual construction in e.g. [2] uses a weak representation of boundary conditions via the choice of numerical fluxes at the boundary. To illustrate the

simplicity of constructing boundary conditions, we instead modify the scaling functions whose support touches the boundary so that the modified scaling functions satisfy the boundary conditions. Then, in the case of finite-difference type discretizations of differential operators, we modify the finite-difference scheme at the boundary to maintain the accuracy of the scheme.

In the case of boundary conditions on the left, we first select an independent orthonormal set of $m+1$ degree polynomials $\{\tilde{\phi}_l(x)\}_{l=0,\ldots,m-1}$ such that $\tilde{\phi}_l(0)$ satisfy the boundary conditions. Then, we construct the orthogonal projection matrix $P = \{p_{k,l}\}_{\substack{k=0,\ldots,m-1 \\ l=0,\ldots,m}}$ using the expansion

$$\tilde{\phi}_k(x) = \sum_{l=0}^{m} p_{k,l}\phi_l(x). \tag{11}$$

We derive modified wavelet functions $\{\tilde{\psi}_k(x)\}_{k=0,\ldots,m}$ by first using the basis

$$\gamma_k(x) = \sum_{i=0}^{m} g_{k,i}^{(0)} \sum_{l=0}^{m-1} p_{l,i}\tilde{\phi}_i(2x) + \sum_{i=0}^{m} g_{k,i}^{(1)}\phi_i(2x-1) \text{ for } k = 0,\ldots,m \tag{12}$$

and then constructing $\{\tilde{\psi}_k(x)\}_{k=0,\ldots,m}$ via Gram-Schmidt orthogonalization of $\{\gamma_k(x)\}_{k=0,\ldots,m}$.

Finally, our modified scaling functions are given by

$$\tilde{\phi}_{l,k}^{j}(x) = \begin{cases} 2^{j/2}\phi_l(2^j x - k) & k > 0 \\ 2^{j/2}\tilde{\phi}_l(2^j x) & k = 0 \end{cases} \tag{13}$$

with the modified wavelets $\tilde{\psi}_{l,k}^{j}(x)$ defined similarly. For boundary conditions on the right, we can follow a very similar construction.

Let us denote by Q the projector matrix onto the basis which satisfies the right boundary conditions. Suppose that we have computed the tridiagonal interior stencil for a Galerkin approximation to d/dx on $\mathbf{V}_j$ such that this stencil is exact for m-th degree polynomials. This stencil is defined by three $(m+1)\times(m+1)$ matrices M_0, M_1, and M_{-1}, where M_0 is the matrix on the diagonal and M_1,M_{-1} are the off-diagonal matrices. Let us define an $(m+1)\times(m+1)$ matrix L which takes the scaling-function coefficients on a given interval and computes the coefficients of this function extended as an m-th degree polynomial into the interval to the left. Similarly, we define a right-extension matrix R. Now, we can construct a matrix $\mathbf{T}_j$ which is an approximation to d/dx with Dirichlet or other homogeneous linear boundary conditions. This matrix is given by

$$\begin{pmatrix} P(M_0 + M_{-1}L)P^* & PM_1 & 0 & \cdots & 0 \\ M_{-1}P^* & M_0 & M_1 & \cdots & 0 \\ 0 & M_{-1} & M_0 & M_1 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ & & M_{-1} & M_0 & M_1Q^* \\ 0 & \cdots & 0 & QM_{-1} & Q(M_0 + M_1R)Q^* \end{pmatrix} \tag{14}$$

Clearly it is the case that for a polynomial of degree m which satisfies the boundary conditions, the operator defined by the above matrix will compute the first derivative of this polynomial exactly. This means that the truncation error of the matrix $\mathbf{T}_j$ as an approximation to d/dx is $\mathcal{O}(h^{m+1})$ at all points, including as those that are close to the boundary.

6 Higher dimensional problems

Finally, we present a brief discussion of higher dimensional problems. The material below applies directly to two-dimensional problems, but the extension to higher dimensions follows by direct extension.

For rectangular-shaped domains in two dimensions, the usual construction is a tensor-product basis, $\phi^j_{k,k',l,l'}(x,y) = \phi^j_{k,l}(x)\phi^{j'}_{k',l'}(y)$. The interpolating, boundary condition, and non-linear function evaluation properties therefore directly extend to this two-dimensional basis. Stretching and shearing preserve the orthogonality of the basis, so we may extend this construction to any parellelogram in the plane.

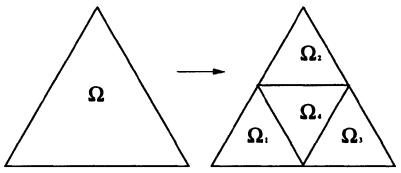


Fig. 1. Subdivision of a triangle.

A construction on triangles is possible, see [8] and [7] for barysymmetric and piecewise-linear examples, respectively. In this case, we require an orthonormal basis of polynomials on a triangle; examples may be found in [8] and [6]. The coarsest scale is defined as m-th degree polynomials restricted to the triangle; the finer scales are achieved by recursively subdividing each triangle at a given scale into four smaller triangles as shown in Figure 1. The wavelet spaces are defined as usual. This construction automatically produces a piecewise-discontinuous system of multiwavelets on *any* triangle in the plane, since an orthogonal set of functions in the plane remains orthogonal under any affine transformation of the plane.

The boundary condition technique of Section 5 extends directly to this triangular construction. However, we do not know of any existing derivations of arbitrary-degree interpolating orthogonal polynomials on a triangle, so the interpolating property in Section 4 can not currently be obtained. The algorithm for multiplication of functions in compressed form is still possible with such a basis, but we no longer can achieve the convenient relation

$(u^2, \phi_k) = d_k(u, \phi_k)^2$. Rather, we must use pre-computed matrices of the coefficients of squares of the basis functions. This requires an extra factor in the computation time, but the algorithm still runs directly on functions in compressed form.

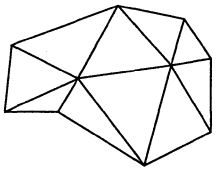


Fig. 2. Multiple triangles pieced together.

As a possible future direction, we see that a logical next step beyond the triangular construction is the piecing together of multiple triangles to form a larger domain with a more complicated shape. Figure 2 shows an example of this. The MRA on this domain is basically defined as the union of the MRA's on the individual triangles. Boundary conditions for the exterior edges may be easily represented using the methodology of Section 5. Matching across the interior boundaries can be achieved in the representation of the operator. This concept closely parallels discontinuous finite element methods for triangles [6].

References

1. B. Alpert. A Class of Bases in l^2 for the Sparse Representation of Integral Operators. *SIAM J. Math. Anal*, 24(1):246–262, 1993.
2. B. Alpert, G. Beylkin, D. Gines, and L. Vozovoi. Adaptive solution of partial differential equations in multiwavelet bases. Preprint, 1999.
3. G. Beylkin, R. R. Coifman, and V. Rokhlin. Fast wavelet transforms and numerical algorithms I. *Comm. Pure and Appl. Math.*, 44:141–183, 1991. Yale University Technical Report YALEU/DCS/RR-696, August 1989.
4. G. Beylkin and J. M. Keiser. On the adaptive numerical solution of nonlinear partial differential equations in wavelet bases. *J. Comput. Phys.*, 132(2):233–259, 1997.
5. I. Daubechies. *Ten Lectures on Wavelets.* CBMS-NSF Series in Applied Mathematics. SIAM, 1992.
6. Spencer J. Sherwin and George Em. Karniadakis. A new triangular and tetrahedral basis for high-order (*hp*) finite element methods. *Internat. J. Numer. Methods Engrg.*, 38(22):3775–3802, 1995.
7. T. von Petersdorff, C. Schwab, and R. Schneider. Multiwavelets for second-kind integral equations. *SIAM J. Numer. Anal.*, 34(6):2212–2227, 1997.
8. T. Yu, K. Kolarov, and W. Lynch. Barysymmetric multiwavelets on triangle. Preprint, March 1997.

The Local Discontinuous Galerkin Method for Contaminant Transport Problems

Clint Dawson[1], Vadym Aizinger[1] and Bernardo Cockburn[2]

[1] Texas Institute for Computational and Applied Mathematics, University of Texas at Austin, Austin, TX 78712
[2] Department of Mathematics, University of Minnesota, Minneapolis, MN 55455

Abstract. We discuss the application of the Local Discontinuous Galerkin method to the approximation of contaminant transport in porous media.

1 Introduction

In this paper, we consider the flow of an incompressible fluid through a homogeneous saturated porous medium, where the fluid is contaminated by one or more solutes with concentrations $c_i, i = 1, ..., n$. We assume the flow to be at steady-state, and the transport to be described by advection, molecular diffusion, mechanical dispersion and chemical reactions (adsorption) between a solute and the surrounding porous skeleton. Mathematically, this process is modeled by a possibly nonlinear partial differential equation, in the case of contamination by one solute, or a non-linear coupled system of PDEs if the fluid is contaminated by several substances. The equation (or system) can be purely hyperbolic or purely parabolic depending on whether we include diffusion and dispersion terms.

A major challenge in designing numerical schemes for porous medium flow is to create a method which gives sharp resolution of fronts for discontinuous or rough problems but, at the same time, provides a high order scheme for smooth problems. Here we will investigate the Local Discontinuous Galerkin (LDG) method [1], which appears to be well-suited for these types of problems. On the one hand, it uses higher order approximating spaces and exhibits optimal (or close to optimal) orders of convergence for smooth problems, on the other, it possesses a capability of resolving steep fronts. It is also easily extendable to systems of equations and multidimensions.

For simplicity, we treat in this paper only the case of one-dimensional flow but the method can be generalized for multidimensional problems. In section 2, we formulate briefly the mathematical model of contaminant transport in porous media; in section 3 we give a description of the method and in section 4 present numerical results for some test problems pertaining to contaminant transport.

* This work was supported in part by National Science Foundation Grant DMS-9805491

2 Mathematical model of contaminant transport in porous medium

For one component, mass-conservation of the contaminant gives us the equation [2]

$$c_t + \frac{1-\epsilon}{\epsilon}\,\tilde{s}_t + u\,c_x - D\,c_{xx} = 0, \tag{1}$$

where c is the concentration of solute in moles per unit volume in the fluid phase, $\tilde{s}\,(x,t)$ is the concentration of contaminant adsorbed on the solid matrix in moles per unit volume of solid, $\epsilon > 0$ is the porosity, u is the effective fluid velocity, and $D > 0$ accounts for molecular diffusion and mechanical dispersion. For simplicity, ϵ, u, and D are assumed constant. We will also assume $u > 0$.

The chemical reactions describing adsorption may be fast (equilibrium) or slow (non-equilibrium) depending on the rate of reaction with respect to the rate of flow. Here we will only consider equilibrium reactions. In this case, the contaminant adsorbed by the solid is generally assumed to be a function of the concentration in the fluid; that is, $\tilde{s} = f(c)$. The function f is called an adsorption isotherm. A common isotherm is the Langmuir isotherm $f(c) = \frac{N\,K\,c}{1+K\,c}$, where N is the saturation concentration of the adsorbed solute and $K > 0$ is a rate constant.

Letting $\phi(c) = \frac{1-\epsilon}{\epsilon} f(c)$ and substituting into (1) we obtain

$$c_t + \phi(c)_t + u\,c_x - D\,c_{xx} = 0. \tag{2}$$

Depending on the particular situation being modeled, a similar equation in each component holds for multicomponent contaminant transport:

$$\mathbf{c}_t + \mathbf{\Phi}(\mathbf{c})_t + u\,\mathbf{c}_x - D\,\mathbf{c}_{xx} = 0,\ 0 < x < L,\ t > 0, \tag{3}$$

$$\mathbf{\Phi}(\mathbf{c}) = \frac{1-\epsilon}{\epsilon}\,\mathbf{f}(\mathbf{c}), \tag{4}$$

where $\mathbf{c} = (c_1, c_2, ..., c_n)^T$ and n is the number of components.

The Langmuir isotherm in the multicomponent case is [2]

$$\mathbf{f}(\mathbf{c}) = (f_1(\mathbf{c}), f_2(\mathbf{c}), ..., f_n(\mathbf{c}))^T, \tag{5}$$

where

$$f_i(\mathbf{c}) = \frac{N_i\,K_i\,c_i}{1 + K_1\,c_1 + K_2\,c_2 + ... + K_n\,c_n},\ K_i > 0,\ \ i = 1, ..., n. \tag{6}$$

N_i stands here for the maximum number of moles of solute i that can be adsorbed per unit volume of adsorbent.

We augment (3) with the initial and "inflow" and "outflow" boundary conditions

$$\mathbf{c}(x,0) = \mathbf{c}^0(x), 0 < x < L, \tag{7}$$

$$u\mathbf{c}(0,t) - D\mathbf{c}_x(0,t) = u\mathbf{c}_I(t), \quad t > 0, \tag{8}$$

$$D\mathbf{c}_x(L,t) = 0 \quad t > 0. \tag{9}$$

3 The Local Discontinuous Galerkin Method

3.1 Formulation of the method

To define the LDG method, we introduce the new variable $\mathbf{q} = \sqrt{D}\,\mathbf{c}_x$, denote

$$\mathbf{s}(x,t) = \mathbf{c}(x,t) + \mathbf{\Phi}(\mathbf{c}(x,t)) \tag{10}$$

and rewrite the problem (3) as follows:

$$\mathbf{s}_t + (u\mathbf{c} - \sqrt{D}\mathbf{q})_x = 0, \tag{11}$$

$$\mathbf{q} - \sqrt{D}\mathbf{c}_x = 0, \qquad 0 < x < L, t > 0. \tag{12}$$

The LDG method for (3) is now obtained by simply discretizing the system above with a discontinuous Galerkin method. We define the flux $\mathbf{h} = (\mathbf{h}_c, \mathbf{h}_q)^T$ as follows:

$$\mathbf{h}(\mathbf{c}, \mathbf{q}) = \begin{pmatrix} u\mathbf{c} - \sqrt{D}\mathbf{q} \\ -\sqrt{D}\mathbf{c} \end{pmatrix}. \tag{13}$$

For each partition of the interval $(0, L)$, $\{ x_{j+\frac{1}{2}} \}_{j=0}^{N}$ we set $I_j = (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ and $\Delta x_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$ for $j = 1, \ldots, N$. We will use the standard notation $(u, v)_{I_j}$ to denote the $L^2(I_j)$ inner product. Furthermore, we define $p^{\pm}_{j+\frac{1}{2}} = p(x^{\pm}_{j+\frac{1}{2}})$ where

$$p^{+}_{j+\frac{1}{2}} = \lim_{\substack{x \to x_{j+\frac{1}{2}} \\ x > x_{j+\frac{1}{2}}}} p(x), \quad p^{-}_{j+\frac{1}{2}} = \lim_{\substack{x \to x_{j+\frac{1}{2}} \\ x < x_{j+\frac{1}{2}}}} p(x).$$

We seek an approximation $\mathbf{w}_h = (\mathbf{C}, \mathbf{Q})^T$ to $\mathbf{w} = (\mathbf{c}, \mathbf{q})^T$, such that for each time t, the components of $\mathbf{C}(t)$ and $\mathbf{Q}(t)$ are polynomials of degree at most k on I_j.

The approximate solution given by the LDG method is defined as the solution of the following weak formulation:

$$\Big(\partial_t \mathbf{S}(x,t)\,,\, \mathbf{v}_{h,c}(x)\Big)_{I_j} - \Big(\mathbf{h}_c(\mathbf{w}_h(x,t))\,,\, \partial_x \mathbf{v}_{h,c}(x)\Big)_{I_j}$$
$$+ \hat{\mathbf{h}}_c(\mathbf{w}_h^-, \mathbf{w}_h^+)_{j+\frac{1}{2}}(t)\, \mathbf{v}_{h,c}(x^-_{j+\frac{1}{2}}) - \hat{\mathbf{h}}_c(\mathbf{w}_h^-, \mathbf{w}_h^+)_{j-\frac{1}{2}}(t)\, \mathbf{v}_{h,c}(x^+_{j-\frac{1}{2}}) = 0, \tag{14}$$

$$\Big(\mathbf{Q}(x,t)\,,\, \mathbf{v}_{h,q}(x)\Big)_{I_j} - \Big(\mathbf{h}_q(\mathbf{w}_h(x,t))\,,\, \partial_x \mathbf{v}_{h,q}(x)\Big)_{I_j}$$
$$+ \hat{\mathbf{h}}_q(\mathbf{w}_h^-, \mathbf{w}_h^+)_{j+\frac{1}{2}}(t)\, \mathbf{v}_{h,q}(x^-_{j+\frac{1}{2}}) - \hat{\mathbf{h}}_q(\mathbf{w}_h^-, \mathbf{w}_h^+)_{j-\frac{1}{2}}(t)\, \mathbf{v}_{h,q}(x^+_{j-\frac{1}{2}}) = 0. \tag{15}$$

Here $\mathbf{S} = \mathbf{C} + \mathbf{\Phi}(\mathbf{C})$ and $\hat{\mathbf{h}}_c$ and $\hat{\mathbf{h}}_q$ are numerical fluxes, given below. Moreover $\mathbf{v}_{h,c}$ and $\mathbf{v}_{h,q}$ are vector functions whose components are polynomials of degree at most k on I_j.

To discuss the numerical flux, we use the notation:

$$[p] = p^+ - p^-, \ \overline{p} = \frac{1}{2}(p^+ + p^-).$$

We write the numerical flux as the sum of an advective flux and a diffusive flux:

$$\hat{\mathbf{h}}(\mathbf{w}_h^-, \mathbf{w}_h^+) = \hat{\mathbf{h}}_{adv}(\mathbf{w}_h^-, \mathbf{w}_h^+) + \hat{\mathbf{h}}_{diff}(\mathbf{w}_h^-, \mathbf{w}_h^+). \tag{16}$$

The advective flux is given by

$$\hat{\mathbf{h}}_{adv}(\mathbf{w}_h^-, \mathbf{w}_h^+) = (\hat{\mathbf{g}}(\mathbf{C}^-, \mathbf{C}^+), 0)^T, \tag{17}$$

where $\hat{\mathbf{g}}(\mathbf{C}^-, \mathbf{C}^+)$ is computed using upwinding. For the i-th component of $\hat{\mathbf{g}}$,

$$\hat{g}_i(\mathbf{C}^-, \mathbf{C}^+)_{j+1/2}(t) = \begin{cases} u\, C_i(x^-_{j+\frac{1}{2}}, t), \text{ if } u \geq 0, \\ u\, C_i(x^+_{j+\frac{1}{2}}, t), \text{ if } u < 0. \end{cases} \tag{18}$$

The diffusive flux is given by

$$\left(\hat{\mathbf{h}}_{diff}(\mathbf{w}_h^-, \mathbf{w}_h^+)_{j+\frac{1}{2}}(t)\right)_i = \left\{ \begin{pmatrix} -\sqrt{D}\,\overline{Q}_i \\ -\sqrt{D}\,\overline{C}_i \end{pmatrix} - \tilde{\mathbf{C}}_{diff}\,[\mathbf{w}_{h,i}] \right\}(x_{j+\frac{1}{2}}, t), \tag{19}$$

where

$$\tilde{\mathbf{C}}_{diff} = \begin{pmatrix} 0 & c_{12} \\ -c_{12} & 0 \end{pmatrix}, \tag{20}$$

$$c_{12} = c_{12}(\mathbf{w}_h^-, \mathbf{w}_h^+) \quad \text{is locally Lipschitz,} \tag{21}$$

$$c_{12} \equiv 0, \quad \text{when } D = 0. \tag{22}$$

At the boundaries $x = 0$ and $x = L$ we define

$$\hat{\mathbf{h}}(\mathbf{w}_h^-, \mathbf{w}_h^+)_{\frac{1}{2}}(t) = (u\mathbf{c}_I(t), -\sqrt{D}\,\overline{\mathbf{C}}(x_{\frac{1}{2}}, t) + c_{12}[\mathbf{C}](x_{\frac{1}{2}}, t))^T, \tag{23}$$

$$\hat{\mathbf{h}}(\mathbf{w}_h^-, \mathbf{w}_h^+)_{N+\frac{1}{2}}(t) = (u\mathbf{C}(x^-_{N+\frac{1}{2}}, t), -\sqrt{D}\mathbf{C}(x^-_{N+\frac{1}{2}}, t))^T, \tag{24}$$

in accordance with the boundary conditions (8) and (9), where

$$\overline{\mathbf{C}}(x_{\frac{1}{2}}) \equiv \frac{1}{2}(\mathbf{C}^+_{\frac{1}{2}} + \mathbf{c}_I),$$

and

$$[\mathbf{C}](x_{\frac{1}{2}}) \equiv \mathbf{C}^+_{\frac{1}{2}} - \mathbf{c}_I.$$

It is important to note that by (15) the degrees of freedom for $\mathbf{Q}$ can be eliminated in terms of degrees of freedom in $\mathbf{C}$, thus giving a system in $\mathbf{C}$ unknowns only.

The method above must be discretized in time. Moreover, for higher order polynomials, the higher order terms in the solution must be controlled to inhibit oscillations. Here we follow earlier work of Cockburn and Shu as cited in [1], and use explicit third order Runge-Kutta procedures in time combined with local projection operators in space to control numerical oscillations.

4 Some numerical results for contaminant transport problems

In this section, we present numerical results obtained using the LDG method applied to a typical contaminant transport problem.

We consider the advection-diffusion system described above with $u = 1$, the isotherm function

$$\Phi(\mathbf{c}) = \begin{pmatrix} \frac{c_1}{1 + c_1 + 10\,c_2} \\ \\ \frac{10\,c_2}{1 + c_1 + 10\,c_2} \end{pmatrix}, \tag{25}$$

the initial condition $\mathbf{c}(x, 0) = (0, 0)$ and the boundary condition $\mathbf{c}_I(0, t) = (1, 1)$.

We compute an approximate solution at times $T \in \{0.2, 0.5, 0.8\}$ using the LDG method with piecewise constant approximating functions on 320 elements for $D = 0$.

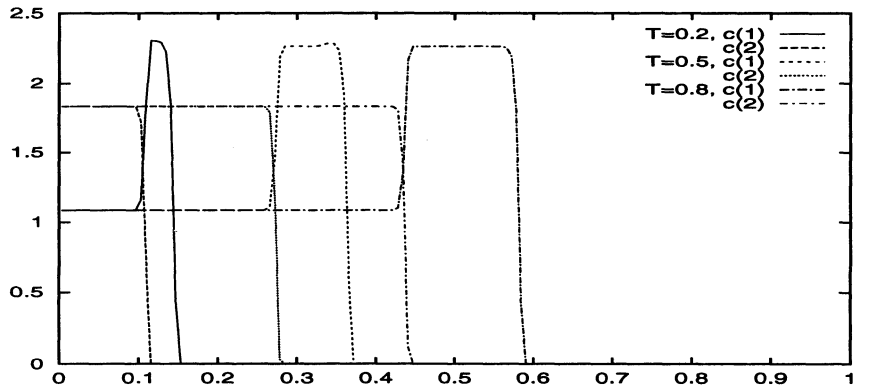


Fig. 1. Contaminant concentration at different times, $D = 0$.

The result, given in Figure 1, is a moving shock wave, which agrees with the theoretical results presented in [2].

In the next numerical experiments we compare results of computations on the same test problem by the LDG method using approximating spaces of different order and with different number of elements for the parabolic system with $D = 0.01$.

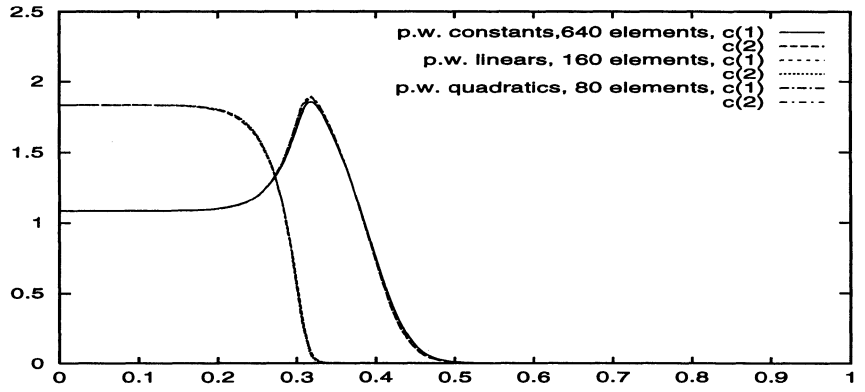


Fig. 2. Comparison of solutions obtained using different approximating spaces, $D = 0.01$.

We see that all three approximate solutions lie very close together, which illustrates a possibility of dramatic improvement in convergence with fewer degrees of freedom by using higher order approximating spaces.

References

1. B. Cockburn and C. W. Shu, *The local discontinuous Galerkin Method for time dependent convection-diffusion systems*, SIAM J. Numer. Anal., Vol. 35, pp. 2440-2463, 1998.
2. H. K. Rhee, R. Aris, N. R. Amundson, *First-Order Partial Differential Equations*, PRENTICE HALL, Englewood Cliffs, New Jersey, 1989.

Discontinuous Galerkin Method for the Numerical Solution of Euler Equations in Axisymmetric Geometry

Bruno Despres

Commissariat à l'Energie Atomique, BP12, 91680, Bruyeres le Chatel, FRANCE, Laboratoire d'analyse numérique, Université PARIS VI, 4 place Jussieu, 75252, Paris, FRANCE.

1 Introduction

This paper is devoted to the presentation of a new family of high order numerical schemes in space (we restrict the scope of this paper to first order in time) for the numerical solution of Euler equations in axisymmetric geometry (1) (see also [Des98a] in 1D and [Des98b]). The unknowns are the density, the two components of the velocity and the specific total energy (ρ, u_1, u_2, e).

$$\left\{\begin{array}{l} \partial_t(y^r\rho) + \partial_x(y^r\rho u_1) + \partial_y(y^r\rho u_2) = 0, \\ \partial_t(y^r\rho u_1) + \partial_x(y^r\rho u_1^2 + y^r p) + \partial_y(y^r\rho u_1 u_2) = 0, \\ \partial_t(y^r\rho u_2) + \partial_x(y^r\rho u_1 u_2) + \partial_y(y^r\rho u_2^2 + y^r p) = r y^{r-1} p, \\ \partial_t(y^r\rho e) + \partial_x(y^r\rho u_1 e + y^r p u_1) + \partial_y(y^r\rho u_2 e + y^r p u_2) = 0. \end{array}\right. \tag{1}$$

For simplicity, the domain is a square $(x, y) \in \Omega =]0,1[\times]0,1[$, and the pressure provided by a γ-law $p = (\gamma - 1)(\rho e - \rho\frac{u_1^2+u_2^2}{2})$. The coefficient r may take essentially two different values. If $r = 0$ we recover the classical Euler equations in 2D written in plane geometry. If $r = 1$ we obtain the Euler equations in axisymmetric geometry : it corresponds to Euler equations in 3D assuming invariance of the flow around the axis of revolution. Note that the equation on u_2 is then non conservative.

Solving "real problems" in plasma physics requires most of the time to solve the axisymmetric case $r = 1$. The problem is then to handle a singularity which is located near the axis $y = 0$. This singularity may have a dramatic effect on the quality of the computations. It appears to be an important issue in an ICF (Inertial Confinement Fusion) context. In this contribution we show that a Lagrange+Remap approach combined with a Discontinuous Galerkin Method (DGM) [CS91]-[AS96] is a good solution.

2 Quadrature formulas and compatibility condition

It is standard in DGM to replace exact integration by quadrature formulas. These quadrature formulas are exact for a discrete basis of polynomials characterized by the Gauss quadrature points ζ_j, $j = 1, ..., J$. If $P(x, y)$ is a given

polynomial which takes the value 0 at every Gauss point $\zeta_{j' \neq j}$ and the value 1 at one Gauss point ζ_j, we write

$$P(x,y) = P_j(x,y), \quad P_j(\zeta_{j'}) = \delta_{jj'}. \tag{2}$$

Given an elementary reference square cell $\Theta =]x_-, x_+[\times]y_-, y_+[$, one question is to choose the best location of the Gauss points ζ_j such that the discrete sum with positive weights approaches the exact integral with maximal accuracy

$$\int_\Theta f(x,y) y^r \, dxdy = \sum_j \omega_j f(\zeta_j) + \text{ small residual}, \quad \text{and } \omega_j > 0.$$

The residual is as small as possible for a smooth function f and equal to 0 for a characteristic polynomial $P_{j'}$

$$\int_\Theta P_{j'}(x,y) y^r \, dxdy = \sum_j \omega_j P_{j'}(\zeta_j). \tag{3}$$

In term of the position of the Gauss points, many choices are possible for (3). We add to (3) a constraint for all characteristic polynomials $P_{j'}$

$$\int_\Theta \nabla(y^r P_{j'}(x)) dxdy = \int_{\partial\Theta} P_{j'}(x) \nu y^r \, d\sigma = \sum_j z_j P_{j'}(\zeta_j) \nu_j, \tag{4}$$

ν being the outgoing normal form Θ and ν_j the outgoing normal at point ζ_j. This condition (4) expresses a kind of compatibility between the exact integration on the boundary of Θ and the discrete integration. It is in some sense the boundary counterpart of (3), ω_j and z_j being the weights in both quadrature formulas. This relation (4) is used in the proof of theorem 1 (section 3). Be careful that in (4), the notation may be abusive. It is the case when a Gauss point $\zeta_{j'}$ is located at the corner of the reference cell Θ : in this case the order of multiplicity of j in (4) is two ; one for the horizontal normal and one for the vertical one.

Given a general arbitrary set of Gauss points ζ_j in the reference cell Θ, it may be a set of characteristic polynomials $P_j(x,y)$ satisfies (2-3) but does not satisfy (4). On the other hand it is straightforward to verify that (2-3) and (4) are true in the following cases

- one Gauss point located at the center of the cell. The characteristic polynomial is $P_1(x,y,) = 1$. We will recover a classical one order scheme.
- four Gauss points located at the corners of the reference cell Θ. The characteristic polynomials are $P_{j=1,\ldots,4}(x,y) = \epsilon_j \frac{(x_\pm + x)(y_\pm + y)}{\Delta x \Delta y}$, with $|\epsilon_j| = 1$.
- four Gauss points located at the corners of the reference cell Θ, plus one Gauss point located at the center of the cell. The fifth characteristic polynomial is $P_5(x,y) = (1 - \frac{2x}{\Delta x}^2)(1 - \frac{2y}{\Delta y}^2)$, while $P_{j=1,\ldots,4}(x,y) = \epsilon_j \frac{(x_\pm + x)(y_\pm + y)}{\Delta x \Delta y} - \frac{1}{4} P_5(x,y)$, with $|\epsilon_j| = 1$.

3 High order scheme

The domain Ω is split in many square cells Ω_k. The value of an unknown f at step n, in the cell Ω_k and at location ζ_{kj} is denoted as f^n_{kj}. In order to define the increment of unknowns, we consider a 2 parts scheme. A Lagrange part where we solve

$$\begin{cases} y^r \rho \, \mathrm{D}_t \tau = \operatorname{div} y^r u, & \mathrm{D}_t = \partial_t + u.\nabla, \\ y^r \rho \, \mathrm{D}_t u = -y^r \operatorname{grad} p, & u = (u_1, u_2), \\ y^r \rho \, \mathrm{D}_t e = -\operatorname{div}(y^r p u), & \end{cases} \tag{5}$$

and a remap part (i.e. transport part) where we solve

$$\begin{cases} \partial_t \rho y^r + \operatorname{div} y^r \rho \tilde{u} = 0, \\ \partial_t \rho y^r u + \operatorname{div} y^r \rho u \otimes \tilde{u} = 0, \\ \partial_t \rho y^r e + \operatorname{div} y^r \rho e \tilde{u} = 0, \end{cases} \tag{6}$$

where $\tilde{u} = u$ is known, given by the Lagrange part.

3.1 Lagrange part of the scheme

Using DGM one obtains from (5)

$$\begin{cases} \left(\int_{\Omega_k} \rho(\partial_t \tau)\tilde{\tau} y^r \, dxdy\right) + \left[\int_{\Omega_k} u.\operatorname{grad} \tilde{\tau} y^r \, dxdy\right] + A_1 = \int_{\partial\Omega_k} (u.\nu_k)\tilde{\tau} y^r \, d\sigma \\ \left(\int_{\Omega_k} \rho(\partial_t u).\tilde{u} y^r \, dxdy\right) - \left[\int_{\Omega_k} p \operatorname{div} \tilde{u} y^r \, dxdy\right] + A_2 = -\int_{\partial\Omega_k} p(\tilde{u}.\nu_k) y^r \, d\sigma \\ \left(\int_{\Omega_k} \rho(\partial_t e)\tilde{e} y^r \, dxdy\right) - \left[\int_{\Omega_k} p \operatorname{div}(u\tilde{e}) y^r \, dxdy + \int_{\Omega_k} u.\operatorname{grad}(p\tilde{e}) y^r \, dxdy\right] \\ +A_3 = -\int_{\partial\Omega_k} p(u.\nu_k)\tilde{e} y^r \, d\sigma, \qquad \tilde{\tau}, \tilde{u}, \tilde{e} \text{ being test functions.} \end{cases} \tag{7}$$

The terms $A_{1,2,3}$ are some approximations of 0, which corresponds to some degrees of freedom we allow in the numerical approximations of (5).

We now use the standard procedure in order to get from (7) a discrete set of equations. The integral (...) are discretized using the explicit value of the density ρ^n_{kj} at the beginning of the time step, the quadrature weights ω_{kj} and explicit differentiation of the derivatives. The integral [...] are discretized using the implicit value of the unknowns. The right hand sides are discretized using the Riemann-solver like formulas

$$\begin{cases} p^*_{kj} = \dfrac{(p^n_{kj} + (\rho^* c^*)_{kj}(u^n_{kj}.\nu_{kj}) + (p^n_{k'j'} + (\rho^* c^*)_{kj}(u^n_{k'j'}.\nu_{k'j'})}{2} \\ (u^*_{kj}.\nu_{kj}) = \dfrac{(p^n_{kj} + (\rho^* c^*)_{kj}(u^n_{kj}.\nu_{kj}) - (p^n_{k'j'} + (\rho^* c^*)_{kj}(u^n_{k'j'}.\nu_{k'j'})}{2(\rho^* c^*)_{kj}} \end{cases}, \tag{8}$$

$(\rho^* c^*)_{kj}$ being some local approximation of the density times the sound velocity, and the quadrature formulas on the boundaries (4). It remains to give the value of $A_{1,2,3}$. These approximations of 0 are chosen in order that both

the conservativity of (5) is satisfied and that theorem 1 holds. It is the case for this choice

$$\left\{\begin{array}{l} A_1 = -\frac{1}{2}\int_{\partial\Omega_k}(u_k^{n+\frac{1}{2}}.\nu_k)\tilde{\tau}_k y^r\,d\sigma + \frac{1}{2}\sum_j z_{kj}(u_{kj}^{n+\frac{1}{2}}.\nu_{kj})\tilde{\tau}_{kj} \\ A_2 = \frac{1}{2}\int_{\partial\Omega_k} p_k^{n+\frac{1}{2}}(\tilde{u}_k.\nu_k)y^r\,d\sigma - \frac{1}{2}\sum_j z_{kj}p_{kj}^{n+\frac{1}{2}}(\tilde{u}_{kj}.\nu_{kj}) \\ A_3 = \frac{1}{2}\int_{\partial\Omega_k} p_k^{n+\frac{1}{2}}(\Pi_k(u_{kj}^{n+\frac{1}{2}}\tilde{e}_{kj}).\nu_k)y^r\,d\sigma - \frac{1}{2}\sum_j z_{kj}(u_{kj}^{n+\frac{1}{2}}.\nu_{kj})p_{kj}^{n+\frac{1}{2}}\tilde{e}_{kj} \\ +\frac{1}{2}\int_{\partial\Omega_k}\Pi_k(p_{kj}^{n+\frac{1}{2}}\tilde{e}_{kj})(u_k^{n+\frac{1}{2}}.\nu_k)y^r\,d\sigma - \frac{1}{2}\sum_j z_{kj}(u_{kj}^{n+\frac{1}{2}}.\nu_{kj})p_{kj}^{n+\frac{1}{2}}\tilde{e}_{kj} \end{array}\right.$$

With all these ingredients the reduced Lagrangian discrete system is

$$\left\{\begin{array}{l} w_{kj}\rho_{kj}^n\frac{\tau_{kj}^{n+\frac{1}{2}}-\tau_{kj}^n}{\Delta t} - \sum_{j'}\alpha_{kjj'}u_{kj'}^{n+\frac{1}{2}} - z_{kj}(u_{kj}^*.\nu_{kj}) = 0 \\ w_{kj}\rho_{kj}^n\frac{u_{kj}^{n+\frac{1}{2}}-u_{kj}^n}{\Delta t} + \sum_{j'}\beta_{kjj'}p_{kj'}^{n+\frac{1}{2}} + z_{kj}p_{kj}^*\nu_{kj} = 0 \\ w_{kj}\rho_{kj}^n\frac{e_{kj}^{n+\frac{1}{2}}-e_{kj}^n}{\Delta t} + (\sum_{j'\neq j}\beta_{kjj'}p_{kj'}^{n+\frac{1}{2}})u_{kj}^{n+\frac{1}{2}} \\ +(\sum_{j'\neq j}\alpha_{kjj'}.u_{kj'}^{n+\frac{1}{2}})p_{kj}^{n+\frac{1}{2}} + z_{kj}p_{kj}^*(u_{kj}^*.\nu_{kj}) = 0 \end{array}\right. \tag{9}$$

$$\alpha_{kj} = \int_{\Omega_k}(P_{k,j'}\nabla P_{k,j})\,y^r\,dxdy,\quad \beta_{kj} = \int_{\Omega_k} P_{k,j'}\nabla(y^r P_{k,j})\,dxdy.$$

It is clear that this system is an implicit discrete system, local in each cell. It is possible to solve it using a Newton like algorithm in each cell. The system (9) satisfies the following remarkable property

Theorem 1. *For any cell Ω_k and for any Gauss-point ζ_j (i.e. for any (k,j)), there exists a constant $c_{kj}^n > 0$ such that if the CFL condition $c_{kj}\frac{\Delta t}{\Delta x} \leq 1$ is true then the entropy increases locally at each Gauss point : $S_{kj}^{n+\frac{1}{2}} \geq S_{kj}^n$. Without any restriction on the time step, the conservative relations (see (9)) are satisfied*

$$\sum_k\sum_j w_{kj}\rho_{kj}^n f_{kj}^{n+\frac{1}{2}} = \sum_k\sum_j w_{kj}\rho_{kj}^n f_{kj}^n,\quad f = \tau, u, e.$$

The proof of these various properties rely on the very particular expression chosen for the approximations of zero A_1, A_2, A_3 and on (4). In practical computations, we never use the fully implicit formulation. We prefer to use a predicted value of the pressure $\bar{p}_{kj}^{n+\frac{1}{2}} = p_{kj} + \Delta t(\frac{dp}{dt})_{kj}^n \approx p_{kj}^{n+\frac{1}{2}}$, where the derivative of the pressure is evaluated explicitly. So doing the local (at each point ζ_{kj}) increase of the entropy is no more true, however we still get a local (at each cell Ω_k) increase $\sum_j \omega_{kj}\rho_{kj}^n S_{kj}^{n+\frac{1}{2}} \geq \sum_j \omega_{kj}\rho_{kj}^n S_{kj}^n$ under CFL. In some sense the increase of the entropy is equivalent to some non linear stability estimate in the Lagrange step.

3.2 Remap part of the scheme

As quoted previously, we base the remap stage on the discretisation of (6). We obtain using the DGM

$$\begin{cases} \sum_j w_{kj} \frac{\rho_{kj}^{n+1}-\rho_{kj}^n}{\Delta t}\tilde{\rho}_{kj} & -\int_{\Omega_k} \rho_k^{n+1} u_k^{n+\frac{1}{2}}.\operatorname{grad}\tilde{\rho}_k y^r\,dxdy \\ & = -\sum_j z_{kj}\rho_{kj,a}^{n+\frac{1}{2}}(u_{kj}^*.\nu_{kj})\tilde{\rho}_{kj} \\ \sum_j w_{kj} \frac{\rho_{kj}^{n+1}u_{kj}^{n+1}-\rho_{kj}^n u_{kj}^{n+\frac{1}{2}}}{\Delta t}.\tilde{u}_{kj} & -\int_{\Omega_k}(\rho_k^{n+1}u_k^{n+1}\otimes u_k^{n+\frac{1}{2}}).\operatorname{grad}\tilde{u}_k y^r\,dxdy \\ & = -\sum_j z_{kj}((\rho u)_{kj,a}^{n+\frac{1}{2}}(u_{kj}^*.\nu_{kj})\tilde{u}_{kj}) \\ \sum_j w_{kj} \frac{\rho_{kj}^{n+1}e_{kj}^{n+1}-\rho_{kj}^n e_{kj}^{n+\frac{1}{2}}}{\Delta t}\tilde{e}_{kj} & -\int_{\Omega_k}(\rho_k^{n+1}e_k^{n+1}u_k^{n+\frac{1}{2}}).\operatorname{grad}\tilde{e}_k y^r\,dxdy \\ & = -\sum_j z_{kj}(\rho e)_{kj,a}^{n+\frac{1}{2}}(u_{kj}^*.\nu_{kj})\tilde{e}_{kj} \end{cases} \quad (10)$$

Here the fluxes on the boundaries are evaluated with upwinding, while as before the integral inside each cell is evaluated with implicit values. This step is L^2 stable.

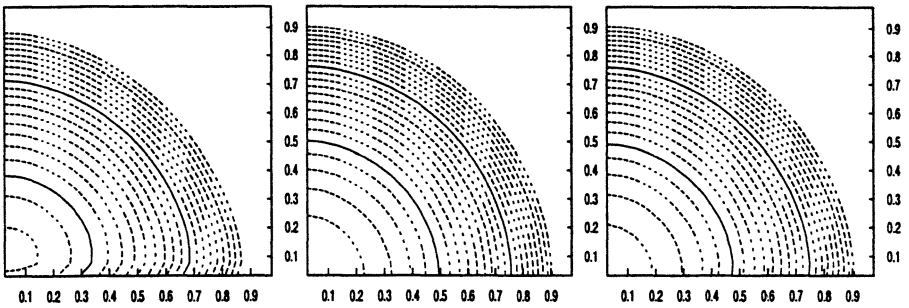
4 Numerical results

We rely on various test cases in order to evaluate the interest of the scheme. All experiments show that the Lagrange step is extremely robust. The oscillations of the scheme are very small in the Lagrange step : the positivity of the density and the internal energy is insured without any limiter. However it is sometimes preferable to add a positivity constraint in the remap step.

We test the robustness of a scheme with a very strong isentropic convergent flow. It is based on an analytic solution of the Euler equation given by Kidder [Kid76]. The analytical solution is self-similar for $\gamma = \frac{5}{3}$. The solution at (r,t) is related to the solution at $(R, t=0)$ through the transformation $r = R\sqrt{1-\frac{t^2}{\tau^2}}$, this transformation is defined for t smaller than the focusing time $0 \le t < \tau$. The initial conditions are

$$\begin{cases} \rho(r,0) = \left(\rho_2^{\gamma-1}\frac{r^2-R_1^2}{R_2^2-R_1^2} + \rho_1^{\gamma-1}\frac{R_2^2-r^2}{R_2^2-R_1^2}\right), \\ u(r,0) = 0, \\ \varepsilon(r,0) = \rho(r,0)^{\gamma-1}, \\ p(r,0) = (\gamma-1)\rho(r,0)^{\gamma}, \\ \tau = \sqrt{\frac{1}{2\gamma}\left(\frac{R_2^2-R_1^2}{\rho_2^{\gamma-1}-\rho_1^{\gamma-1}}\right)}, \end{cases} \quad \text{and then} \quad \begin{cases} \rho(r,t) = \frac{\rho(R)}{h(t)^3}, \\ u(r,t) = \frac{dr}{dt}, \\ \varepsilon(r,t) = \rho(r,t)^{\gamma-1}, \\ p(r,t) = \frac{\rho(R)}{h(t)^{3\gamma}}. \end{cases}$$

This analytical solution is symmetric around the axis of revolution (the left bottom corner of the picture). The problem stressed by this test case is that in general the discrete solution is no more symmetric. In numerical experiments we take $(\rho_1, \rho_2, R_1, R_2) = (1, 2, 0, 1)$. The time is $T = .95 \times \tau$, with 20×20 cells. In following figures, the isolines of the density are plotted for the



one Gauss point four Gauss points five Gauss points

Kidder problem. The numerical artifact on the axis with the one order scheme can not be reduced with an increase of the number of cells. This loss of symmetry may have a disastrous effect for "real" problems. It is the case for ICF computations for instance, where it is preferable to preserve the symmetry in order to control hydrodynamics instabilities. It is only with higher order scheme that this artifact may be reduced. It is clear that the use of DGM with ($\geq$) four Gauss points improves a lot the quality of the computation, regarding to this criterion. On this very particular test case the five points scheme seems to give poorer result than the four points scheme, but still much better than the one point scheme.

Through numerical experiments, we have computed that the error follows more or less the law $||U - U_{\Delta x}||_{L^2} \approx C_1 \Delta t + C_2 \Delta x^d$ with $d = 1$ for one Gauss point and $d = 1.7$ for four Gauss points.

References

[AS96] H.L. Atkins and C-W. Shu. Quadrature-free implementation of discontinuous galerkin method for hyperbolic equations. AIAA Journal, 36:775-782, 1998.

[CS91] B. Cockburn and C. W. Shu. The Runge-Kutta local projection P^1 discontinuous Galerkin method for scalar conservation laws. M^2AN, 25:337–361, 1991.

[Des98a] B. Després. Entropy inequality for high order discontinuous Galerkin approximation of Euler equations. In *VII conference on hyperbolic problems.* ETHZ-Zurich, 1998.

[Des98b] B. Després. *Systèmes hyperboliques: nouveaux schémas et nouvelles applications*, chapter : Inégalités entropiques pour un solveur de type Lagrange+convection des équations de l'hydrodynamique. INRIA-France, 1998.

[Kid76] R. E. Kidder. Laser-driven compression of hollow shells : power requirements and stability limitations. *Nuclear Fusion*, 16(1):3–14, 1976.

Ten Years Using Discontinuous Galerkin Methods for Polymer Processing Problems

A. Fortin, A. Béliveau, M.C. Heuzey and A. Lioret

Département de Mathématiques et de Génie Industriel
École Polytechnique de Montréal
C.P. 6079, Succursale Centre-Ville, Montréal, Canada, H3C 3A7

1 Introduction

The numerical simulation of polymer processing problems requires important computer ressources making essential the development of very efficient numerical techniques. Among the difficulties, the viscoelastic nature of polymers is the most important. Nonlinear multi-mode differential models exist to describe their behaviour, but their hyperbolic (convective) nature makes numerical simulations quite difficult. The presence of free surfaces in problems such as injection molding process, extrusion die swell, coextrusion, etc. further enhance the complexity.

The discontinuous Galerkin method (also known as the Lesaint-Raviart method [LR]) allows the solution of convective equations in a very efficient manner. We have introduced this method in this field in 1989 (see Fortin-Fortin [FF]). Since then, it has been succesfully used thoughout the world.

2 Description of the problem

Let $\boldsymbol{\sigma}$ denote the Cauchy stress-tensor defined as $\boldsymbol{\sigma} = -p\boldsymbol{I} + 2\eta_s\dot{\boldsymbol{\gamma}}(\boldsymbol{u}) + \boldsymbol{\tau} = -p\boldsymbol{I} + 2\eta_s\dot{\boldsymbol{\gamma}}(\boldsymbol{u}) + \sum_{i=1}^{n}\boldsymbol{\tau}_i$, where p is the pressure, η_s a shear viscosity called "solvent" viscosity, $\boldsymbol{u}$ is the velocity field and $\dot{\boldsymbol{\gamma}}(\boldsymbol{u})$ the rate of strain tensor. The extra-stress tensor $\boldsymbol{\tau}$ has been decomposed into n modes $\boldsymbol{\tau}_i$ so that the characterization of the polymer can be more accurate. A modified elastic-viscous stress splitting (DEVSS) method [GF] is introduced to ease the choice of the discretization spaces. This requires the explicit introduction (and discretization) of a tensor variable $\boldsymbol{d}$ which is nothing but the gradient tensor:

$$\boldsymbol{d} = \nabla\boldsymbol{u} \qquad \left(\text{so that } \dot{\boldsymbol{\gamma}}(\boldsymbol{u}) = \frac{(\boldsymbol{d}+\boldsymbol{d}^t)}{2}\right) \tag{1}$$

The fluid is assumed to be incompressible and inertia forces are neglected. The conservation equations are then expressed as :

$$-\nabla\cdot(2(\eta_s+\alpha)\dot{\boldsymbol{\gamma}}(\boldsymbol{u})) + \nabla p - \nabla\cdot\boldsymbol{\tau} = -\nabla\cdot\left(2\alpha\left(\frac{\boldsymbol{d}+\boldsymbol{d}^t)}{2}\right)\right) + \boldsymbol{f} \tag{2}$$

$$\nabla \cdot \boldsymbol{u} = 0 \tag{3}$$

The vector $\boldsymbol{f}$ stands for mass forces usually neglected. Note that in the conservation of momentum equation 2, the α-terms on both sides cancel out. However, in the discrete problem, the α-terms will have different discretization and the slight difference between these terms has a stabilization effect on the discretization. The parameter α is *a priori* arbitrary but if properly chosen, it will improve the overall convergence of the algorithm (see [FGP]).

To close the above system, we introduce a constitutive law that relates the velocity field to the extra-stress tensor. We have selected the Phan-Thien and Tanner (PTT) model [PTT] where each mode $\boldsymbol{\tau}_i$ of the viscoelastic extra-stress tensor $\boldsymbol{\tau}$ is a solution of:

$$\lambda_i \left(\frac{\partial \boldsymbol{\tau}_i}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{\tau}_i - \nabla \boldsymbol{u} \cdot \boldsymbol{\tau}_i - \boldsymbol{\tau}_i \cdot \nabla \boldsymbol{u}^t \right) + \left(1 + \frac{\epsilon \lambda_i}{\eta_{v_i}} tr(\boldsymbol{\tau}_i) \right) \boldsymbol{\tau}_i = 2\eta_{v_i} \dot{\boldsymbol{\gamma}}(\boldsymbol{u}) \tag{4}$$

In equation (4), λ_i and η_{v_i} are the relaxation time and the corresponding viscosity of the i^{th} mode while ϵ is a characteristic parameter (fluid dependent). The Newtonian case is obtained by using only one mode ($n = 1$) and by setting $\lambda = 0$.

In presence of free surfaces, a pseudo-concentration method is used requiring the solution of the extra equation:

$$\frac{\partial F}{\partial t} + \boldsymbol{u} \cdot \nabla F = 0 \tag{5}$$

3 Discretization

3.1 The four-field Stokes problem

The discretization of the different variables is not a simple task since compatibility conditions exists. Indeed, even in the one mode Newtonian case, system 1-4 is a four-field Stokes problem for the discretized variables $\boldsymbol{u}_h, p_h, \boldsymbol{\tau}_h$ and $\boldsymbol{d}_h$. Their respective discretization must therefore satisfy a generalization of Brezzi's condition [DBF] for the usual velocity-pressure formulation of the Stokes problem. For a complete discussion, we refer the reader to Fortin-Guénette-Pierre [FGP]. The chosen discretization is illustrated in figure 1.

When a free surface is present, system 1-4 is completed with equation 5. The variable F_h is approximated by piecewise linear polynomials as for τ_h and d_h. Note the similarity between the constitutive equation 4 and the pseudo-concentration equation 5. Both these equations are suitable for the discontinuous Galerkin method which is now briefly described.

3.2 The discontinuous Galerkin method

We will only describe the method for the pseudo-concentration keeping in mind that a similar treatment is applied to equation 4. We consider the more

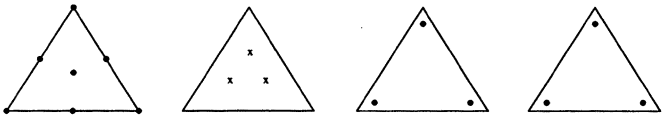


Figure1. *The P_2^+-P_1-P_1-P_1 for $\boldsymbol{u}_h, p_h, \boldsymbol{\tau}_h$ and $\boldsymbol{d}_h$.*

general time-dependent case. A fully implicit second order Gear scheme is used for the time derivative. In stationary problems, the time derivative is simply removed.

We suppose that the velocity field $\boldsymbol{u}_h$ is known (from a previous iteration for example) and we define the inflow boundary ∂K^- with respect to the velocity field as:

$$\partial K^- = \{(x_1, x_2) \in \partial K | \boldsymbol{u} \cdot \boldsymbol{n}_K(x_1, x_2) < 0\}$$

where $\boldsymbol{n}_K$ is the outward unit normal vector to the boundary ∂K.

$$\int_K \left(\frac{3F_h - 4F_h^{n-1} + F_h^{n-2}}{2\Delta t} + \boldsymbol{u} \cdot \nabla F_h \right) G \, dx = \int_{\partial K^-} \boldsymbol{u}_h \cdot \boldsymbol{n}_K (F_h - F_h^-) G \, ds$$

where $F_h^n = F_h$ is the solution at time n (droping the superscript n), F_h^- is the value of F_h at time t_n in the element adjacent to the inflow boundary and F_h^{n-1} and F_h^{n-2} are the values of F_h at previous time steps. This results in a small linear system on each element K. However, the resolution of this system requires the knowledge of the quantity F_h^- on elements adjacent to ∂K^- so that a particular numbering of the elements is necessary. A perfect numbering is not always possible. In presence of recirculation zones for example, the best possible numbering is provided and the elements are swept many times so that the resolution can be seen as a block relaxation method.

Similarly, for each mode τ_i in the constitutive equation, we have to solve the following non linear system of size 9×9 in 2D, 12×12 for axisymmetric problem, and this on each element:

$$\begin{aligned}
&\int_K \left\{ \frac{3\boldsymbol{\tau}_{i_h}}{2\Delta t} + \lambda_i \left(\boldsymbol{u} \cdot \nabla \boldsymbol{\tau}_{i_h} - \left(\nabla \boldsymbol{u}_h \cdot \boldsymbol{\tau}_{i_h} + \boldsymbol{\tau}_{i_h} \cdot \nabla \boldsymbol{u}_h^t \right) \right) \right\} : \boldsymbol{\Psi}_h dx \\
&\quad + \int_K \left\{ \left(1 + \frac{\epsilon \lambda_i}{\eta_{v_i}} tr(\boldsymbol{\tau}_{i_h}) \right) \boldsymbol{\tau}_{i_h} \right\} : \boldsymbol{\Psi}_h dx - \lambda_i \int_{\partial K^-} \boldsymbol{u}_h \cdot \boldsymbol{n}_K \boldsymbol{\tau}_{i_h} : \boldsymbol{\Psi}_h ds \\
&= \int_K \left(\frac{4\boldsymbol{\tau}_{i_h}^{n-1} - \boldsymbol{\tau}_{i_h}^{n-2}}{2\Delta t} + 2\eta_{v_i} \dot{\boldsymbol{\gamma}}(\boldsymbol{u}_h) \right) : \boldsymbol{\Psi}_h dx - \lambda_i \int_{\partial K^-} \boldsymbol{u}_h \cdot \boldsymbol{n}_K \boldsymbol{\tau}_{i_h}^- : \boldsymbol{\Psi}_h ds
\end{aligned}$$

Each mode τ_{i_h} is computed separately. A Newton method is used to linearize the system on each element.

3.3 Solution of the global system

System 1-4 can be very large. In the next section, numerical results will be presented with up to 6 modes in the constitutive equation so the number of degrees of freedom increases very quickly, even for two-dimensional problems. It thus prohibits the Newton-Raphson method as a solver because it would require the construction of a huge Jacobian matrix.

We have thus introduced Newton-Krylov iterative schemes for the solution of the global system (see Fortin-Zine-Agassant [FZA]). The idea is to use the Generalized Minimal Residual (GMRES) method to solve the linear system in the Newton method. Iterative methods do not require the Jacobian matrix itself but only its product $J\boldsymbol{d}$ with various descent directions $\boldsymbol{d}$. This product is then approximated by a finite difference, avoiding the explicit construction of J.

4 Numerical results

4.1 Stationary fluid flow

The first application is a classical benchmark problem for viscoelastic fluid flow problems. The geometry is a planar 8 to 1 contraction. Meshes with 1277 to 2976 elements were used to assess the convergence with mesh size. Moreover, birefringence fringes were obtained in order to make comparison between numerical and experimental results. A high density polyethylene was characterized and the resulting spectrum had 6 modes which were used in the numerical simulations. Figure 2 provides the comparison for an apparent wall shear rate of 48s^{-1}, showing good qualitative agreement with experimental results.

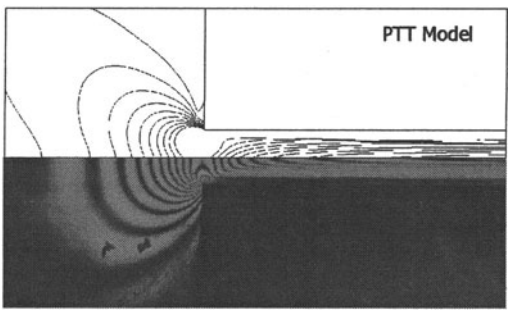


Figure2. *Comparison between experimental and numerical results* ($\gamma_{a_w} = 48s^{-1}$).

4.2 Die swell

This is a first example with a free surface. Here again, it is a classical problem typical of the behaviour of viscoelastic fluid. In this case, the geometry is axisymmetric and is represented in figure 3. A polymer is extruded at the exit of a die and tends to swell. For Newtonian fluid, the swell ratio is between 12 to 13%. However, for polymer, the swelling ratio can increase up to 300%.

The pseudo-concentration F allows the computation of the interface on a fixed mesh. Figure 3 shows a 21% swelling for an Oldroyd B ($\epsilon = 0, \eta_s/(\eta_s + \eta_v) = 1/9$) fluid at $We = 1$.

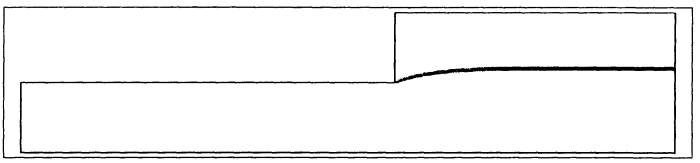


Figure3. *Pseudo-concentration at* $We = 1$.

4.3 Deformation of droplets

This is an example of a time-dependent free surface problems (see Béliveau [DBF]) for more details). A circular cylindrical drop of radius R_d is placed inside a two dimensional rectangular geometry and is subjected to a shearing velocity field of the form described in figure 4. The parameter $\dot{\gamma}$ is a shearing coefficient. The capillary number Ca is defined by $Ca = \frac{\dot{\gamma}\eta_m R_d}{C}$ where η_m is

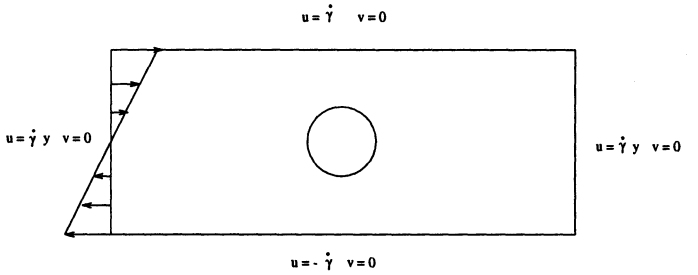


Figure4. *Initial geometry and boundary conditions.*

the viscosity of the surrounding fluid and C is the surface tension coefficient.

When the capillary number is greater than a critical value, the drop breaks more or less rapidly into smaller droplets. Figure 5 illustrates the case where Ca is slightly larger than the critical value.

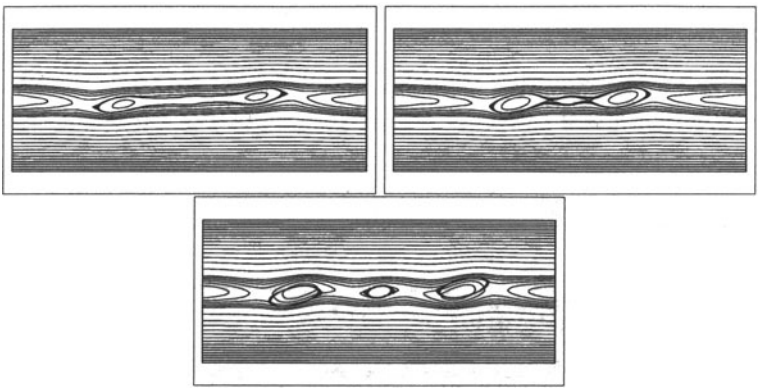


Figure5. *Interface* $F = 1/2$ *and streamlines at* $t = 8$, $t = 17$ *and* $t = 25$.

5 Conclusion

In this paper, we have briefly discussed the numerical strategy and presented numerical results in various problems showing the potential of the proposed method based on the Discontinuous Galerkin method.

References

[PTT] Phan-Thien N., Tanner R.I., Journal for Non-Newtonian Fluid Mechanics, **2**, p.353, (1977).

[FF] Fortin M., Fortin A., *A New Approach for the FEM Simulation of Viscoelastic Flows*, Journal for Non-Newtonian Fluid Mechanics, JNNFM, **32**, pp.295-310, (1989).

[FZA] Fortin A., Zine A., Agassant J.F., *Computing Viscoelastic Fluid Flow at Low Cost*, JNNFM, 45, 209-229, (1992).

[DBF] Demay Y., Béliveau A., Fortin A., *A Numerical Method for the Deformation of Two-dimensional Drops with Surface Tension*, Int. J. Comp. Fluid Dynamics, Vol.10, 225-240, (1998).

[LR] Lesaint P., Raviart P.A., C. de Boor, Academic Press, p.89, (1974).

[GF] Guénette, R., Fortin M., *Numerical Analysis of the Modified EVSS Method*, JNNFM, **60**, (1995).

[FGP] Fortin A., Guénette R., Pierre R., *On the discrete EVSS Method*, submitted to Comp. Meth. Appl. Mech. Eng., (1999).

Using Krylov-Subspace Iterations in Discontinuous Galerkin Methods for Nonlinear Reaction-Diffusion Systems

Donald J. Estep[1] and Roland W. Freund[2]

[1] School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, USA

[2] Bell Laboratories, Room 2C–420, 700 Mountain Avenue, Murray Hill, NJ 07974–0636, USA

Abstract. We consider discontinuous in time and continuous in space Galerkin finite-element methods for the numerical solution of reaction-diffusion differential equations. These are implicit methods that require the solution of a system of nonlinear equations at each time node. In this paper, we explore the use of Krylov-subspace techniques for the iterative solution of the linear systems that arise when these nonlinear systems are solved by means of Newton-type methods. It is shown how these linear systems depend on the choice of the basis functions used for the time discretization. We demonstrate that Krylov-subspace methods can be sped up considerably by employing an orthogonal basis for the time discretization and by combining the Krylov iteration with a suitable block preconditioner. Results of numerical experiments are reported.

1 Introduction

Many fundamental models in science take the form of time-dependent nonlinear reaction-diffusion differential equations. Examples include the modeling of domain walls in ferromagnetic materials, predator-prey models, the description of superconductivity in liquids, the modeling of the famous Belousov-Zhabotinsky reaction in chemical kinetics, the modeling of flame propagation, and the modeling of the spread of rabies in foxes. As well as being important for physical modeling purposes, solutions of reaction-diffusion equations can also exhibit complex and beautiful behavior arising primarily from the competition between reaction and diffusion and the nonlinear nature of the equations. However, the numerical solution of reaction-diffusion equations is correspondingly difficult.

In this paper, we consider systems of D reaction-diffusion equations consisting of d, $1 \leq d \leq D$, parabolic equations and $D - d$ ordinary equations for the $\mathbb{R}^D$-valued function $\mathbf{u} = [u_i]_{1 \leq i \leq D}$:

$$\begin{aligned} \frac{\partial u_i}{\partial t} - \nabla \cdot (\epsilon_i(\mathbf{u}, \mathbf{x}, t) \nabla u_i) &= f_i(\mathbf{u}, \mathbf{x}, t), \quad (\mathbf{x}, t) \in \Omega \times \mathbb{R}^+, \ 1 \leq i \leq D, \\ u_i(\mathbf{x}, t) &= 0, \quad (\mathbf{x}, t) \in \partial\Omega \times \mathbb{R}^+, \ 1 \leq i \leq d, \\ \mathbf{u}(\mathbf{x}, 0) &= \mathbf{u}_0(\mathbf{x}), \quad \mathbf{x} \in \Omega. \end{aligned} \tag{1}$$

Here, Ω is a convex polygonal domain in $\mathbb{R}^N$, where $N \in \{1,2,3\}$, with boundary $\partial\Omega$. We assume that there is a constant $\epsilon > 0$ such that

$$\epsilon_i(\mathbf{u},\mathbf{x},t) \geq \epsilon \quad \text{for} \quad 1 \leq i \leq d \quad \text{and} \quad \epsilon_i(\mathbf{u},\mathbf{x},t) \equiv 0 \quad \text{for} \quad d < i \leq D.$$

Note that, in (1), there is no parabolic smoothing in the $D-d$ "singular" equations. To compute numerical solutions of systems of the form (1), we employ the finite-element space-time discretization of (1) called the *discontinuous Galerkin method* (dG method); see, e.g., [2,3,10] and the references given there. The reason for employing space-time finite element methods is to take advantage of the new approach to computational error estimation based on residuals and variational analysis; see [3,4]. Let $q \geq 1$ be an integer. The dG(q) method uses discontinuous (in time) approximations that are piecewise polynomials of degree at most q in time and piecewise linear polynomials in space. The approximations are allowed to be discontinuous at the time nodes, but for fixed time, they are continuous in space. It is well known that dG(q) methods offer a number of advantages over their continuous counterparts. These advantages include an increased convergence order at the time nodes, better behavior for long-time integration of parabolic problems, and the easy implementation of adaptive spatial grids that are changed with time. On the other hand, dG(q) methods can be significantly more expensive than the corresponding continuous Galerkin methods, especially when no attention is paid to the linear algebra problems that actually dominate the computational costs. In this paper, we explore the use of Krylov-subspace methods for the linear algebra problems arising in discontinuous Galerkin methods.

2 Discontinuous Galerkin Discretizations

For simplicity, from now on we write $\epsilon_i(\mathbf{u}) = \epsilon_i(\mathbf{u},\mathbf{x},t)$ and $f_i(\mathbf{u}) = f_i(\mathbf{u},\mathbf{x},t)$. We partition $[0,\infty)$ as $0 = t_0 < t_1 < t_2 < \cdots < t_n < \cdots$, denoting each time interval by $I_n := (t_{n-1}, t_n]$ and time step by $k_n := t_n - t_{n-1}$. To each interval I_n, we associate a triangulation $\mathcal{T}_n$ of Ω. Note that $\mathcal{T}_n$ is allowed to change across time nodes.

The dG(q) approximations are polynomials of degree at most q in time and piecewise polynomials in space on each space-time "slab" $S_n = \Omega \times I_n$. In space, we let $\mathbf{V}_n \subset (H_0^1(\Omega))^d \times (H^1(\Omega))^{D-d}$ denote the space of piecewise linear continuous vector-valued functions $\mathbf{v}(\mathbf{x}) \in \mathbf{R}^D$ defined on $\mathcal{T}_n$, where the first d components of $\mathbf{v}$ are zero on $\partial\Omega$. Then on each slab, we define

$$\mathbf{W}_n^q := \{\, \mathbf{w}(\mathbf{x},t) \mid \mathbf{w}(\mathbf{x},t) = \sum_{j=0}^{q} t^j \mathbf{v}_j(\mathbf{x}),\ \mathbf{v}_j \in \mathbf{V}_n,\ (\mathbf{x},t) \in S_n \,\}. \quad (2)$$

Furthermore, we let $\mathbf{W}^q$ denote the space of functions defined on the space-time domain $\Omega \times \mathbf{R}^+$ such that $\mathbf{v}|_{S_n} \in \mathbf{W}_n^q$ for $n \geq 1$. Note that functions in $\mathbf{W}^q$ are generally discontinuous across the discrete time levels, and we denote

the jump across t_n by $[\mathbf{w}]_n = \mathbf{w}_n^+ - \mathbf{w}_n^-$ where $\mathbf{w}_n^\pm = \lim_{s \to t_n \pm} \mathbf{w}(s)$. Finally, $\mathbf{P}_n : L^2(\Omega) \to \mathbf{V}_n$ denotes the L^2-projection operator onto $\mathbf{V}_n$, and $(\cdot, \cdot)$ is the $L_2(\Omega)$-inner product.

The dG(q) approximation to the solution of (1) is then defined as the function $\mathbf{U} = [U_i]_{1 \le i \le D} \in \mathbf{W}^q$ that satisfies $\mathbf{U}_0^- = \mathbf{P}_0 \mathbf{u}_0$ and for $n \ge 1$,

$$\int_{t_{n-1}}^{t_n} \left((\dot{U}_i, v_i) + (\epsilon_i(\mathbf{U}) \nabla U_i, \nabla v_i) \right) dt + \left([U_i]_{n-1}, v_i^+ \right) = \int_{t_{n-1}}^{t_n} (f_i(\mathbf{U}), v_i)\, dt \quad \text{for all } \mathbf{v} = [v_i]_{1 \le i \le D} \in \mathbf{W}_n^q,\ 1 \le i \le D. \tag{3}$$

In practice, some of the integrals in (3) are computed using quadrature. In particular in space, we evaluate the integrals involving some form of mass matrix, i.e., $(\dot{U}_i, v_i)$, $\left([U_i]_{n-1}, v_i^+ \right)$, and $(f_i(\mathbf{U}), v_i)$, using the lumped mass, or composite trapezoidal rule, quadrature. The choice of this quadrature rule is dictated by stability considerations; see [4].

Equations (3) defining the dG(q) approximation result in a large sparse system of nonlinear equations that needs to be solved on each space-time slab S_n. For this task, we use an inexact Newton method combined with preconditioned Krylov-subspace iterations to obtain an approximate solution of the large sparse linear system for the Newton direction that needs to be solved at each Newton iteration. In the next section, we discuss the structure of these linear systems for the special case of the dG(1) method.

3 Structure of the Linear Systems for the Case $q = 1$

For simplicity, from now on, we assume that the ϵ_i's in (1) are all constant, and that the triangulation $\mathcal{T} = \mathcal{T}_n$ of Ω is the same for all time nodes. The usual nodal basis associated with $\mathcal{T}$ is used for the spatial functions in the ansatz space (2). In the sequel, we consider only the dG(1) method. Extensions to dG(q) methods with $q \ge 2$ will be described elsewhere.

For each space-time slab S_n, we need to solve the nonlinear system

$$\mathbf{F}_n(\mathbf{U}) = \mathbf{0}, \quad \text{where} \quad \mathbf{U} = \begin{bmatrix} \mathbf{U}_n^- \\ \mathbf{U}_{n-1}^+ \end{bmatrix}, \tag{4}$$

for the approximate solutions $\mathbf{U}_n^-$ and $\mathbf{U}_{n-1}^+$ corresponding to the right and left end point of the interval $I_n = (t_{n-1}, t_n]$. Note that (4) is a system of $2Dm$ equations for a total of $2Dm$ unknowns, where m denotes the the number of grid points in the triangulation $\mathcal{T}$ of Ω.

Each step of a Newton method applied to (4) requires the solution of a linear system with the Jacobian matrix $\mathbf{J} = D\mathbf{F}_n(\mathbf{U})$ of $\mathbf{F}_n(\mathbf{U})$. It turns out that the structure of $\mathbf{J}$ depends on the choice of the basis function for the time discretization in the ansatz space (2). Recall that for dG(1), the approximation is linear on each time interval I_n so there are two basis functions for each I_n.

First, we consider the usual nodal basis functions that are 1 at one end point and 0 at the other end point. In this case, the Jacobian matrix $\mathbf{J}$ is of the form

$$\mathbf{J}_{\mathcal{N}} = \frac{1}{2}\begin{bmatrix} \mathbf{B} & -\mathbf{B} \\ \mathbf{B} & \mathbf{B} \end{bmatrix} + \frac{k_n}{3}\begin{bmatrix} \epsilon\,\mathbf{A} & \epsilon\,\mathbf{A}/2 \\ \epsilon\,\mathbf{A}/2 & \epsilon\,\mathbf{A} \end{bmatrix} - k_n\,\mathbf{R}_{\mathcal{N}}. \tag{5}$$

Here, $\mathbf{A}$ and $\mathbf{B}$ are the stiffness and mass matrix associated with the spatial discretization, and $\epsilon := \mathrm{diag}(\epsilon_1, \epsilon_2, \ldots, \epsilon_D)$. Note that, in (5), the first term corresponds to the time-derivative part in (1), the second term to the diffusion part in (1), and the third term to the reaction part in (1). In practice, mass lumping is used, so that $\mathbf{B}$ is a diagonal matrix. Moreover, for equidistant regular spatial grids, $\mathbf{B} = \mathbf{I}$ is the identity matrix. The matrix $\mathbf{R}_{\mathcal{N}}$ has nonzero elements only in a total of $4D - 1$ diagonals.

A second standard choice for the time basis functions is an orthogonal (on I_n) basis where one function is constant on I_n and the other is the linear function with the values 1 and -1 at the end points t_{n-1} and t_n of I_n. For the general dG(q) method, the orthogonal basis is given by the first $q+1$ Legendre polynomials (translated from $[-1, 1]$ to I_n). In this case, the Jacobian matrix is of the form

$$\mathbf{J}_{\mathcal{O}} = \begin{bmatrix} \mathbf{B} & -\mathbf{B} \\ \mathbf{B} & \mathbf{B} \end{bmatrix} + k_n\begin{bmatrix} \epsilon\,\mathbf{A} & \mathbf{0} \\ \mathbf{0} & \epsilon\,\mathbf{A}/3 \end{bmatrix} - k_n\,\mathbf{R}_{\mathcal{O}}. \tag{6}$$

Finally, a third choice is to use a "mixed" basis with the nodal basis for the solution $\mathbf{U}$ and the orthogonal basis for the test functions $\mathbf{v}$ in (3). The resulting Jacobian is now of the form

$$\mathbf{J}_{\mathcal{M}} = \begin{bmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix} + \frac{k_n}{6}\begin{bmatrix} \epsilon\,\mathbf{A} & -\epsilon\,\mathbf{A} \\ 3\,\epsilon\,\mathbf{A} & 3\,\epsilon\,\mathbf{A} \end{bmatrix} - k_n\,\mathbf{R}_{\mathcal{M}}. \tag{7}$$

Note that the mixed basis makes the time-derivative part of $\mathbf{J}_{\mathcal{M}}$ block-diagonal. The mixed basis is used in the reaction-diffusion solver Cards [5,6]. The motivation for this choice is that in (7), $J_{\mathcal{M}}$ becomes a very simple matrix as the time step k_n is reduced. In fact, for equidistant regular spatial grids, $\mathbf{B} = \mathbf{I}$ and thus $\mathbf{J}_{\mathcal{M}} \to \mathbf{I}$ as $k_n \to 0$. This is exploited in Cards, in the sense that the time step is reduced every time the Krylov-subspace linear solver does not converge or converges too slowly, so that the linear system becomes easier to solve. However, this strategy often results in the use of exceedingly tiny time steps that are dictated solely by the linear algebra, but not by the dG(1) discretization. We now demonstrate that such an "artificial" reduction of the time step can be easily avoided by employing the orthogonal basis for the time discretization and by combining the Krylov iteration with a suitable block preconditioner.

The convergence of Krylov-subspace methods depends on the spectral properties of the coefficient matrix of the linear system to be solved; see, e.g., [9] and the references given there. For the numerical results presented in

this paper, we use the TFQMR Krylov-subspace method [8]. Convergence results for TFQMR can be found in [7]. It turns out that the spectral properties of the three matrices (5)–(7), which result from the three described choices of the basis functions for the time discretization, are vastly different. In particular, the matrix (6) corresponding to the orthogonal basis usually has a spectrum that is most amenable to Krylov-subspace iterations. Moreover, we note that the sparsity of the matrices (5)–(7) is dominated by the stiffness matrix $\mathbf{A}$ in the diffusion part. Thus the orthogonal basis also leads to the sparsest Jacobian matrix (6) with roughly half as many nonzero entries as the two other matrices (5) and (7). This helps to further reduce the computational costs of the Krylov-subspace method. Therefore, we recommend the choice of the orthogonal basis.

4 Preconditioning

In order to speed up Krylov-subspace methods, the basic iteration is combined with preconditioning. The key idea is that instead of the original linear system, $\mathbf{J}\,\mathbf{u} = \mathbf{b}$, the Krylov-subspace iteration is applied to the preconditioned system

$$\left(\mathbf{M}_1^{-1}\,\mathbf{J}\,\mathbf{M}_2^{-1}\right)\,\mathbf{u}' = \mathbf{M}_1^{-1}\,\mathbf{b},$$

and the solution of the original system is then obtained as $\mathbf{u} = \mathbf{M}_2^{-1}\,\mathbf{u}'$. The matrix $\mathbf{M} = \mathbf{M}_1 \cdot \mathbf{M}_2$ is called the preconditioner. It needs to be such that systems with $\mathbf{M}_1$ and $\mathbf{M}_2$ are "easy" to solve and that the spectrum of the preconditioned matrix, $\mathbf{M}_1^{-1}\,\mathbf{J}\,\mathbf{M}_2^{-1}$, is more amenable to the employed Krylov-subspace method. The latter requirement usually means that the eigenvalues of $\mathbf{M}_1^{-1}\,\mathbf{J}\,\mathbf{M}_2^{-1}$ should be clustered around the point 1 and bounded away from the origin 0 in the complex plane.

We now discuss two simple preconditioners for the Jacobian matrices arising in the dG(1) method. As explained, in the previous section, we use the orthogonal basis. Furthermore, we scale the second block row and the second block column of $\mathbf{J}_{\mathcal{O}}$ in (6) by $\sqrt{3}$ so that the stiffness matrix is equally weighted. The resulting Jacobian matrix is

$$\mathbf{J} = \begin{bmatrix} \mathbf{B} & -\sqrt{3}\,\mathbf{B} \\ \sqrt{3}\,\mathbf{B} & 3\,\mathbf{B} \end{bmatrix} + k_n \begin{bmatrix} \epsilon\,\mathbf{A} & \mathbf{0} \\ \mathbf{0} & \epsilon\,\mathbf{A} \end{bmatrix} - k_n\,\mathbf{R}. \tag{8}$$

A simple, but often efficient preconditioner is SSOR. It has the advantage that due to the so-called Eisenstat trick [1], this preconditioner can be implemented so that it requires only little extra computational work, compared with the work per iteration for the unpreconditioned system. The idea is to split the matrix $\mathbf{J} = \mathbf{L} + \boldsymbol{\Delta} + \mathbf{U}$ into its strictly lower part $\mathbf{L}$, its diagonal part $\boldsymbol{\Delta}$, and its strictly upper part $\mathbf{L}$. The SSOR preconditioner is then given by

$$\mathbf{M}_1 = (\mathbf{L} + \boldsymbol{\Delta})\,\boldsymbol{\Delta}^{-1}, \quad \mathbf{M}_2 = \mathbf{U} + \boldsymbol{\Delta}.$$

However, for the linear systems arising in the dG(1) method applied to reaction-diffusion systems (1), SSOR only leads to marginal speed-ups.

The key to obtain an efficient preconditioner is to first reorder the linear system and then to use a block variant of SSOR. Note that in $\mathbf{J}\,\mathbf{u} = \mathbf{b}$ with $\mathbf{J}$ given in (8), there are $2D$ unknowns per spatial grid point. However, in (8), these unknowns are not grouped together, but rather they are spaced $m-1$ apart, where m is the number of spatial grid points. We now reorder $\mathbf{J}$ so that all unknowns associated with each spatial grid point are grouped together. This reordering is done by means of a permutation matrix $\mathbf{P}$. We then consider the following block-wise splitting of the resulting reordered matrix:

$$\mathbf{P}^{\mathrm{T}}\,\mathbf{J}\,\mathbf{P} = \mathbf{L}_{\mathcal{B}} + \boldsymbol{\Delta}_{\mathcal{B}} + \mathbf{U}_{\mathcal{B}}.$$

Here, $\boldsymbol{\Delta}_{\mathcal{B}}$ is a block-diagonal matrix with $2D \times 2D$ diagonal blocks, $\mathbf{L}_{\mathcal{B}}$ is strictly lower block-triangular, and $\mathbf{U}_{\mathcal{B}}$ is strictly upper block-triangular. The block-SSOR preconditioner is then given by

$$\mathbf{M}_1 = \mathbf{P}^{\mathrm{T}}\,(\mathbf{L}_{\mathcal{B}} + \boldsymbol{\Delta}_{\mathcal{B}})\,\boldsymbol{\Delta}_{\mathcal{B}}^{-1}, \quad \mathbf{M}_2 = \mathbf{P}\,(\mathbf{U}_{\mathcal{B}} + \boldsymbol{\Delta}_{\mathcal{B}}).$$

5 Numerical Examples

In this section, we report numerical results for two representative examples. In both cases, the spatial domain is the two-dimensional unit square $\Omega = [0,1] \times [0,1]$, and a regular equidistant mesh is used for Ω. We stress that the preconditioners discussed in this paper can also be employed in the more general situation when the mesh points change across time nodes.

The first example is a predator-prey model; see, e.g., [11,12]. More precisely, we use a model that is analyzed in [12]. This is a system of the form (1) with $D = d = 2$ and functions

$$f_1(\mathbf{u}) = u_1\,(-(u_1 - a_1)\,(u_1 - 1) - a_2 u_2)\,, \quad f_2(\mathbf{u}) = u_2\,(-a_3 - a_4 u_2 + a_2 u_1)\,,$$

where $a_1, a_2, a_3, a_4 > 0$ are real parameters. For our numerical tests, we set $a_1 = 0.25$, $a_2 = 2$, $a_3 = 1$, and $a_4 = 3.4$. Furthermore, in (1), the diffusion coefficients are the constants $\epsilon_1 = \epsilon_2 = 0.01$. A 64×64 grid is used, resulting in $m = 63^2 = 3969$ spatial grid points and a total matrix size of $2Dm = 15876$. In Figure 1, we show the relative residuals for four TFQMR runs with the same matrix. The first run is without preconditioning, the other three runs are with SSOR, block-diagonal ($\mathbf{M}_1 = \mathbf{I}$, $\mathbf{M}_2 = \boldsymbol{\Delta}_{\mathcal{B}}$), and block-SSOR preconditioning. Clearly, block-SSOR results in the most efficient preconditioner.

The second example arises form the bistable equation; see, e.g., [4] and the references given there. This is an equation of the form (1) with $D = d = 1$ and $f_1(u_1) = \alpha\,(u_1 - u_1^3)$. For our tests, we chose the constant $\alpha = 1111.111$ and the constant diffusion coefficient $\epsilon_1 = 1$. We use a small 16×16 grid with $m = 15^2 = 225$ grid points, resulting in a total matrix size of $2Dm = 450$. The

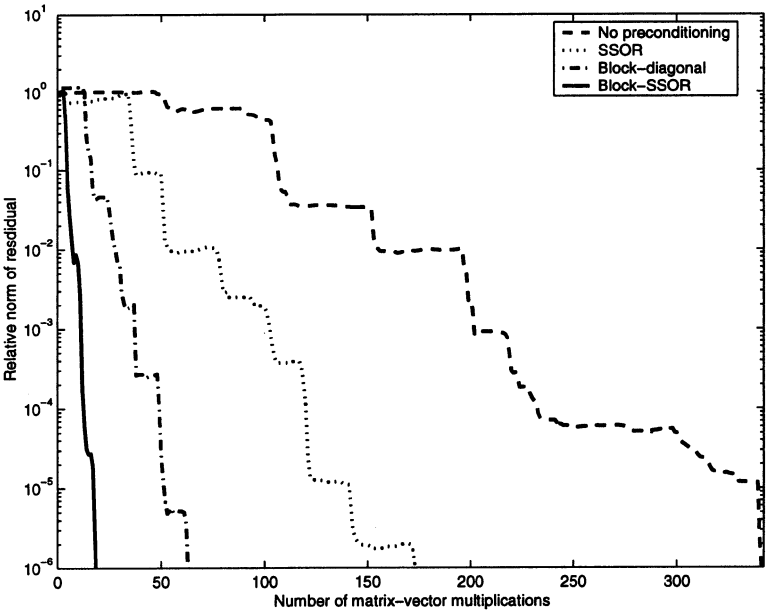


Fig. 1. TFQMR residual history for predator-prey example

small grid was chosen so that we can compute the complete spectrum of the unpreconditioned and the block-SSOR preconditioned matrix. In Figure 2, we show the resulting spectra. Note that the unpreconditioned matrix has eigenvalues on both sides of the origin, causing TFQMR to converge slowly, while block-SSOR nicely clusters the eigenvalues about 1. In Figure 3, we show the relative residuals for TFQMR runs with the unpreconditioned matrix, SSOR, block-diagonal, and block-SSOR preconditioning.

Acknowledgment. The research of the first author was partially supported by the National Science Foundation, DMS 9805748.

References

1. Eisenstat, S.C.: Efficient implementation of a class of preconditioned conjugate gradient methods. SIAM J. Sci. Statist. Comput. **2** (1981) 1–4
2. Eriksson, K., Estep, D., Hansbo, P., Johnson, C.: Introduction to adaptive methods for differential equations. Acta Numerica **4** (1995) 105–158
3. Eriksson, K., Estep, D., Hansbo, P., Johnson, C.: Computational Differential Equations. Cambridge University Press, New York, 1996
4. Estep, D., Larson, M., Williams, R.: Estimating the error of numerical solutions of systems of reaction-diffusion equations. Mem. Amer. Math. Soc., 1999, to appear

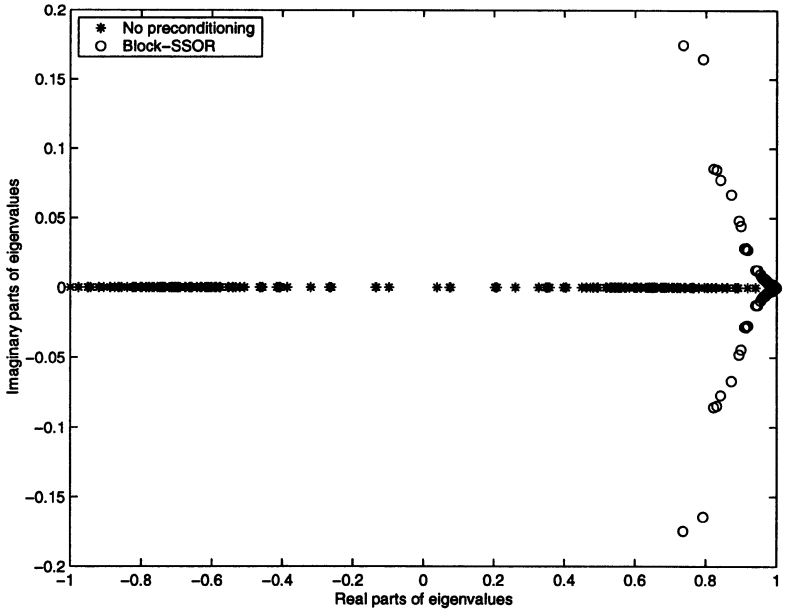


Fig. 2. Spectra for bistable example

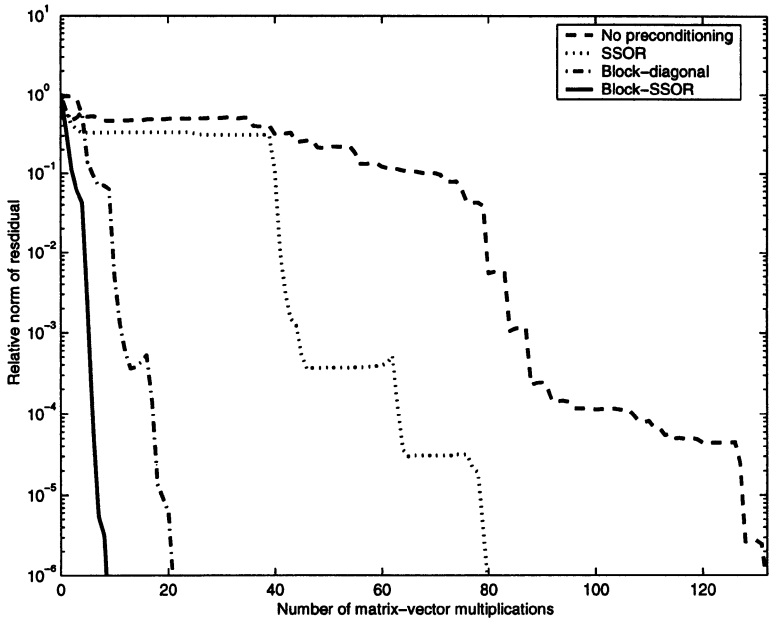


Fig. 3. TFQMR residual history for bistable example

5. Estep, D., Williams, R.: Accurate parallel integration of large sparse systems of differential equations. Math. Models Meth. Appl. Sci. **6** (1996) 535–568
6. Estep, D., Williams, R.: Cards: Concurrent Adaptive Reaction-Diffusion Solver, Version 2.0 (1997)
7. Freund, R.W.: Quasi-kernel polynomials and convergence results for quasi-minimal residual iterations. In: Numerical Methods of Approximation Theory (D. Braess and L. L. Schumaker, eds.). Birkhäuser, Basel, 1992, pp. 77–95
8. Freund, R.W.: A transpose-free quasi-minimal residual algorithm for non-Hermitian linear systems. SIAM J. Sci. Comput. **14** (1993) 470–482
9. Freund, R.W., Golub, G.H., Nachtigal, N.M.: Iterative solution of linear systems. Acta Numerica **1** (1992) 57–100
10. Makridakis, Ch.G., Babuška, I.: On the stability of the discontinuous Galerkin method for the heat equation. SIAM J. Numer. Anal. **34** (1997) 389–401
11. Murray, J.D.: Mathematical Biology, 2nd edition. Springer-Verlag, New York, 1993
12. Smoller, J.A.: Shock Waves and Reaction-Diffusion Equations, 2nd edition. Springer-Verlag, New York, 1994

An Abridged History of Cell Discretization

John Greenstadt

NASA-Ames Research Center
Moffett Field, CA
greensta@worldnet.att.net

Introduction

This brief account will call attention to a line of research which stretches over 40 years, and which appears now to be joining the mainstream of work on the discretization of linear partial differential equations. The first in a series of papers [1] describing the Method of Cells, now called the Cell Discretization Algorithm (CD or CDA), was published in 1959.

H. Swann, starting in the mid-'80's, has made many major contributions to CD, perhaps most significantly in his proofs of convergence and his error estimates for many important paradigm problems in PDE. We shall, however, restrict this sketch mostly to the author's own efforts, since he is not able to do justice to the mathematics developed by Swann, who indeed has described his own work in many papers, e.g., [11].

The Mortar Element Method (MEM), introduced in recent years by Maday et al., [8] is one of the Discontinuous Galerkin Methods (DGM) and, as such, is an outgrowth of the Finite Element Method (FEM). The FEM was originally formulated in terms of an ensemble of elementwise approximations assembled in such a way that the global approximation of the solution is continuous across element interfaces. The MEM relaxes these constraints in a particular way which is quite close to that used in CD. Approaches using similar ideas were put forward in 1977 by Raviart and Thomas [5] and by Dorr in 1989 [7]. In the sections which follow, we shall describe the successive formulations of CD which were reported or published by the author on the dates shown.

The problem domain Ω is subdivided into a set of K subdomains (cells) $\{\Omega_k\}$ of arbitrary shape, with $k = 1, \ldots, K$. Between each pair of cells, Ω_k and Ω_m there is an interface Γ_{km}. The labels (k, m) are restricted to pairs of *contiguous* cells, i.e., neighboring cells that share an interface with nonvanishing measure. In each cell, the true solution is approximated by a specified function of the main argument x and a finite set of parameters $\{\theta_{k\mu}\}$, with $\mu = 1, \ldots, M_k$.

$$u_k = f_k(x, \theta_{k1}, \theta_{k2}, \ldots, \theta_{kM_k}) \equiv u_k(x, \theta_k) \tag{1}$$

In Ω_k, the PDE to be solved is:

$$\mathcal{E}_k(u_k) \equiv -\nabla \cdot (\vec{a}_k \cdot \nabla u_k) + \boldsymbol{b}_k \cdot \nabla u_k + c_k\, u_k - d_k = 0 \tag{2}$$

The double arrow over the a indicates that a is a diadic. On Γ_{km}, the general interface condition is given in terms of::

$$\mathcal{B}_{km}(u_k) \equiv \mathcal{B}^0_{km}(u_k) - R_{km} \equiv \left(P_{km}u_k + Q_{km}\frac{\partial u_k}{\partial n}\right) - R_{km} = 0 \qquad (3)$$

and it is:

$$\Delta\mathcal{B}_{km} \equiv \mathcal{B}_{km}(u_k) - \mathcal{B}_{mk}(u_m) = 0 \qquad (4)$$

1959 [1]

The basic functional Φ_0 first tried was of least-squares form and was equal to

$$\Phi_0 \equiv \sum_{k=1}^{K} \int_{\Omega_k} [\mathcal{E}_k(u_k)]^2 \, d\Omega_k \qquad (5)$$

which generates K separate problems. To connect these separate problems together to form one global problem, we impose the constraints (4) in a least-squares sense by forming the composite functional:

$$\Phi = \sum_{k=1}^{K} \int_{\Omega_k} [\mathcal{E}_k(u_k)]^2 \, d\Omega_k + \lambda \sum_{km} \int_{\Gamma_{km}} [\Delta\mathcal{B}_{km}]^2 \, d\Gamma_{km} \qquad (6)$$

This approach was abandoned in the early 1960's because (1) there was no obvious way to choose a (scaled) value for λ and (2) the Euler equation resulting from the variation of the original functional leads to fourth-order equations having too many solutions.

1967 [2]

The classical variational functional for a self-adjoint PDE (i.e., for which $b_k = 0$), is

$$\Phi_0 = \sum_k \int_{\Omega_k} \left[\tfrac{1}{2}\nabla u_k \cdot \overleftrightarrow{a}_k \cdot \nabla u_k + \tfrac{1}{2} c\, u_k^2 - d_k\, u_k\right] d\Omega_k \qquad (7)$$

The interface constraints are now imposed *weakly*, in the sense of functional analysis. We introduce a linearly independent set of *weight* functions defined on each interface, viz., $\{w_{km\alpha}(x)\}$ with $\alpha = 1, \ldots, L_{km}$ and $x \in \Gamma_{km}$. ($w_{mk\alpha} \equiv w_{km\alpha}$). The interface constraints are then:

$$\int_{\Gamma_{km}} w_{km\alpha}\, \Delta\mathcal{B}_{km}\, d\Gamma_{km} = 0 \quad ; \quad \alpha = 1, \ldots, L_{km} \tag{8}$$

We next assume linear approximations for the u_k's:

$$u_k(x, \theta_k) = \sum_{\mu=1}^{M_k} \theta_{k\mu} \phi_{k\mu}(x) \tag{9}$$

and with these approximations, the problem is then that of finding a stationary point of the quadratic Φ_0 in the space of the degrees of freedom $\{\theta_k\}$, subject to a system of linear constraints on them.

In matrix-vector form, the discrete functional is:

$$\Phi_0 = \sum_{k=1}^{K} \left[\tfrac{1}{2} \theta_k^T S_k \theta_k - \theta_k^T T_k \right] \tag{10}$$

with

$$[S_k]_{\mu\nu} \equiv \int_{\Omega_k} \left\{ \nabla \phi_{k\mu} \cdot \vec{a}_k \cdot \nabla \phi_{k\nu} + c_k\, \phi_{k\mu}\, \phi_{k\nu} \right\} d\Omega_k; \quad [T_k]_\mu \equiv \int_{\Omega_k} \phi_{k\mu}\, d_k\, d\Omega_k \tag{11}$$

The interface conditions become, in matrix-vector form:

$$U_{km}^T \theta_k - U_{mk}^T \theta_m = W_{km} - W_{mk} \tag{12}$$

with

$$[U_{km}]_{\alpha\mu} \equiv \int_{\Gamma_{km}} w_{km\alpha} \mathcal{B}_{km}^0(\phi_{k\mu}) d\Gamma_{km} \quad ; \quad [W_{km}]_\alpha \equiv \int_{\Gamma_{km}} w_{km\alpha} R_{km} d\Gamma_{km} \tag{13}$$

with corresponding formulas for k and m interchanged.

A composite functional incorporating the constraints, using Lagrange multipliers, is:

$$\Phi \equiv \Phi_0 + \sum_{km} \sum_{\alpha=1}^{L_{km}} \lambda_{km\alpha} \int_{\Gamma_{km}} w_{km\alpha}\, \Delta\mathcal{B}_{km}\, d\Gamma_{km} \tag{14}$$

1971 [3]

The foregoing formulation of CD was reported at the Dundee Biennial Conference in 1971. The complete functional Φ in (14) is indefinite in the θ's and λ's. An iterative method to solve the discrete equations, reported at the Dundee Biennial Conference of 1971, was not satisfactory. but was subsequently replaced by an effective procedure.

1972 [4]

To avoid the difficulties of indefinite system matrices associated with the use of Lagrange multipliers, we reduce the constraints to identities. We compute from $\{U_{km}\}$ and S_k a set of matrices $\{V_{km}\}$ and Z_k with the following properties:

$$U_{kp}^T V_{kq} = \delta_{pq}; \quad U_{kp}^T Z_k = 0; \quad Z_k^T S_k V_{kq} = 0 \tag{15}$$

where it is understood that p and q label *contiguous* neighbor cells Ω_p and Ω_q of the cell Ω_k. By introducing new variables, $\{\rho_k\}$ and $\{\sigma_{kp}\}$, and replacing the θ's according to

$$\theta_k = \sum_{p[k]} V_{kp} [\sigma_{kp} + W_{kp}] + Z_k \rho_k \tag{16}$$

(where $p[k]$ labels the *contiguous* neighbors of Ω_k, etc.) The constraint equation (12) reduces to $\sigma_{km} - \sigma_{mk} = 0$. so that the constraint amounts only to identifying σ_{km} and σ_{mk} as the same variable in subsequent differentiations.

The resulting discrete equations reduce to two disjoint sets in the two sets of unknowns $\{\rho_k\}$ and $\{\sigma_{km}\}$:

$$\begin{gathered} A_k \rho_k = Z_k^T T_k \\ \Lambda_{km} \sigma_{km} + \sum_{p[k] \neq m} H_{kmp} \sigma_{kp} + \sum_{q[m] \neq k} H_{mkq} \sigma_{mq} = G_{km} \end{gathered} \tag{17}$$

where

$$\begin{gathered} A_k \equiv Z_k^T S_k Z_k; \quad H_{kpq} \equiv V_{kp}^T S_k V_{kq}; \quad \Lambda_{km} \equiv H_{kmm} + H_{mkk} \\ G_{km} \equiv -\sum_{p[k]} H_{kmp} W_{kp} - \sum_{q[m]} H_{mkq} W_{mq} + V_{km}^T T_k + V_{mk}^T T_m \end{gathered} \tag{18}$$

The first equation in (17) may be solved *cell-by-cell*, since all these equations are disjoint. Hence, the ρ's may be regarded as *intracell* variables. However, since the second equation of (17) links each σ_{km} (on Γ_{km}) with the σ's on all the faces of the two contiguous cells of which Γ_{km} is the interface, the

σ's may be regarded as *interface* variables. This system of equations is not in general p-cyclic.

1982 [6]

The 1982 paper, including the results of the 1972 report, was based on the same algorithm described there, but the discrete equation system for the σ's was solved with the Generalized Conjugate Gradient algorithm, using the diagonal blocks $\{\Lambda_{km}\}$ as preconditioners, instead of SOR. New studies were made here of the effect on accuracy of choice of basis set, cell size, number of moment collocations and number of intracell degrees of freedom (the θ's). Two and three dimensional problems were solved.

1991 [9]

The basic algorithm was extended to nonselfadjoint equations (and systems), which made it necessary to introduce a functional depending on primal and dual unknowns $\{u_{\mathbf{p}k}\}$ and $\{u_{\mathbf{d}k}\}$, basis sets $\{\phi_{\mathbf{p}k\mu}(x)\}$ and $\{\phi_{\mathbf{d}k\mu}(x)\}$, etc. Several convection-diffusion problems with boundary layers were solved with this setup.

1994 [12]

To allow for variable coefficients of the PDE and of the boundary conditions, a nested adaptive Gauss quadrature procedure was included to do the quadratures where needed. It thus became feasible not only to deal with the arbitrary coefficients, but also to admit basis and weight functions much more general than polynomials. By using basis sets which include asymptotic features, boundary singularities, such as those arising from boundary layers, cracks, boundary condition discontinuities, reentrant corners, etc., the overshoot, or "ringing" can be entirely, or almost entirely eliminated without the need for local refinement.

1998 [13]

Full discretization in space-time was done by the standard CD procedure. The domain of the problem was decomposed into rectangular cells so that the space-time problem domain was regarded as having been divided into "spacelike" slabs (in the sense of relativity), and the procedure was to step through the slabs, one-by-one, each time solving the (implicit) spacelike problem. Various test problems and classical wave-propagation problems were solved in this way.

References

1. J. Greenstadt, "On the reduction of continuous problems to discrete form", IBM Jour. Res. Dev., vol. 3, pp. 355-363 (1959)
2. J. Greenstadt, "Cell discretization I - variational basis", IBM New York Sci. Ctr. Report (1967)

3. J. Greenstadt, "Cell discretization" in Conference on Applications of Numerical Analysis, Dundee, Scotland, ed. J. Ll. Morris, Springer-Verlag (1971)
4. J. Greenstadt, "Some numerical tests of cell discretization", IBM Palo Alto Sci. Ctr. Report (1972)
5. P.A. Raviart, J.M. Thomas, "Primal hybrid finite element methods for second-order elliptic equations", Math. Comp., 31 (138) Rpp. 391-413 (1977)
6. J. Greenstadt, "The cell discretization algorithm for elliptic partial differential equations", Siam J. Sci. Stat. Comput., 3, pp. 261-288 (1982)
7. M.R. Dorr, "On the discretization of interdomain coupling in elliptic boundary value problems", in: *Domain Decomposition Methods*, Eds., T.F. Chan, R. Glowinski, J. Periaux, O.B. Widlund, SIAM, Philadelphia, PA 1989
8. C. Bernardi, Y. Maday & A.T. Patera, "A new nonconforming approach to domain decomposition; the Mortar Element Method", *College de France Seminar*, (1990), Pitman, eds., H. Brezis, J.-L. Lions.
9. J. Greenstadt, "Cell discretization of nonselfadjoint linear elliptic PDE's", SIAM J. Sci. Stat. Comput., 12 (1991), pp. 1074-1108.
10. J. Greenstadt, " Solution of elliptic systems of partial differential equations by cell discretization", SIAM J. Sci. Stat. Comput., 14, pp. 627-653 (1993)
11. H. Swann, " On the use of Lagrange multipliers in domain decomposition for solving elliptic problems", Math. of Comp., vol. 60, no. 201, pp. 49-78 (1993)
12. J. Greenstadt, "The removal of overshoot in P.D.E. solutions by the use of special basis functions", Computer Meth. in Appl. Mech. and Eng., (1994)
13. J. Greenstadt, "The application of the cell discretization method to time-dependent problems", Computer Meth. in Appl. Math. and Engineering, (1998)

The Effect of the Least Square Procedure for Discontinuous Galerkin Methods for Hamilton-Jacobi Equations

Changqing Hu[1], Olga Lepsky[2], and Chi-Wang Shu[1]

[1] Division of Applied Mathematics, Brown University, Providence, RI 02912.
[2] Department of Mathematics, Brown University, Providence, RI 02912.

Abstract. In this presentation, we perform further investigation on the least square procedure used in the discontinuous Galerkin methods developed in [2] and [3] for the two-dimensional Hamilton-Jacobi equations. The focus of this presentation will be upon the influence of this least square procedure to the accuracy and stability of the numerical results. We will show through numerical examples that the procedure is crucial for the success of the discontinuous Galerkin methods developed in [2] and [3], especially for high order methods. New test cases using P^4 polynomials, which are at least fourth order and often fifth order accurate, are shown, in addition to the P^2 and P^3 cases presented in [2] and [3]. This addition is non-trivial as the least square procedure plays a more significant role for the P^4 case.

1 Introduction

The two-dimensional Hamilton-Jacobi (HJ) equations we consider in this presentation is given by

$$\varphi_t + H(\varphi_x, \varphi_y) = 0, \qquad \varphi(x, y, 0) = \varphi^0(x, y). \tag{1}$$

In [2], a discontinuous Galerkin finite element method for solving (1) is constructed from the Runge-Kutta discontinuous Galerkin finite element methods developed in [1] for solving conservation laws. It is further investigated in [3]. This method has the flexibility of treating complicated geometry by using arbitrary triangulations, can achieve high order accuracy with a local, compact stencil, and are well suited for efficient parallel implementation.

Equation (1) is in some sense equivalent to the following conservation law system

$$\begin{cases} u_t + H(u, v)_x = 0, \\ v_t + H(u, v)_y = 0, \\ (u, v)(x, y, 0) = (u, v)^0(x, y), \end{cases} \tag{2}$$

if we identify

$$(u, v) = (\varphi_x, \varphi_y). \tag{3}$$

However, one should be careful to boundary conditions when this equivalency is used.

From this equivalence, we have used in [2] φ as the solution variable (i.e. we approximate φ in each cell K by a polynomial of degree at most k), updating

the degrees of freedom of this polynomial via (2) and (3) by the Runge-Kutta discontinuous Galerkin finite element methods in [1] for solving conservation laws.

One important step in the method is the least square procedure. We face an over-determined system when we update the degrees of freedom of the polynomial ($k > 1$) for φ via (2) and (3). This system is solved in a least-square sense:

$$||(\varphi_x - u)^2 + (\varphi_y - v)^2||_{L^1(K)} = \min_{\psi \in P^k(K)} ||(\psi_x - u)^2 + (\psi_y - v)^2||_{L^1(K)} \qquad (4)$$

at each time step.

In [3], it is proven that the above least-square procedure does not destroy the L^2 stability of the discontinuous Galerkin method applied to the system of conservation laws. In fact,

$$||\varphi_x||^2_{L^2(K)} + ||\varphi_y||^2_{L^2(K)} \leq ||u||^2_{L^2(K)} + ||v||^2_{L^2(K)}.$$

In this presentation, we will take a careful look at this least-square procedure. We remark that numerical tests for the P^2 and P^3 cases in [2] and [3] have shown the important role of the least square procedure on the accuracy and stability of the method. We develop and test the method for the P^4 case in this presentation, which has a much greater influence from the least square procedure, as 6 degrees of freedom will be eliminated by this procedure. Examples for smooth problems as well as those with discontinuities are presented.

2 Numerical Examples

Because of space limitation we will only show two numerical examples in this section for the P^4 case. The first example is chosen to show that the accuracy of the solution is not compromised by the least square procedure (which eliminates 6 degrees of freedom). It also shows the non-oscillatory property in the presence of discontinuities in the derivatives. Notice that no nonlinear limiters are used in these runs. The second example shows the importance of the least square procedure: without it the method is not stable or convergent.

2.1 Two-Dimensional Burgers' Equation

The two dimensional Burgers' equation is

$$\begin{cases} \varphi_t + \frac{(\varphi_x + \varphi_y + 1)^2}{2} = 0, & -2 < x < 2, -2 < y < 2, \\ \varphi(x, y, 0) = -\cos\left(\frac{\pi(x+y)}{2}\right) \end{cases} \qquad (5)$$

with periodic boundary conditions.

We use this problem to test the accuracy for P^4 case. Four types of meshes are tested, the first one is uniform rectangular meshes; the second is non-uniform rectangular meshes obtained from a tensor product of one dimensional nonuniform meshes (the meshes in two directions are independent) which is obtained by randomly shifting the cell boundaries in a uniform mesh in the range $[-0.1h, 0.1h]$; the third type is uniform triangular meshes, shown in Fig. 1, left, for the case $h = 1/4$; and the last one is non-uniform triangular meshes, shown in Fig. 1, right, for the

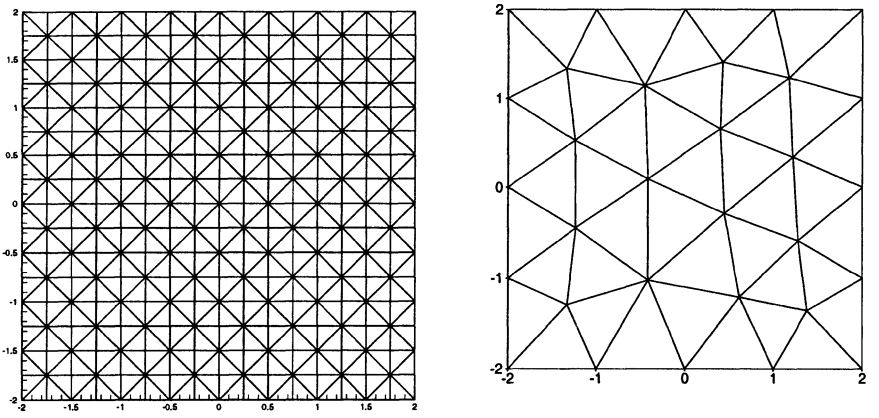


Fig. 1. Triangulation for 2D Burgers' equation. Left: uniform mesh with $h = \frac{1}{4}$; Right: non-uniform mesh with $h = 1$.

coarsest case $h = 1$, where h is just an average length, the refinement of the meshes is done in a uniform way, namely by cutting each triangle into 4 smaller similar ones.

For uniform rectangular meshes, the errors and orders of numerical accuracy are listed in Table 1.

Table 1. Accuracy: Burgers' equation, uniform rectangular mesh.

P^4	with least-square				w/o least-square			
$N \times N$	L^1 error	order	L^∞ error	order	L^1 error	order	L^∞ error	order
10×10	9.73E-04	—	2.43E-03	—	5.75E-04	—	3.02E-03	—
20×20	3.55E-05	4.78	1.66E-04	3.87	4.07E-05	3.82	2.16E-04	3.81
40×40	1.86E-06	4.25	1.89E-05	3.14	2.41E-06	4.08	2.26E-05	3.26
80×80	1.54E-07	3.60	1.16E-06	4.03	1.53E-07	3.98	1.42E-06	3.99
160×160	1.16E-08	3.73	7.04E-08	4.04	9.54E-09	4.00	8.83E-08	4.01

For non-uniform rectangular meshes, the errors and orders of numerical accuracy are listed in Table 2.

For uniform triangular meshes, the errors and orders of numerical accuracy are listed in Table 3.

For non-uniform triangular meshes, the errors and orders of numerical accuracy are listed in Table 4. Here, we also continue the computation to $t = 1.5/\pi$ when the derivative of φ develops discontinuities; results are in Fig. 2.

Table 2. Accuracy: Burgers' equation, non-uniform rectangular mesh.

P^4	with least-square				w/o least-square			
$N \times N$	L^1 error	order	L^∞ error	order	L^1 error	order	L^∞ error	order
10×10	9.14E-04	—	3.03E-03	—	5.70E-04	—	3.87E-03	—
20×20	4.32E-05	4.40	3.01E-04	3.33	4.25E-05	3.75	3.14E-04	3.62
40×40	2.05E-06	4.40	3.16E-05	3.25	2.49E-06	4.09	3.50E-05	3.17
80×80	1.54E-07	3.74	2.17E-06	3.86	1.58E-07	3.98	2.40E-06	3.87
160×160	2.11E-08	3.76	1.48E-07	3.88	9.84E-09	4.01	1.66E-07	3.85

Table 3. Accuracy: Burgers' equation, uniform triangular mesh.

P^4	with least-square				w/o least-square			
h	L^1 error	order	L^∞ error	order	L^1 error	order	L^∞ error	order
1/2	1.90E-04	—	8.54E-04	—	4.33E-04	—	1.49E-03	—
1/4	1.03E-05	4.21	7.08E-05	3.59	4.27E-05	3.34	2.42E-04	2.62
1/8	3.16E-07	5.03	2.09E-06	5.08	2.71E-06	3.98	1.77E-05	3.77
1/16	1.99E-08	3.99	1.63E-07	3.68	3.20E-07	3.08	1.71E-06	3.37
1/32	1.52E-09	3.71	9.22E-09	4.14	2.85E-08	3.49	1.60E-07	3.42

Table 4. Accuracy: Burgers' equation, non-uniform triangular mesh.

P^4	with least-square				w/o least-square			
h	L^1 error	order	L^∞ error	order	L^1 error	order	L^∞ error	order
1/2	1.01E-04	—	7.95E-04	—	1.16E-04	—	8.46E-04	—
1/4	4.34E-06	4.54	7.20E-05	3.47	5.18E-06	4.49	7.21E-05	3.55
1/8	1.85E-07	4.55	3.52E-06	4.35	2.15E-07	4.59	3.86E-06	4.22
1/16	1.03E-08	4.17	1.41E-07	4.64	1.24E-08	4.11	1.88E-07	4.36
1/32	6.56E-10	3.97	7.30E-09	4.27	9.21E-10	3.75	2.63E-08	2.84

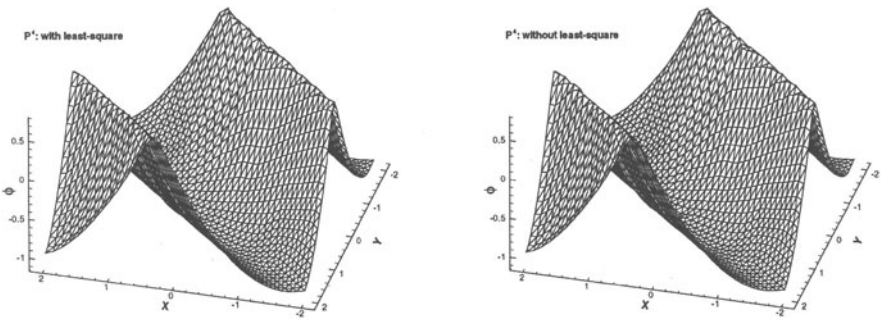


Fig. 2. Two-dimensional Burgers' equation, t=1.5/π^2. Left: with least square; Right: without least square.

We can clearly see that the errors and numerical order of accuracy at least do not deteriorate when the least square procedure is applied. Notice that the least square procedure eliminates 6 degrees of freedoms in this case. From Fig. 2 we can also see that the numerical solution is non-oscillatory when discontinuities appear in the derivatives. Notice that no nonlinear limiters are applied here.

We have also tested the accuracy for a two dimensional nonconvex equation

$$\begin{cases} \varphi_t - \cos(\varphi_x + \varphi_y + 1) = 0, & -2 < x < 2, -2 < y < 2, \\ \varphi(x, y, 0) = -\cos\left(\frac{\pi(x+y)}{2}\right) \end{cases} \tag{6}$$

with periodic boundary conditions. The results are similar to that of the Burgers' equation. We will not present them here to save space.

2.2 A Problem From Computer Vision

The equation is given by

$$\begin{cases} \varphi_t + I(x, y)\sqrt{1 + \varphi_x^2 + \varphi_y^2} - 1 = 0, & -1 < x < 1, -1 < y < 1 \\ \varphi(x, y, 0) = 0 \end{cases} \tag{7}$$

with $\varphi = 0$ as the boundary condition, see [4].

The steady state solution of this problem is the shape lighted by a source located at infinity with vertical direction. Take $I(x, y) = 1/\sqrt{1 + (1 - |x|)^2 + (1 - |y|)^2}$; the exact steady solution is $\varphi(x, y, \infty) = (1 - |x|)(1 - |y|)$. A uniform rectangular mesh of 40 × 40 elements is used.

We use this problem to show that the least-square procedure is important for our method. Fig. 3 contains the history of iterations to the steady state for the methods with and without least-square, we observe that the method without least-square (shown in the right) diverges while the method with least-square (shown in the left) converge to the correct steady state solution. Fig. 4 contains the exact solution and the converged numerical solution with least-square procedure applied; the numerical solution is shown at left, the exact steady state solution at right.

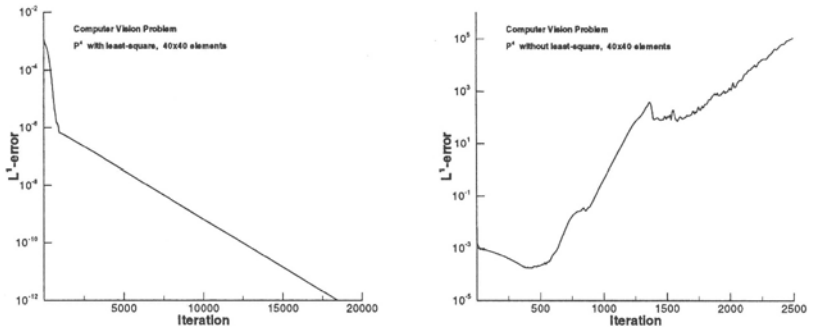


Fig. 3. Computer vision problem, history of convergence. Left: with least square; Right: without least square.

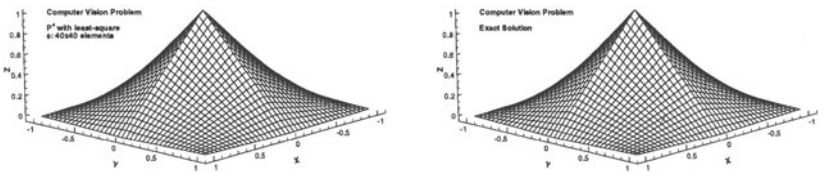


Fig. 4. Computer vision problem: solution φ. Left: numerical solution; Right: exact solution.

Acknowledgments: This research is partially supported by ARO grant DAAG55-97-1-0318, NSF grants DMS-9804985, ECS-9627849 and INT-9601084, NASA Langley grant NAG-1-2070 and AFOSR grant F49620-99-1-0077.

References

1. B. Cockburn, S. Hou and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws IV: the multidimensional case*, Math. Comp., v54 (1990), pp.545-581.
2. C. Hu and C.-W. Shu, *Discontinuous Galerkin finite element method for Hamilton-Jacobi equations*, To appear in SIAM J. Sci. Comput.
3. O. Lepsky, C. Hu and C.-W. Shu, *Analysis of the discontinuous Galerkin method for Hamilton-Jacobi equations*, To appear in Appl. Numer. Math.
4. E. Rouy and A. Tourin, *A viscosity solutions approach to shape-from- shading*, SIAM J. Numer. Anal., v29 (1992), pp.867–884.

A Posteriori Error Estimate in the Case of Insufficient Regularity of the Discrete Space

Guido Kanschat[1], Franz-Theo Suttmeier[2]

[1] Institut für Angewandte Mathematik, Universität Heidelberg, INF 294, 69120 Heidelberg, Germany, Kanschat@iwr.uni-heidelberg.de
[2] Universität Dortmund, Fachbereich Mathematik, Lehrstuhl X, Vogelpothsweg 87, 44221 Dortmund, Germany, Suttmeier@math.uni-dortmund.de

Abstract. We derive a posteriori error estimates for the nonconforming rotated bilinear element. The estimates are residual based and make use of weight factors obtained by a duality argument. Galerkin orthogonality requires the introduction of additional local trial functions. We show that their influence is of higher order and that they can be neglected. The validity of the estimate is demonstrated by computations for the Laplacian and for Stokes' equations.

1 Introduction

Nonconforming finite elements of Crouzeix-Raviart type have become quite popular for solving flow problems. They allow for an inf-sup-stable discretization of the Stokes problem with mixed linear/constant finite elements. Furthermore, they arise naturally from the elimination of weak continuity conditions in some discontinuous Galerkin methods.

We take for example the variational formulation for the Stokes problem on a domain $\Omega \subset \mathbb{R}^2$: Find a velocity field $u = (u_1, u_2)$ and a scalar pressure p, such that

$$a(\{u,p\},\{\varphi,q\}) \equiv (\nabla u, \nabla\varphi) + (p, \operatorname{div}\varphi) - (\operatorname{div} u, q) = (f,\varphi) \qquad \forall\,\{\varphi,q\} \in W \times Q, \tag{1.1}$$

where $W = (V \times V)$, V the Sobolev space $H_0^1(\Omega)$ and $Q = L^2(\Omega)/\mathbb{R}$.

On quadrilateral meshes with possible local refinement, we use the finite element space generated by rotated bilinear element $\tilde{Q}_1$ described in [6] for the discretization W_h^N of W. The pressure is approximated by piecewise constants $Q_h \subset Q$. With these spaces, we want to find approximate solutions to Stokes' equation $(u_h^N, p_h) \in W_h^N \times Q_h$. This discretization is known to fulfil the Babuška-Brezzi stability condition. Since we do not require functions in W_h^N to be in $H_0^1(\Omega;\mathbb{R}^2)$, the form $a(.,.)$ is replaced by a discrete analogue. The discrete solution of (1.1) is thus determined by

$$a_h(\{u_h^N, p\}, \{\varphi, q\}) \equiv \sum_{K\in\mathbb{T}_h} (\nabla u_h^N, \nabla\varphi)_K + \sum_{K\in\mathbb{T}_h} (p_h, \operatorname{div}\varphi)_K - \sum_{K\in\mathbb{T}_h} (\operatorname{div} u_h^N, q)_K = (f,\varphi) \quad \forall\{\varphi, q\} \in W_h^N \times Q_h. \tag{1.2}$$

The forms $a_h(.,.)$ and $a(.,.)$ coincide on the solution space W of the continuous problem. Therefore, we have the Galerkin-orthogonality relation

$$a_h(\{u - u_h^N, p - p_h\}, \{\varphi, q\}) = 0 \quad \forall\{\varphi, q\} \in (W_h^N \cap W) \times Q_h. \tag{1.3}$$

The main goal of this article is to ensure, that the space $W_h^N \cap W$ is large enough, such that (1.3) is feasible.

A posteriori error estimates exploiting local auxiliary problems are presented in [2,3] for triangular elements. We take a different approach and derive weighted residual type estimates as described in [1,4,7].

2 Derivation of the error estimate

First, we illustrate the basic concepts of error estimates for nonconforming adaptive finite element schemes for Poisson's problem with homogeneous Dirichlet boundary conditions

$$a(u,\varphi) \equiv (\nabla u, \nabla\varphi) = (f,\varphi) \qquad \forall\varphi \in V. \tag{2.1}$$

Different boundary conditions cause no further difficulties and are omitted just for simplicity. For a general error measure denoted by the linear functional $J(.)$ on the space V, we define $z \in V$, the solution of the dual problem

$$J(\varphi) = (\varphi, -\Delta z) \qquad \forall\varphi \in V.$$

Then, the standard approach to a posteriori error estimates for conforming finite element solution u_h^C uses the error representation

$$\begin{aligned} J(e) &= (e, -\Delta z) \\ &= \sum_{K\in\mathbb{T}_h} \left\{ (f + \Delta u_h^C, z - z_h^C)_K + \tfrac{1}{2}([\partial_n u_h^C], z - z_h^C)_{\partial K} \right\}, \end{aligned} \tag{2.2}$$

where $e = u - u_h^C$ is the error function and z_h^C denotes a suitable projection of the dual solution z into the discrete space V_h^C. The brackets [.] denote the jumps across the element edges.

The Galerkin orthogonality $a(e, z_h^C) = 0$ and partial integration are essential ingredients to this representation. These have to be modified for nonconforming elements.

While in the triangular case the conforming finite element space is a subset of the nonconforming, this is not true for quadrilaterals. Even worse, the intersection $V_h^N \cap V$ is rather small and does not allow for proper approximation. Therefore, we enrich the element $\tilde{Q}_1$ by a bulb function yielding the trial space V_h^E. On the unit square $K_0 = [-1,1]^2$ we have base functions $\psi_i(x,y) = 1, x, y, x^2 - y^2$ and $\psi_B(x,y) = xy$. The additional functions are chosen cellwise and discontinuous. All node functionals vanish on ψ_B, thus it does not affect the continuity condition across cell boundaries. On the other hand, it introduces the missing base function of the Q_1 element. A simple symmetry argument yields

$$(\psi_B, \psi_i)_{K_0} = (\nabla\psi_B, \nabla\psi_i)_{K_0} = 0 \qquad i = 1, \ldots, 4. \tag{2.3}$$

Since $V_h^N \not\subset V$ and $a(.,.)$ is not defined on $V_h^N \times V_h^N$, we introduce the modified form a_h for the discrete solution $u_h^N \in V_h^N$

$$a_h(u_h^N, \varphi) \equiv \sum_K (\nabla u_h^N, \nabla\varphi)_K = (f, \varphi) \quad \forall\, \varphi \in V_h^N. \tag{2.4}$$

Due to the definition of the spaces V and V_h^E their intersection is the standard conforming finite element space of continuous cellwise bilinear functions V_h^C. Analogously to (1.3), we have the Galerkin-orthogonality relation

$$a_h(u - u_h^N, \varphi) = 0 \qquad \forall\, \varphi \in V_h^C. \tag{2.5}$$

Choosing $\varphi = e$, integrating by parts and Galerkin orthogonality (2.5) yield the modified error representation

$$J(e) = \sum_K \Big\{ (f + \Delta u_h^N, z - z_h^C)_K - \tfrac{1}{2}([\partial_n u_h^N], z - z_h^C)_{\partial K} - \tfrac{1}{2}([u_h^N], \partial_n z)_{\partial K} \Big\},$$

for a suitable interpolation $z_h^C \in V \cap V_h^E$ of z. $\partial_n z$ is not necessarily constant along edges. Therefore, the additional jump term occurs in the corresponding estimate reading

$$|J(e)| \le \eta = \sum_{K \in \mathbb{T}_h} \big(\omega_K \varrho_K + \tilde{\omega}_K \gamma_K + j_K(z) \big), \tag{2.6}$$

with the local residuals ϱ_K, "roughness" γ_K and weights ω_K, $\tilde{\omega}_K$,

$$\varrho_K = h_K \|f + \Delta u_h^N\|_K, \qquad \omega_K = h_K^{-1} \|z - z_h^C\|_K, \tag{2.7}$$

$$\gamma_K = \tfrac{1}{2} h_K^{1/2} \|n \cdot [\nabla u_h^N]\|_{\partial K}, \qquad \tilde{\omega}_K = h_K^{-1/2} \|z - z_h^C\|_{\partial K},$$

$$j_K(z) = \tfrac{1}{2}([u_h^N], \partial_n z)_{\partial K}. \tag{2.8}$$

Since the mean value of $[u_h^N]$ is zero along each edge, we can subtract the constant $\overline{\partial_n z}$ on the right hand side, yielding higher order convergence for

$j_K(z)$. In practice, this term is evaluated with an approximate solution z_h, similarly to the weight factors. For a treatment of the unknown weights ω_K and $\tilde{\omega}_K$ see [1,4,5,7]

Computations show, that the improvement of the discretization by the additional bulb is insignificant. Thus, it should be possible to extend the result to the standard rotated bilinear element. Because of (2.3) in the case of parallelograms, we can represent the solution u_h^N with rotated bilinear elements as $u_h^N = u_h^E - u_h^B$, where u_h^E is the solution with the enriched element above. This may easily be extended to general quadrilaterals. We make use of the error estimate for the enhanced element and modify it by a correction term.

Starting with the "orthogonality" relation

$$a_h(u - u_h^N, \varphi_h^C) = a_h(u_h^B, \varphi_h^B), \tag{2.9}$$

where φ_h^B is the orthogonal projection of φ_h^C into the bulb space, we derive the error representation

$$J(e) = \sum_K \Big\{ \big(\nabla e, \nabla(z - z_h^C)\big)_K - (e, \partial_n z)_{\partial K} \Big\} + a_h(u_h^B, z_h^B). \tag{2.10}$$

The first two terms are treated like in the last section. The last one measures the lack of Galerkin-orthogonality between V_h^N and V_h^C and needs further consideration. In order to estimate its size, we remark that the coeffi-

	V_h^N, Cartesian		V_h^N, distorted	
Cells	$a(e, \varphi_h^C)$	red.	$a(e, \varphi_h^C)$	red.
64	1.88e-06	13.86	1.97e-06	13.11
256	1.22e-07	15.46	1.24e-07	15.96
1024	7.69e-09	15.86	7.74e-09	15.95

	V_h^N, Cartesian		V_h^N, distorted	
Cells	$a(e, \varphi_h^C)$	red.	$a(e, \varphi_h^C)$	red.
4096	4.81e-10	15.97	4.90e-10	15.79
16384	3.01e-11	15.99	3.12e-11	15.72
65536	1.88e-12	16.03	1.89e-12	16.54

Table 1. Error in Galerkin-orthogonality on rectangular and distorted meshes with reduction factors and for the enhanced element for comparison

cients u_K^B and z_K^B defined by $u_h^B|_K = u_K^B \psi_B^K$ and $z_h^B|_K = z_K^B \psi_B^K$, respectively, behave like h^2 each. Since $(\nabla\psi_B^K, \nabla\psi_B^K) = \mathcal{O}(h^2)$, we have by summation over the whole mesh

$$a_h(u_h^B, z_h^B) = \mathcal{O}(h^4). \tag{2.11}$$

The reduction factors in Table 1 confirm fourth order convergence on rectangular *and* arbitrary meshes. Furthermore, we observe, that the constant involved is sufficiently small on both types of meshes. Therefore, the additional term in (2.10) may be neglected in practical computations and u_h^B does not enter the estimate at all.

Cells	$\lvert J(e)\rvert$	η	$C_i\eta/J(e)$	Cells	$\lvert J(e)\rvert$	η	$C_i\eta/J(e)$
256	5.99e-05	1.49e-03	4.96	4903	6.24e-06	5.65e-05	1.81
448	5.47e-05	6.59e-04	2.41	8683	4.40e-06	3.08e-05	1.40
814	3.86e-05	3.30e-04	1.71	15823	1.81e-06	1.65e-05	1.83
1489	2.10e-05	1.92e-04	1.83	28210	1.25e-06	9.58e-06	1.54
2659	1.38e-05	9.92e-05	1.44	50425	5.98e-07	4.92e-06	1.65

Table 2. Error, estimate and efficiency during adaptive solution.

We investigate the efficiency of the error estimate (2.6) in an adaptive iteration (see Table 2). We display the error $J(e)$ of the evaluation, the estimate η and the efficiency index given by the ratio of both. The efficiency is $\mathcal{O}(1)$ with a small constant independent of the mesh size.

3 Stokes' Problem

The error estimate for Stokes' problem is derived in the same way as for the Laplacian. Let the bilinear form $a(.,.)$ again be given as in (1.1). The discrete equations on quadrilateral meshes read

$$\begin{aligned} a_h(\{u_h^N, p_h\}, \{\varphi, q\}) \equiv \sum_K \{ &(\nabla u_h^N, \nabla\varphi)_K + (p, \operatorname{div}\varphi)_K \\ &-(\operatorname{div} u_h^N, q)_K\} = (f, \varphi) \quad \forall\ \{\varphi, q\} \in W_h^N \times Q_h, \end{aligned} \tag{3.1}$$

Remark, that solutions $\{u, p\}$ to the continuous equation (1.1) solve equation (3.1) for all test functions in $W \times Q$ and $W_h^N \times Q_h$, respectively. We omitted the additional bulb functions, since they did not improve convergence of the Laplacian sufficiently. Furthermore,

$$(\operatorname{div}\psi_B^K, q)_K = 0 \quad \forall\, q \in Q_h, \tag{3.2}$$

yielding again a decoupling of the systems for u_h^N and u_h^B and an "orthogonality" relation analogous to (2.9),

$$a_h(\{u - u_h^N, p - p_h\}, \{\varphi_h^C, q_h\} = (\nabla u_h^B, \nabla\varphi_h^B) \quad \forall\, \varphi_h^C \in W_h^C, q_h \in Q_h.$$

Therefore, we apply the same method of error estimation as above. The dual problem determines $\{z, \zeta\} \in W \times Q$, such that

$$J(\{\varphi, q\}) = (\varphi, -\Delta z) + (q, \operatorname{div} z) + (\varphi, \nabla\zeta) \quad \forall\ \{\varphi, q\} \in W \times Q, \tag{3.3}$$

yielding the error estimate

$$|J(\{e, \varepsilon\})| \le \eta = \sum_K (\omega_K \cdot \varrho_K + \tilde{\omega}_K \cdot \gamma_K + j_K(\{z, \zeta\}), \tag{3.4}$$

where the components are defined by

$$\varrho_K = \begin{pmatrix} h_K \|f + \Delta u_h^N - \nabla p_h\|_K \\ h_K \|\operatorname{div} u_h^N\|_K \end{pmatrix} \qquad \omega_K = \begin{pmatrix} h_K^{-1}\|z - z_h^C\|_K \\ h_K^{-1}\|\zeta - \zeta_h\|_K \end{pmatrix},$$

$$\gamma_K = \begin{pmatrix} \frac{1}{2} h_K^{1/2} \|[\partial_n u_h^N]\|_{\partial K} \\ \frac{1}{2} h_K^{1/2} \|[p_h]\| \end{pmatrix} \qquad \tilde{\omega}_K = \begin{pmatrix} h_K^{-1/2}\|z - z_h^C\|_{\partial K} \\ h_K^{-1/2}\|n \cdot (z - z_h^C)\|_{\partial K} \end{pmatrix},$$

and

$$j_K(\{z,\zeta\}) = \tfrac{1}{2}([n \cdot u_h^N], \zeta)_{\partial K} - \tfrac{1}{2}([u_h^N], \partial_n z)_{\partial K}.$$

Again, by introducing appropriate mean values, it is clear that the jump terms do not destroy the order of the estimate. The performance of this error estimator is shown in Table 3. Like for the Laplacian, the efficiency indices are about constant and moderate.

Cells	$\lvert J(e)\rvert$	η	$C_i\eta/J(e)$	Cells	$\lvert J(e)\rvert$	η	$C_i\eta/J(e)$
256	4.96e-03	9.07e-02	1.83	3754	5.07e-04	9.12e-03	1.80
505	2.62e-03	4.98e-02	1.90	7180	2.49e-04	5.31e-03	2.13
1009	1.79e-03	3.72e-02	2.08	13741	1.24e-04	2.56e-03	2.07
1963	6.84e-04	1.59e-02	2.32				

Table 3. Error, estimate and efficiency for Stokes' equation.

References

1. R. Becker and R. Rannacher. A feed–back approach to error control in finite element methods: Basic analysis and examples. *East–West J. Num. Math.*, 4:237–264, 1997.
2. E. Dari, R. Duran, C. Padra, and V. Vampa. A posteriori error estimators for nonconforming finite element methods. *RAIRO M^2AN*, 30(4):385–400, 1996.
3. R. H. W. Hoppe and B. Wohlmuth. Element-oriented and edge-oriented local error estimators for nonconforming finite element methods. *RAIRO, M^2AN*, 30(2):237–263, 1996.
4. G. Kanschat. *Parallel and Adaptive Galerkin Methods for Radiative Transfer Problems.* Dissertation, Universität Heidelberg, 1996.
5. G. Kanschat and F.-T. Suttmeier. A posteriori error estimates for nonconforming finite element schemes. *Calcolo*, 1999. to appear.
6. R. Rannacher and S. Turek. Simple nonconforming quadrilateral stokes element. *Num. Meth. PDE*, 8(2):97–111, March 1992.
7. F.-T. Suttmeier. *Adaptive Finite Element Approximation of Problems in Elasto-Plasticity Theory.* Dissertation, Universität Heidelberg, 1996.

Discontinuous Spectral Element Approximation of Maxwell's Equations

David A. Kopriva, Stephen L. Woodruff, and M.Y. Hussaini

Program in Computational Science and Engineering
The Florida State University, Tallahassee, FL 32306

Abstract. Two discontinuous spectral element methods for the solution of Maxwell's equations are compared. The first method is a staggered-grid Chebyshev approximation. The second is a spectral element (collocation) form of the discontinuous Galerkin method. In both methods, the approximations are discontinuous at element boundaries, making them suitable for propagating waves through multiple materials. Solutions are presented for propagation of a plane wave through a plane dielectric interface, and for scattering off a coated perfectly conducting cylinder.

1 Introduction

High order methods are needed to solve computational electromagnetics problems accurately in the time domain [6]. Nonetheless, low order methods are still typically used, even though they have large dispersion errors (c.f. *IEEE Trans. Antennas Propagat.* **45**(1997) No. 3). The development of new methods has traditionally concentrated either on low order unstructured algorithms that have poor phase properties but can handle both complex geometries and material interfaces, or on high order finite difference methods that have good phase error properties, but are difficult to apply in complex geometries. High order finite difference methods have not found significant application because they are difficult to apply to complex geometries or when material discontinuities are present [6].

Recent advances in the development of spectral multidomain and spectral element methods for CFD and computational aeroacoustics have resulted in methods that can use arbitrarily high order approximations in complex geometries [3,2,4]. Spectral methods, which use high order orthogonal polynomial approximations for the solution unknowns, are well known to have excellent phase properties [1]. By using spectral methods in a multidomain context, where the polynomial approximations are applied locally within subdivisions of a complex geometry, the methods can be made geometrically flexible.The use of spectral approximations that are discontinuous at domain interfaces enables the computation of wave propagation in discontinuous media, while retaining the compactness of the approximation at subdomain interfaces.

In this paper we show how high order discontinuous spectral element methods (DSEM's) can be used effectively for the solution of Maxwell's equations when material discontinuities are present. We will consider two similar

spectral approximations for solving the equations on unstructured quadrilateral grids. The first method is the staggered-grid approximation of Kopriva and Kolias[3], which is a collocation method that uses Chebyshev polynomial expansions and different approximation spaces for the fluxes and solutions. The second is a spectral element (collocation) form of the discontinuous Galerkin method.

The methods are applied to two problems. The first is the propagation of a plane electromagnetic wave through a planar material discontinuity. Both methods are shown to give exponential convergence for this problem. The second problem is the computation of the radar cross-section of a perfectly conducting cylinder that is coated with a thin dielectric. The solutions to this problem are compared to the exact series solution.

2 Problem formulation

Maxwell's equations can be written in conservation form as

$$\mathbf{Q}_t + \nabla \cdot \mathbf{F} = \mathbf{S}, \tag{1}$$

where

$$\mathbf{Q} = \begin{bmatrix} \mathbf{B} \\ \mathbf{D} \end{bmatrix}, \quad \mathbf{F}_i = \begin{bmatrix} \mathbf{e}_i \times \mathbf{E} \\ -\mathbf{e}_i \times \mathbf{H} \end{bmatrix}, \quad \begin{matrix} \mathbf{D} = \epsilon \mathbf{E} \\ \mathbf{B} = \mu \mathbf{H} \end{matrix} \tag{2}$$

and $\mathbf{e}_i$ represents the unit vector in the i^{th} cartesian coordinate direction. The source term, $\mathbf{S}$, is zero when the total wave is computed. For scattering problems, it is more convenient to subtract the incident wave from the solution and solve only for the scattered wave. In the scattered-wave formulation, the source term becomes

$$\mathbf{S} = \begin{bmatrix} (\mu - \mu_{ref})\, \partial \mathbf{H}^i / \partial t \\ (\epsilon - \epsilon_{ref})\, \partial \mathbf{E}^i / \partial t \end{bmatrix} \tag{3}$$

where the vector $[\mathbf{H}^i \quad \mathbf{E}^i]^T$ is the incident wave propagating in the reference medium denoted by ref.

To approximate the solution of this system, we subdivide the domain of interest into non-overlapping quadrilateral subdomains, or elements. The elements can have a general quadrilateral shape, to enable the accurate resolution of curved boundaries with a minimum number of elements.

Each element is mapped individually onto a square/cube by a local isoparametric transformation, $\mathbf{x} = \mathbf{r}(\xi)$.Under this transformation, the equations become

$$\tilde{\mathbf{Q}}_t + \nabla_\xi \cdot \tilde{\mathbf{F}} = \tilde{\mathbf{S}} \tag{4}$$

The new variables in (4) are

$$\tilde{\mathbf{Q}} = J\mathbf{Q}, \quad \tilde{\mathbf{S}} = J\mathbf{S}$$
$$\tilde{\mathbf{F}} = (\mathbf{a}_j \times \mathbf{a}_k) \cdot \mathbf{F} = \left(\tilde{\mathbf{F}}, \tilde{\mathbf{G}}, \tilde{\mathbf{H}}\right) \tag{5}$$

where $\mathbf{a}_i = \frac{\partial \mathbf{r}}{\partial \xi^i}$ and J is the Jacobian of the transformation.

3 Spectral discretization

The approximation of the system (4) by the staggered-grid Chebyshev method is presented in detail in [3], so it will not be described here. In two space dimensions, the discontinuous Galerkin approximation approximates the solution and the fluxes by the polynomials

$$\tilde{\mathbf{Q}}(\xi, \eta) = \sum_{\mu,\nu=0}^{N} \tilde{\mathbf{Q}}_{\mu,\nu} \phi_{\mu,\nu}, \quad \tilde{\mathbf{F}}(\xi, \eta) = \sum_{\mu,\nu=0}^{N} \tilde{\mathbf{F}}_{\mu,\nu} \phi_{\mu,\nu} \tag{6}$$

where $\phi_{\mu,\nu} = \ell_\mu(\xi)\,\ell_\nu(\eta)$. The Lagrange interpolating polynomials,ℓ_i, are defined at the Legendre Gauss quadrature points. The nodal values of the flux are computed from the nodal values of the solution, i.e. $\tilde{\mathbf{F}}_{i,j} = \tilde{\mathbf{F}}\left(\tilde{\mathbf{Q}}_{i,j}\right)$

The residual is required to be orthogonal to the approximation space locally within an element, so

$$\left(\tilde{\mathbf{Q}}_t, \phi_{i,j}\right) + \left(\nabla_\xi \cdot \tilde{\mathbf{F}}, \phi_{i,j}\right) = \left(\tilde{\mathbf{S}}, \phi_{i,j}\right) \quad i,j = 0, 1, \ldots, N \tag{7}$$

where $(\cdot, \cdot)$ represents the usual L^2 inner product.

Integration by parts gives

$$\left(\tilde{\mathbf{Q}}_t, \phi_{i,j}\right) + \int_{\partial E} \phi_{i,j} \tilde{\mathbf{F}} \cdot \hat{N} dS - \left(\tilde{\mathbf{F}}, \nabla_\xi \phi_{i,j}\right) = \left(\tilde{\mathbf{S}}, \phi_{i,j}\right) \quad i,j = 0, 1, \ldots, N \tag{8}$$

where ∂E represents the boundary of the element.

To solve these equations as a collocation method, the integrals in (8) are replaced by Legendre-Gauss quadrature

$$\int_{-1}^{1} v(\xi, \eta)\, d\xi d\eta = \sum_{i,j=0}^{N} v(\xi_i, \eta_j)\, w_i w_j \quad \forall v \in P_{2N+1,2N+1} \tag{9}$$

This replacement is exact provided that the element sides are straight. For straight sides, the transformation $x = r(\xi)$ is bilinear. The products of the metric derivatives or Jacobian with $\phi_{i,j}\phi_{\mu,\nu}$ are polynomials in $P_{2N+1,2N+1}$, making all quadratures exact. If the sides are curved, however, an additional quadrature error is incurred, just as in the C^0 spectral element method. The advantage gained by using quadrature is the ability to keep the mass matrix

diagonal and hence make practical the use of high order elements that are efficient for wave-propagation problems.

After some manipulation, the final approximation in two space dimensions is

$$\frac{d\tilde{\mathbf{Q}}_{i,j}}{dt} + \left[\tilde{\mathbf{F}}(1,\eta_j) \frac{\ell_i(1)}{w_i} + \tilde{\mathbf{F}}(-1,\eta_j) \frac{\ell_i(-1)}{w_i} - \sum_{\mu} \tilde{\mathbf{F}}_{\mu,j} \frac{(\ell_i',\ell_\mu)_N}{w_i} \right] + \left[\tilde{\mathbf{G}}(\xi_i,1) \frac{\ell_j(1)}{w_j} + \tilde{\mathbf{G}}(\xi_i,-1) \frac{\ell_j(-1)}{w_j} - \sum_{\mu} \tilde{\mathbf{G}}_{\mu,j} \frac{(\ell_j',\ell_\nu)_N}{w_j} \right] = \tilde{\mathbf{S}}_{i,j} \tag{10}$$

where the discrete inner product is the Gauss quadrature

$$(u,v)_N = \sum_{i=0}^{N} u_i v_i w_i \tag{11}$$

Both spectral approximations require the evaluation of the flux along the element edges. Since the solution is not collocated on the edges, the values come from evaluating the interpolants of the solution. The interpolants from each side differ at the element faces. The difference is resolved by solving a Riemann problem for the flux. The Riemann problem for Maxwell's equations is discussed in detail in [5]. Given two states $\mathbf{Q}^L$ and $\mathbf{Q}^R$, the resolved flux for a dielectric interface or continuous medium can be written as

$$\mathbf{F}\cdot\hat{n} = \begin{cases} \hat{n} \times \dfrac{(Y^L\mathbf{E}^L - \hat{n}\times\mathbf{H}^L) + (Y^R\mathbf{E}^R + \hat{n}\times\mathbf{H}^R)}{Y^L+Y^R} \\ -\hat{n} \times \dfrac{(Z^L\mathbf{H}^L + \hat{n}\times\mathbf{E}^L) + (Z^R\mathbf{H}^R - \hat{n}\times\mathbf{E}^R)}{Z^L+Z^R} \end{cases} \tag{12}$$

In eq. (12), $Z = \sqrt{\mu/\epsilon}$ and $Y = 1/Z$. Once the normal flux is computed, the fluxes in (10) can be easily constructed.

Boundary conditions can be applied just as they are in a finite volume method (e.g. [5]) by applying appropriate external conditions to the Riemann solver. For the solutions presented here we use an external layer of damping elements to treat external radiation boundary conditions. These damping elements use a rest state as the external solution to the Riemann problem at the outer boundary. Within the elements, the solution is damped by adding a term $-\alpha R(\xi,\eta)Q$ to the right side of (4). Here, α is a positive constant and R is a function that increases as the outer boundary is approached. This term causes waves propagating through the damping elements to decay before and after the waves reflect off the outermost boundary.

Both spatial approximations were integrated in time with a fourth order low-storage Runge-Kutta method.

4 Examples

In this section, we present two examples to show the effectiveness of discontinuous spectral element methods for the solution of Maxwell's equations in

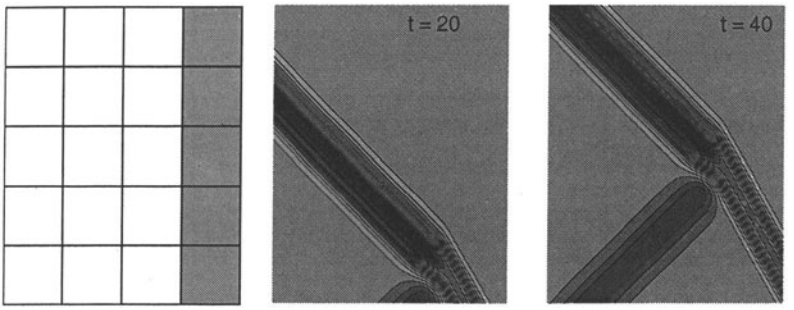


Fig. 1. Propagation of a plane electromagnetic pulse through a material interface. The mesh is shown at the left, with the two materials differentiated by white and gray. Contours of E_z are shown on the right for two times.

inhomogenous materials. The first example is the propagation of a plane wave through a planar dielectric interface. The second problem is the computation of the radar cross-section of a perfectly conducting cylinder that has been coated with a dielectric material.

The first example solves the propagation of a Gaussian plane wave through a discontinuous dielectric interface where $\epsilon^R/\epsilon^L = 2$. Fig. 1 shows the mesh and computed solution at two times. For this problem, the exact solution was used at the exterior boundaries so that only approximation errors would be present. The convergence of the error of the two approximations is shown in Fig. 2. We see that both methods are exponentially convergent but that the staggered-grid approximation requires about two polynomial orders higher to get the same error as the discontinuous Galerkin solution.

In the second example, the radar cross-section (RCS) of a perfectly conducting cylinder of radius a coated with a dielectric of thickness $a/2$ is computed. Within the coating, the dielectric constant, ϵ, is 2.54 the free-space value. Figure 3 shows the geometry and mesh in the neighborhood of the cylinder. Beyond the area shown is a single layer of large damping elements that extends out to 14 cylinder diameters. Fig. 4 shows the radar cross-section of the TM mode at $ka = 1$ for an eighth order computation, where k is the wave number. The computed solutions are compared to the exact series solution. We find that the maximum relative error in the RCS is 1.8×10^{-3} for the discontinuous Galerkin approximation and 4.7×10^{-3} for the staggered grid solution. Thus, the discontinous Galerkin approximation is about a factor of two and a half times more accurate, even with the loss of accuracy coming from the curved element boundaries.

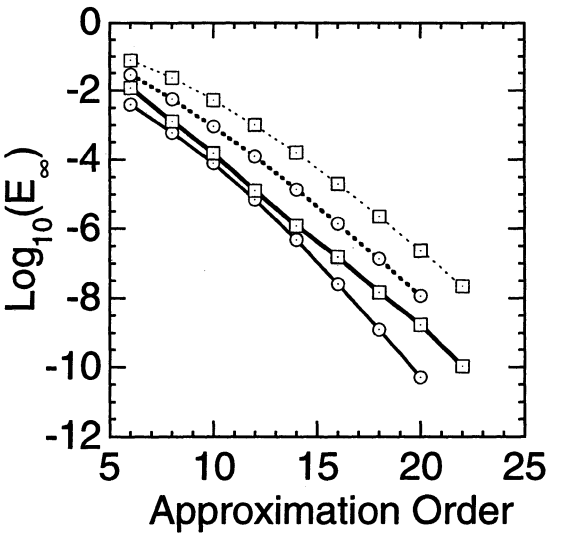


Fig. 2. Maximum error for propagation of a plane electromagnetic pulse through a material interface. Circles represent the discontinuous Galerkin solutions, squares the staggered-grid Chebyshev errors. Solid lines represent errors on the left of the interface, dashed lines represent errors on the right.

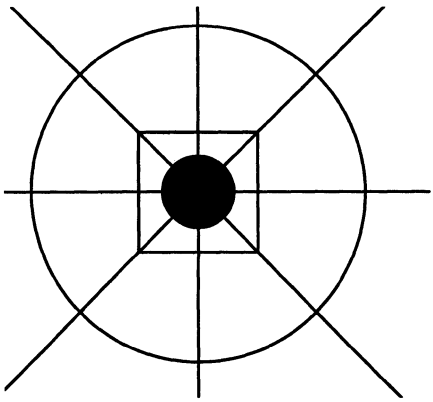


Fig. 3. Element mesh for the computation of the radar cross section of a perfectly conducting cylinder coated with a dielectric in the neighborhood of the cylinder. The coating is shown in gray.

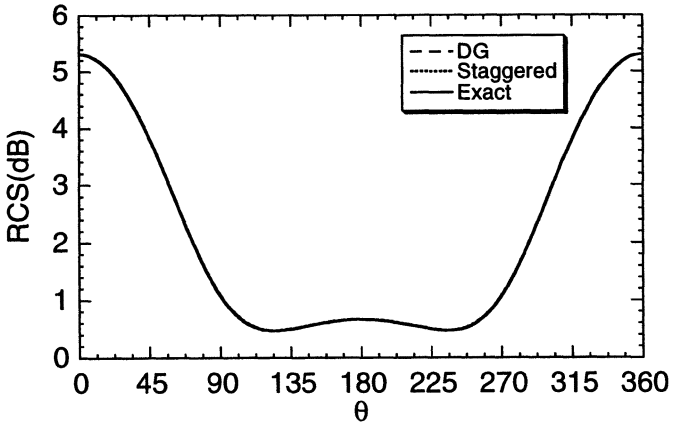


Fig. 4. Radar cross-section for a dielectric-coated perfectly conducting cylinder for $ka = 1$. Theta is the scattering angle relative to the positive x-axis

5 Conclusions

Two similar spectral collocation methods have been used to compute solutions to Maxwell's equations. The first method is a staggered-grid Chebyshev collocation approximation. The second is a collocation form of the discontinuous Galerkin method. Examples were presented showing that the methods are exponentially convergent for the computation of waves propagating through material discontinuities. It was found that the spectral discontinuous Galerkin approximation is more accurate than the spectral staggered-grid approximation, even with curved element sides.

References

1. C. Canuto, M.Y. Hussaini, A. Quarteroni, and T.A. Zang. *Spectral Methods in Fluid Dynamics.* Springer-Verlag, New York, 1987.
2. J.S. Hesthaven. A stable penalty method for the compressible Navier-Stokes equations.II. One dimensional domain decomposition schemes. *SIAM J. Sci. Comp.*, 18:658, 1997.
3. David A. Kopriva and John H. Kolias. A conservative staggered-grid Chebyshev multidomain method for compressible flows. *J. Comp. Phys.*, 125:244–261, 1996.
4. I. Lomtev, C.W. Quillen, and G. Karniadakis. Spectral/hp methods for viscous compressible flows on unstructured 2d meshes. *J. Comp. Phys*, 144:325–357, 1998.
5. A. H. Mohammadian, V. Shankar, and W. F. Hall. Computation of electromagnetic scattering and radiation using a time-domain finite-volume discretization procedure. *Computer Physics Communications*, 68:175–196, 1991.
6. A. Taflove. *Computational Electrodynamics: The Finite-Difference Time-Domain Method.* Artech House, Boston, MA, 1995.

A Posteriori Error Estimation for Adaptive Discontinuous Galerkin Approximations of Hyperbolic Systems

Mats G. Larson[1] and Timothy J. Barth[2]

[1] Stanford University, Mechanics and Computation, Stanford, CA 94305, USA
larson@sccm.stanford.edu
[2] NASA Ames Research Center, NAS Division, Moffett Field, CA 94035, USA
barth@nas.nasa.gov

Abstract. This article considers a posteriori error estimation of specified functionals for first-order systems of conservation laws discretized using the discontinuous Galerkin (DG) finite element method. Using duality techniques, we derive exact error representation formulas for both linear and nonlinear functionals given an associated bilinear or nonlinear variational form. Weighted residual approximations of the exact error representation formula are then proposed and numerically evaluated for Ringleb flow, an exact solution of the 2-D Euler equations.

1 Introduction

A frequent objective in numerically solving partial differential equations is the subsequent calculation of certain derived quantities of particular interest, e.g., aerodynamic lift and drag coefficients, stress intensity factors, etc. Consequently, there is a considerable interest in constructing a posteriori error estimates for such derived quantities so as to improve the reliability and efficiency of numerical computations. For an introduction to a posteriori error analysis see Eriksson et al. [9], related work by Estep et al. [13], Parashivoiu et al. [15], and the recent report of Oden and Prudhomme [14]. For hyperbolic problems and applications in fluid mechanics see Johnson et al. [12], Giles et al. [10], Becker and Rannacher [4] and Süli [16].

This article revisits the topic of a posteriori error estimation of prescribed functionals with special emphasis and consideration given to nonlinear systems of conservation laws discretized using the discontinuous Galerkin (DG) finite method, see for example Johnson and Pitkäranta [11], Bey and Oden [5], and Cockburn et al. [7,8]. In a departure from this previous work, our DG formulation for systems of conservation laws uses entropy symmetrization variables as discussed in detail in the companion papers by the second author [3,2,1].

In Section 2, we briefly review the abstract model for a posteriori error estimation of functionals. Next, we consider the DG method for nonlinear systems of conservation laws and derive concrete error estimates in terms of element residual and weight formulas. Section 4 numerically assesses the sharpness of these estimates for the specific example of Ringleb flow which has a known exact solution via hodograph transformation.

2 A Posteriori Error Estimation of Functionals

Abstract model problem. In this section, we give an abstract presentation of a posteriori error estimation for functionals based on duality techniques. Consider the following abstract variational problem: find $\boldsymbol{u} \in X$ such that

$$\mathcal{A}(\boldsymbol{g}; \boldsymbol{u}, \boldsymbol{v}) = 0 \quad \forall\, \boldsymbol{v} \in X, \tag{2.1}$$

and the corresponding discrete problem: find $\boldsymbol{u}_h \in X_h$ such that

$$\mathcal{A}(\boldsymbol{g}; \boldsymbol{u}_h, \boldsymbol{v}_h) = 0 \quad \forall\, \boldsymbol{v}_h \in X_h. \tag{2.2}$$

Here X is a suitable function space, $X_h \subset X$ is a discrete space, for instance, discontinuous piecewise polynomials of degree k, and $\boldsymbol{g}$ some prescribed data. Note that boundary conditions are *weakly* imposed in the variational statement thus permitting both $\boldsymbol{u}$ and $\boldsymbol{v}$ to reside in X. For brevity, we sometimes write $\mathcal{A}(\boldsymbol{u}_h, \boldsymbol{v}_h) = \mathcal{A}(\boldsymbol{g}; \boldsymbol{u}_h, \boldsymbol{v}_h)$. Our objective is to estimate the error

$$M(\boldsymbol{u}) - M(\boldsymbol{u}_h), \tag{2.3}$$

in a given functional $M(\cdot)$. The first step is to derive an error representation formula.

Error representation: linear case. We first assume that $\mathcal{A}(\cdot,\cdot)$ and $M(\cdot)$ are both linear. To derive a representation formula for the error (2.3), we introduce the dual problem: find $\Phi \in X$ such that

$$\mathcal{A}(\boldsymbol{v}, \Phi) = M(\boldsymbol{v}) \quad \forall\, \boldsymbol{v} \in X. \tag{2.4}$$

Setting $\boldsymbol{v} = \boldsymbol{u} - \boldsymbol{u}_h$ in (2.4) yields

$$\begin{aligned}
M(\boldsymbol{u}) - M(\boldsymbol{u}_h) &= M(\boldsymbol{u} - \boldsymbol{u}_h) && \text{(linearity of } M\text{)} \\
&= \mathcal{A}(\boldsymbol{u} - \boldsymbol{u}_h, \Phi) && \text{(2.4)} \\
&= \mathcal{A}(\boldsymbol{u} - \boldsymbol{u}_h, \Phi - \pi_h\Phi) && \text{(orthogonality)} \\
&= \mathcal{A}(\boldsymbol{u}, \Phi - \pi_h\Phi) - \mathcal{A}(\boldsymbol{u}_h, \Phi - \pi_h\Phi) && \text{(linearity of } \mathcal{A}\text{)} \\
&= -\mathcal{A}(\boldsymbol{u}_h, \Phi - \pi_h\Phi) && \text{(2.1)},
\end{aligned}$$

where $\pi_h\Phi \in X$ is an interpolant of Φ. Thus we have the error representation formula

$$M(\boldsymbol{u}) - M(\boldsymbol{u}_h) = -\mathcal{A}(\boldsymbol{u}_h, \Phi - \pi_h\Phi). \tag{2.5}$$

Error representation: nonlinear case. Consider now the case of a nonlinear variational form $\mathcal{A}(\cdot,\cdot)$ and functional $M(\cdot)$. To perform the analysis given above, we introduce the following mean value linearizations

$$\mathcal{A}(\boldsymbol{g}; \boldsymbol{u}, \boldsymbol{v}) = \mathcal{A}(\boldsymbol{g}; \boldsymbol{u}_h, \boldsymbol{v}) + \overline{\mathcal{A}}(\boldsymbol{g}, \boldsymbol{u}_h, \boldsymbol{u}; \boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{v}) \quad \forall\, \boldsymbol{v} \in X \tag{2.6}$$

$$M(\boldsymbol{u}) = M(\boldsymbol{u}_h) + \overline{M}(\boldsymbol{u}_h, \boldsymbol{u}; \boldsymbol{u} - \boldsymbol{u}_h), \tag{2.7}$$

and the dual linearized problem: find $\Phi \in X$ such that

$$\overline{\mathcal{A}}(g, u_h, u; v, \Phi) = \overline{M}(u_h, u; v) \quad \forall\, v \in X. \tag{2.8}$$

In addition, we have the following orthogonality relation

$$\overline{\mathcal{A}}(g, u_h, u; u - u_h, v_h) = 0 \quad \forall\, v_h \in X_h. \tag{2.9}$$

Proceeding in the same fashion as above, using simplified notation for brevity,

$$\begin{aligned} M(u) - M(u_h) &= \overline{M}(u - u_h) && (2.7)\\ &= \overline{\mathcal{A}}(u - u_h, \Phi) && (2.4)\\ &= \overline{\mathcal{A}}(u - u_h, \Phi - \pi_h\Phi) && (2.9)\\ &= \mathcal{A}(u, \Phi - \pi_h\Phi) - \mathcal{A}(u_h, \Phi - \pi_h\Phi) && (2.6)\\ &= -\mathcal{A}(u_h, \Phi - \pi_h\Phi), && (2.1) \end{aligned}$$

thus yielding the following final error representation formula

$$M(u) - M(u_h) = -\mathcal{A}(g; u_h, \Phi - \pi_h\Phi). \tag{2.10}$$

Abstract a posteriori error estimates. Starting from (2.5) or (2.10), we derive various error estimates by estimating the right hand side of (2.10) using standard inequalities. Later, the sharpness of these inequalities is numerically assessed. Consider the following sequence of direct estimates

$$\begin{aligned} |M(u) - M(u_h)| &= \Big|\sum_T \mathcal{A}_T(g; u_h, \Phi - \pi_h\Phi)\Big| && (2.11)\\ &\le \sum_T |\mathcal{A}_T(g; u_h, \Phi - \pi_h\Phi)| && (2.12)\\ &\le \sum_T R_T(u_h) \cdot W_T(\Phi), && (2.13) \end{aligned}$$

where $\mathcal{A}_T(\cdot,\cdot)$ denotes the restriction of $\mathcal{A}(\cdot,\cdot)$ to the element T. Further $R_T(u_h)$ is a computable estimate of the residual of u_h on T, and $W_T(\Phi)$ is a weight on T describing the local influence of Φ, both are defined below.

3 A Posteriori Error Estimates for the DG Method

First-order nonlinear system. Consider the prototype conservation law problem: find $u : \Omega \to \mathbb{R}^m$ such that

$$\begin{aligned} L(u) = f^i(u)_{,x_i} &= 0 \quad \text{in } \Omega, \\ \tilde{A}^-(n; g, u)(g - u) &= 0 \quad \text{on } \Gamma, \end{aligned} \tag{3.14}$$

where Ω is a domain in $\mathbb{R}^d$ with boundary Γ with local exterior normal vector n and $\tilde{A}(n; u) \equiv n_i f^i_{,u}$ is the flux Jacobian matrix. In addition, $\tilde{A}(n; g, u)$ denotes the mean value matrix obtained from the Volpert path integration

$$\tilde{A}(n; s, t) = \int_0^1 \tilde{A}\,(n; t + \theta(s - t))\; d\theta \tag{3.15}$$

and $P^{\pm}(\boldsymbol{n};\boldsymbol{s},\boldsymbol{t})$ the associated characteristic projectors. Throughout, we assume that $\boldsymbol{u}$ denotes the symmetrization variables so that the matrices $\tilde{A}$ are necessarily symmetric.

Next, consider a finite element tessellation $\mathcal{T}$ of Ω composed of nonoverlapping elements T_i, $\mathcal{T} = \cup T_i$, $T_i \cap T_j = \emptyset$, $i \neq j$ and Γ_T the tessellated boundary. The prototype system can be restated in variational form[1]: find $\boldsymbol{u} \in X$ such that $\forall\, \boldsymbol{v} \in X$

$$\mathcal{A}(\boldsymbol{g};\boldsymbol{u},\boldsymbol{v}) = \sum_{T\in\mathcal{T}} \Big((L(\boldsymbol{u}),\boldsymbol{v})_T + \langle \tilde{A}^-(\boldsymbol{n};\boldsymbol{g},\boldsymbol{u})(\boldsymbol{g}-\boldsymbol{u}),\boldsymbol{v}\rangle_{\partial T\cap\Gamma_T} \qquad (3.16)$$

$$+ \langle P^-(\boldsymbol{n};\boldsymbol{u}_-,\boldsymbol{u}_+)[\boldsymbol{f}(\boldsymbol{n};\boldsymbol{u})]^+_-,\boldsymbol{v}_-\rangle_{\partial T\setminus\Gamma_T} \Big)$$

Note that other mathematically equivalent formulations are possible by grouping together terms element-wise and edge-wise. The above particular grouping has been chosen as it reflects a discrete balance of conserved quantities on an element-by-element basis. In Section 4, we briefly revisit the possibility of alternate groupings although our numerical results show that the element-wise grouping presented above yields superior estimates.

A posteriori error estimate residuals and weights. A straightforward application of Cauchy-Schwarz inequality (with $\tilde{A}_0$ introduced from entropy symmetrization theory for dimensional consistency) (3.16) yields the following element residuals R_T and weights W_T for use in (2.13)

$$R_T(\boldsymbol{u}_h) = \begin{bmatrix} \|L(\boldsymbol{u}_h)\|_{\tilde{A}_0^{-1},T} \\ \|P^-(\boldsymbol{n};\boldsymbol{u}_{-,h}\boldsymbol{u}_{+,h})[\boldsymbol{f}(\boldsymbol{n};\boldsymbol{u}_h)]^+_-\|_{\tilde{A}_0^{-1},\partial T\setminus\Gamma_T} \\ \|\tilde{A}^-(\boldsymbol{n},\boldsymbol{g};\boldsymbol{u}_h)(\boldsymbol{g}-\boldsymbol{u}_h)\|_{\tilde{A}_0^{-1},\Gamma_T} \end{bmatrix} \qquad (3.17)$$

$$W_T(\Phi) = \begin{bmatrix} \|\Phi - \pi_h\Phi\|_{\tilde{A}_0,T} \\ \|\Phi - \pi_h\Phi\|_{\tilde{A}_0,\partial T\setminus\Gamma_T} \\ \|\Phi - \pi_h\Phi\|_{\tilde{A}_0,\Gamma_T} \end{bmatrix} \qquad (3.18)$$

Approximating the dual problem. The weight formulas (3.18) require the calculation of the quantity $\Phi - \pi_h\Phi$ from the dual problem which requires a priori knowledge of both $\boldsymbol{u}$ and $\boldsymbol{u}_h$ for use in the mean value linearizations (2.6) and (2.7). Since $\boldsymbol{u}$ is not generally known, we supplant this calculation with the approximate discrete counterpart to (2.8): find $\Phi_h \in X_h$ such that

$$\overline{\mathcal{A}}(\boldsymbol{g},\boldsymbol{u}_h,\boldsymbol{u}_h;\boldsymbol{v}_h,\Phi_h) = \overline{M}(\boldsymbol{u}_h,\boldsymbol{u}_h;\boldsymbol{v}_h) \quad \forall\, \boldsymbol{v} \in X_h. \qquad (3.19)$$

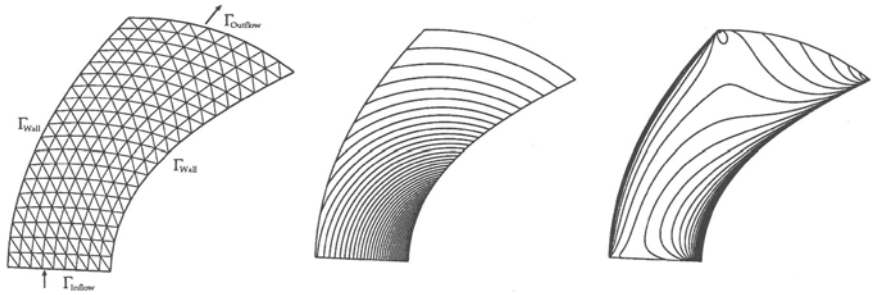
Observe that $\overline{\mathcal{A}}(\boldsymbol{g},\boldsymbol{u}_h,\boldsymbol{u}_h;\boldsymbol{v},\Phi_h)$ and $\overline{M}(\boldsymbol{u}_h,\boldsymbol{u}_h;\boldsymbol{v}_h)$ are precisely the Jacobian linearized forms of the respective operators. Using the techniques described in Barth [3,2], exact Jacobian derivatives of the DG scheme for systems of conservation laws have been derived and used in all subsequent calculations. We have investigated the computation of the needed dual solution terms using two different techniques:

[1] In actual implementations it is desirable to use an integrated-by-parts form (see for example [3,2]) so that exact discrete conservation is achieved on elements with inexact quadrature and/or path integration (3.15).

(1) High-order approximation. Suppose $\boldsymbol{u}_h^{(k)}$ denotes a numerical solution computed in $X_h^{(k)}$. Embed $\boldsymbol{u}_h^{(k)}$ in $X_h^{(l)}$, $l > k$ and approximate $\Phi - \pi_h^{(k)}\Phi \approx \Phi_h^{(l)} - \pi_h^{(k)}\Phi_h^{(l)}$. This technique is employed in the calculations given below.
(2) Recovery post-processing. Let $\mathcal{R}_h^{(l)}\Phi_h^{(k)} : X_h^{(k)} \mapsto X_h^{(l)}$ denote a recovery operator, approximate $\Phi - \pi_h^{(k)} \approx \mathcal{R}_h^{(l)}\Phi_h^{(k)} - \Phi^{(k)}$, $l > k$. Recovery operators based on local compact supported least-squares fitting are considered in a forthcoming report by the present authors.

4 Numerical Results

To evaluate the accuracy of the error representation formulas given in Sect. 2, Ringleb flow (an exact solution of the 2-D Euler equations obtained via hodograph transformation, see [6]) is computed in the channel geometry shown in Fig. 4.1(a). Next, the vertical force component exerted on the channel walls



(a) Coarsest triangulation (342 elements). (b) Primal solution iso-density contours. (c) Dual solution iso-density contours.

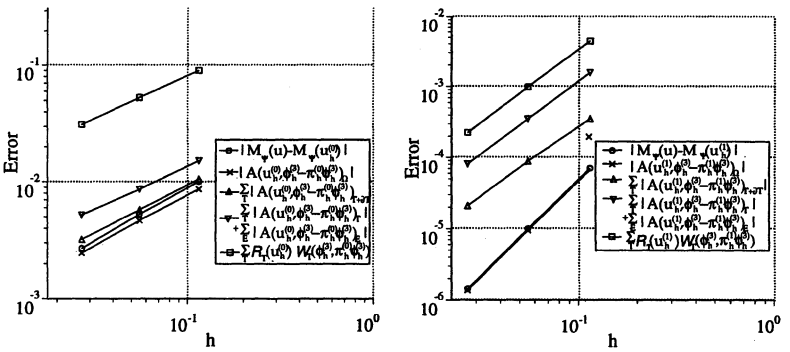
Fig. 4.1. Ringleb flow test problem. Primal and dual solutions calculated using the DG discretization with cubic elements for the vertical force functional $M_\Psi(u)$.

is computed from the functional

$$M_\Psi(\boldsymbol{u}) = \int_{\Gamma_{\text{Wall}}} (\Psi \cdot \mathbf{n})\, p(\boldsymbol{u})\, dl \tag{4.20}$$

with $p(\boldsymbol{u})$ the fluid pressure and Ψ a constant vertical vector. Iso-density contours of the Ringleb primal and dual solutions are given in Figs. 4.1(b-c).

We now evaluate the validity and sharpness of the error estimate formulas (2.11)–(2.13) and (3.18). In Fig. 4.2 we graph for constant (a) and linear (b) approximation: ($\circ$) the exact error; ($\times$) estimate (2.11); (Δ) estimate (2.12); (∇) estimate obtained from element-edge form of (3.16); ($\square$) estimate (3.18). In all cases the dual problem is defined by (3.19) and solved using cubic polynomials. The difference between ($\circ$) and ($\times$) is caused by linearization, i.e., replacing $\boldsymbol{u}$ in (2.8) by $\boldsymbol{u}_h$ to get (3.19). This appears to be a very small error. Next, the more significant loss due to use of the Triangle Inequality is graphed in (Δ). This prevents cancellation between elements. Further error is introduced via Cauchy-Schwarz ($\square$) thus preventing cancellation within the element. Finally, note that the element based estimate (Δ) is notably superior to the element-edge based estimate (∇), where in the latter case contributions are grouped together in such a way that element conservation is violated.



(a) Piecewise constant elements. (a) Piecewise linear elements.

Fig. 4.2. Ringleb flow problem. Sharpness of error estimate inequalities for the vertical force functional (4.20).

Based on our numerical experimentation, we propose the adaptive method:

- Evaluate a *stopping criterion* via $|A(\boldsymbol{u}_h, \phi - \pi_h\phi)_\Omega|$.
- Evaluate an *adaptation criterion* via $|A(\boldsymbol{u}_h, \phi - \pi_h\phi)_T|$.

In addition, the adaptation criterion may be further improved by the use of sharpened variants of the Triangle and Cauchy-Schwarz Inequalities. We consider these topics further in a forthcoming paper.

References

1. T. J. Barth. Simplified discontinuous Galerkin methods for systems of conservation laws with convex extension, 1999. Proceedings of the 1st International Conference on Discontinuous Galerkin Methods.
2. T. J. Barth. Simplified discontinuous Galerkin methods for systems of conservation laws with convex extension. *Math. Comp.*, submitted 1999.
3. T.J. Barth. Numerical methods for gasdynamic systems on unstructured meshes. In *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws*, Vol 5 of LNCSE, pages 195–285. Springer-Verlag, Heidelberg, 1998.
4. R. Becker and R. Rannacher. Weighted a posteriori error control in FE methods. In *Proc. ENUMATH-97, Heidelberg*. World Scientific Pub., Singapore, 1998.
5. K. Bey. A Runge-Kutta discontinuous finite element method for high speed flows. Technical Report 91-1575, AIAA, Honolulu, Hawaii, 1991.
6. G. Chiocchia. Exact solutions to transonic and supersonic flows. Technical Report AR-211, AGARD, 1985.
7. B. Cockburn, S.Y. Lin, and C.W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: 1-D systems. *J. Comp. Phys.*, 84:90–113, 1989.
8. B. Cockburn and C.W. Shu. The Runge-Kutta discontinuous Galerkin method for conservation laws V: Multi-D systems. *J. Comput. Phys.*, 141:199-224, 1998.
9. K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. Introduction to numerical methods for differential equations. *Acta Numerica*, pages 105–158, 1995.
10. M. Giles, M. Larson, M. Levenstam, and E. Süli. Adaptive error control for finite element approximations of the lift and drag coefficients in viscous flow. preprint NA-97/06, Comlab, Oxford University, 1997.
11. C. Johnson and J. Pitkäranta. An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comp.*, 46:1–26, 1986.
12. C. Johnson, R. Rannacher, and M. Boman. Numerics and hydrodynamics stability theory: towards error control in CFD. *SIAM J. Numer. Anal.*, 32:1058–1079, 1995.
13. M. G. Larson. *Analysis of Adaptive Finite Element Methods*. PhD thesis, Department of Applied Mathematics, Chalmers University, Göteberg, Sweden, 1996.
14. J. T. Oden and S. Prudhomme. Goal-oriented error estimation and adaptivity for the finite element method. Technical Report 99-015, TICAM, U. Texas, Austin,TX, 1999.
15. M. Parashivoiu, J. Peraire, and A. Patera. A posteriori finite element bounds for linear-functional outputs of elliptic partial differential equations. *Comput. Meth. Appl. Mech. Engrg.*, 150:289–312, 1997.
16. E. Süli. A posteriori error analysis and adaptivity for finite element approximations of hyperbolic problems. In *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws*, Vol. 5 of LNCSE, pp. 122–194. Springer-Verlag, Heidelberg, 1998.

A Numerical Example on the Performance of High Order Discontinuous Galerkin Method for 2D Incompressible Flows

Jian-Guo Liu[1] and Chi-Wang Shu[2]

[1] Department of Mathematics, University of Maryland, College Park, Maryland 20742

[2] Department of Mathematics, Brown University, Providence, Rhode Island 02912.

Abstract. In this presentation we explore a recently introduced high order discontinuous Galerkin method for two dimensional incompressible flow in vorticity streamfunction formulation [8]. In this method, the momentum equation is treated explicitly, utilizing the efficiency of the discontinuous Galerkin method. The streamfunction is obtained by a standard Poisson solver using continuous finite elements. There is a natural matching between these two finite element spaces, since the normal component of the velocity field is *continuous* across element boundaries. This allows for a correct upwinding gluing in the discontinuous Galerkin framework, while still maintaining total energy conservation with no numerical dissipation and total enstrophy stability. The method is suitable for inviscid or high Reynolds number flows. In our previous work, optimal error estimates are proven and verified by numerical experiments. In this presentation we present one numerical example, the shear layer problem, in detail and from different angles to illustrate the resolution performance of the method.

1 Introduction

The two-dimensional time dependent incompressible Navier-Stokes and Euler equations in vorticity streamfunction formulation we consider in this presentation is given by

$$\begin{aligned} &\omega_t + \nabla \cdot (\mathbf{u}\omega) = \frac{1}{\mathrm{Re}} \Delta \omega \\ &\Delta \psi = \omega, \qquad \mathbf{u} = \nabla^{\perp} \psi, \end{aligned} \tag{1}$$

with periodic boundary conditions (Re=∞ for the Euler equations). Other boundary conditions can also be handled, see [8].

A discontinuous Galerkin method for solving (1) has been developed by the authors in [8]. We first emphasize that, for Euler equations and high Reynolds number (Re $\gg$ 1) Navier-Stokes equations (1), it is advantageous to treat both the convective terms and the viscous terms explicitly. The methods discussed in [8] are stable under standard CFL conditions. Since the momentum equation (the first equation in (1)) is treated explicitly in the discontinuous Galerkin framework, there is no global mass matrix to invert, unlike conventional finite element methods. This makes the method highly efficient for parallel implementation. As any finite element

method, our approach has the flexibility for complicated geometry and boundary conditions. The method is adapted from the Runge-Kutta discontinuous Galerkin methods discussed by Cockburn et al. in a series of papers, see for example [3], [4], [6] and other papers in this volume.

The main difficulties in solving incompressible flows are the incompressibility condition and boundary conditions. The incompressibility condition is global and is thus solved by the standard Poisson solver for the streamfunction ψ using continuous finite elements. One advantage of our approach is that there is no matching conditions needed for the two finite element spaces for the vorticity ω and for the streamfunction ψ. The incompressibility condition, represented by the streamfunction ψ, is exactly satisfied pointwise, and is naturally matched with the convective terms in the momentum equation. The normal velocity $\mathbf{u}\cdot\mathbf{n}$ is automatically *continuous* along any element boundary, allowing for correct upwinding for the convective terms and still maintaining a total energy conservation and total enstrophy stability.

In [8], a proof is given for L^2 stability, both in the total enstrophy (L^2 norm of the vorticity) and in the total energy (L^2 norm of the velocity), which does not depend on the regularity of the exact solutions. For smooth solutions error estimates are also obtained in [8].

We briefly describe the setup of the method here. More details can be found in [8].

We start with a triangulation $\mathcal{T}_h$ of the domain Ω, consisting of polygons of maximum size (diameter) h, and the following two approximation spaces

$$V_h^k = \left\{ v : v \Big|_K \in P^k(K),\ \forall K \in \mathcal{T}_h \right\}, \qquad W_{0,h}^k = V_h^k \cap C_0(\Omega), \tag{2}$$

where $P^k(K)$ is the set of all polynomials of degree at most k on the cell K.

For the Euler equations ((1) with Re=∞), the numerical method is defined as follows: find $\omega_h \in V_h^k$ and $\psi_h \in W_{0,h}^k$, such that

$$\langle \partial_t \omega_h\, v\rangle_K - \langle \omega_h\, \mathbf{u}_h \cdot \nabla v\rangle_K + \sum_{e\in\partial K} \langle \mathbf{u}_h \cdot \mathbf{n}\, \widehat{\omega_h}\, v^-\rangle_e = 0, \quad \forall v \in V_h^k, \tag{3}$$

$$-\langle \nabla\psi_h \cdot \nabla\varphi\rangle = \langle \omega_h\, \varphi\rangle, \quad \forall \varphi \in W_{0,h}^k, \tag{4}$$

with the velocity field obtained from the stream function by

$$\mathbf{u}_h = \nabla^\perp \psi_h. \tag{5}$$

Here $\langle\cdot\rangle$ is the usual integration over either the whole domain Ω or a subdomain denoted by a subscript. Same thing for $\|\cdot\|$ for the L^2 norm.

Notice that the normal velocity $\mathbf{u}_h\cdot\mathbf{n}$ is *continuous* across any element boundary e, but both the solution ω_h and the test function v are discontinuous there. We take the values of the test function from within the element K, denoted by v^-. The solution at the edge is taken as a single valued flux $\widehat{\omega_h}$, which can be either a central or preferably a upwind biased average. In this presentation we use the (global) Lax-Friedrichs upwind biased flux defined by

$$\mathbf{u}_h \cdot \mathbf{n}\, \widehat{\omega_h} = \tfrac{1}{2}\left[\mathbf{u}_h \cdot \mathbf{n}\left(\omega_h^+ + \omega_h^-\right) - \alpha\left(\omega_h^+ - \omega_h^-\right)\right] \tag{6}$$

where α is the maximum of $|\mathbf{u}_h \cdot \mathbf{n}|$ globally. Other fluxes can be found in [8].

Navier-Stokes equations (1) can be handled in a similar way, with the additional viscous terms treated by the local discontinuous Galerkin technique in [4]. The detail can be found in [8], together with stability proofs, error estimates and some numerical examples.

2 A Numerical Example

In this section we describe a detailed numerical study of the double shear layer problem. This problem, proposed in [1], has become a popular numerical benchmark problem for numerical methods of incompressible flows. The problem is to solve the Euler or Navier-Stokes equations (1) in the domain $[0, 2\pi] \times [0, 2\pi]$ with a periodic boundary condition and an initial condition:

$$\omega(x, y, 0) = \begin{cases} \delta \cos(x) - \frac{1}{\rho} sech^2((y - \pi/2)/\rho) & y \leq \pi \\ \delta \cos(x) + \frac{1}{\rho} sech^2((3\pi/2 - y)/\rho) & y > \pi \end{cases} \tag{7}$$

where ρ and δ are two small parameters measuring the width of the shear and the magnitude of the perturbation. We take $\delta = 0.05$ in all the tests in this section.

We refer to [5] for a comparison with nonlinear ENO schemes, and to [7] for a comparison with upwind-central schemes.

The following four scenarios are considered in this section.

The first scenario is the thin shear layer defined by Brown and Minion in [2], corresponding to $\rho = \pi/100$. This is in contrast to the thick shear layer corresponding to $\rho = \pi/15$ in their paper. We perform the numerical simulation for this thin shear layer with Reynolds number Re=$2000/2\pi$, as was used in [2]. A uniform rectangular mesh of 256×256 cells is used for the P1/Q1 (referring to P1 in the vorticity space and Q1 in the stream function space). And a uniform rectangular mesh of 162×162 cells is used for the P2/Q2. We can see from Fig. 1 that the flow is completely resolved for time up $t = 12$. This is at a much later time than the numerical experiments performed in [2], where they showed that a second order Godunov upwind projection method with 256×256 points produces spurious nonphysical vortices.

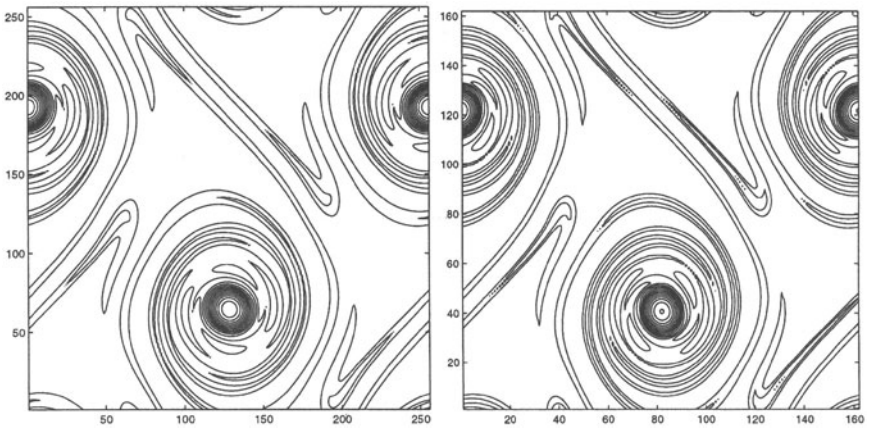


Fig. 1. Contour of vorticity ω at $t = 12$. 30 equally spaced contour lines between $\omega = -12$ and $\omega = 12$. Thin shear layer corresponding to $\rho = \pi/100$. Re=$2000/2\pi$. Left: P1/Q1 result with 256×256 cells; Right: P2/Q2 result with 162×162 cells;

The second scenario is the same thin shear layer $\rho = \pi/100$ with a much higher Reynolds number Re=$70000/2\pi$. The simulation result with a uniform rectangular mesh of 256×256 cells with P2/Q2 method up to $t = 8$ is shown in Fig. 2. We notice that our numerical method is still stable in this case. A time history for energy and enstrophy in Fig. 3 shows that the physical viscosity is still dominating the numerics at such high Reynolds numbers, according to the decay of energy and enstrophy. This indicates that the built-in numerical viscosity of the methods is very small.

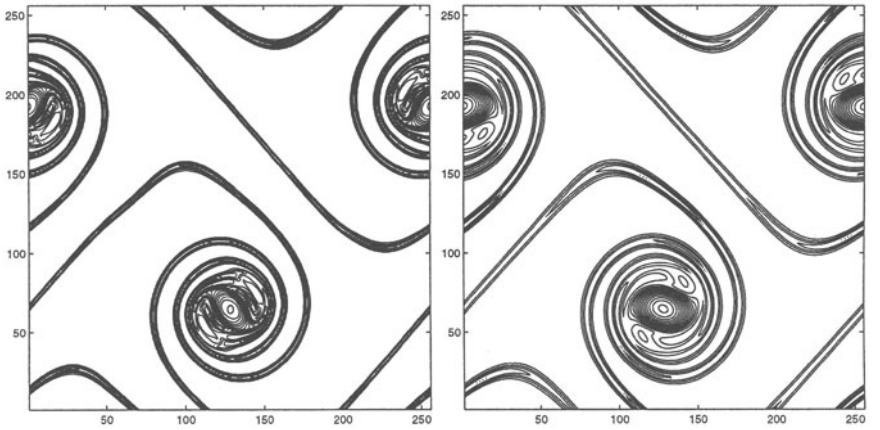


Fig. 2. Contour of vorticity ω at $t = 8$. 30 equally spaced contour lines between $\omega = -14.5$ and $\omega = 14.5$. Thin shear layer corresponding to $\rho = \pi/100$. Re=$70000/2\pi$. 256×256 cells with P2/Q2 scheme (left). As a comparison, we also plot the result of 256×256 cells with P1/Q1 scheme computation with Re=$20000/2\pi$ (right).

The third scenario is a ultra thin shear layer with $\rho = \pi/200$ with Reynolds number Re=$20000/2\pi$. The simulation result with a uniform rectangular mesh of 512×512 cells with P1/Q1 method up to $t = 8$ is shown in Fig. 4. We notice that our numerical method is still stable in this case.

The last scenario is for the same thin layer case as in the first scenario, but this time we keep coarsening our mesh for any given polynomial degree in our method until the method becomes unstable. This way we can find the minimum number of cells needed for each case. We find out that we would need 175×175 cells for the P1/Q1 method to resolve the flow in this case, 150×150 cells for the P2/Q2 method, and 64×64 cells for the P3/Q3 method. The results of the P2/Q2 and P3/Q3 cases are shown in Fig. 5.

Acknowledgments: The research of the first author is supported by NSF grant DMS-9805621. The research of the second author is supported by ARO grant DAAG55-97-1-0318, NSF grants DMS-9804985, ECS-9627849 and INT-9601084, NASA Langley grant NAG-1-2070 and AFOSR grant F49620-99-1-0077.

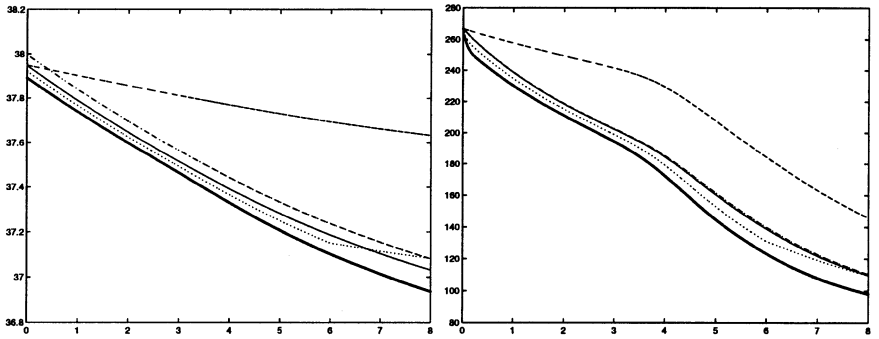


Fig. 3. The time history of energy at left (square of the L_2 norm of velocity, computed by rectangle rules) and enstrophy at right (square of the L_2 norm of vorticity, computed by rectangle rules). Thin shear layer corresponding to $\rho = \pi/100$. 175×175 cells with P^1/Q^1 with Re=$2000/2\pi$ in dots; 150×150 cells with P^2/Q^2 with Re=$2000/2\pi$ in solid line; 64×64 cells with P^3/Q^3 with Re=$2000/2\pi$ in dash-dot line; 256×256 cells with P^1/Q^1 with Re=$6000/2\pi$ in dotted line; 256×256 cells with P^2/Q^1 with Re=$7000/2\pi$ in dashed line

References

1. J. Bell, P. Colella and H. Glaz, *A second order projection method for the incompressible Navier-Stokes equations*, J. Comput. Phys., v85 (1989), pp.257–283.
2. D. L. Brown and M. Minion, *Performances of under-resolved two dimensional incompressible flow simulations*, J. Comput. Phys., v122 (1995), pp.165–183.
3. B. Cockburn, S. Hou and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws IV: the multidimensional case*, Math. Comp., v54 (1990), pp.545-581.
4. B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection diffusion systems*, SIAM J. Numer. Anal., v35 (1998), pp.2440–2463.
5. W. E and C.-W. Shu, *A numerical resolution study of high order essentially non-oscillatory schemes applied to incompressible flow*, J. Comput. Phys., v110 (1994), pp.39–46.
6. G. Jiang and C.-W. Shu, *On cell entropy inequality for discontinuous Galerkin methods*, Math. Comp., v62 (1994), pp.531–538.
7. D. Levy and E. Tadmor, *Non-oscillatory central schemes for the incompressible 2-D Euler equations*, Math. Research Letters, v4 (1997), pp.1–20.
8. J.-G. Liu and C.-W. Shu, *A high order discontinuous Galerkin method for 2D incompressible flows*, J. Comput. Phys., submitted.

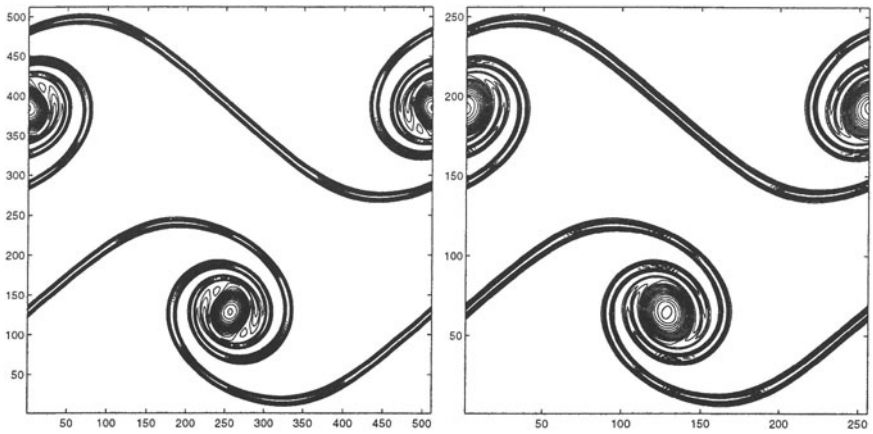


Fig. 4. Contour of vorticity ω at $t = 12$. 30 equally spaced contour lines between $\omega = -17$ and $\omega = 17$. Ultra thin shear layer corresponding to $\rho = \pi/200$. Re=$20000/2\pi$. 512×512 cells with P1/Q1 method (Left); As a comparison we plot a 256×256 cells with P1/Q1 method for $\rho = \pi/100$ and Re=$20000/2\pi$ on the right.

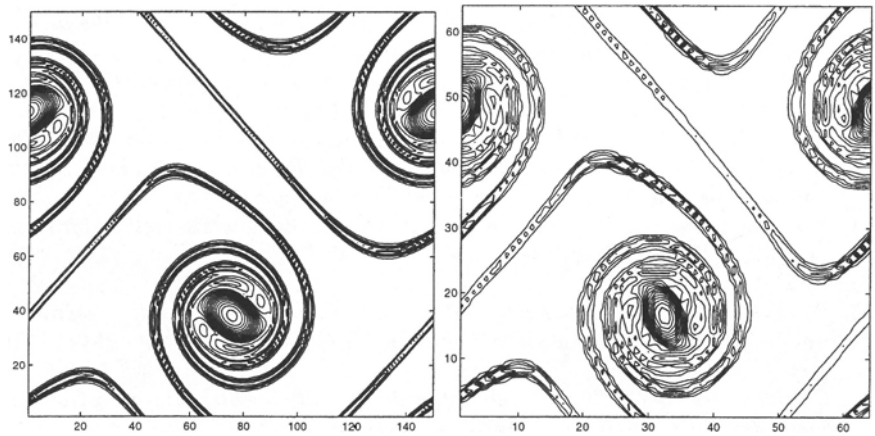


Fig. 5. Contour of vorticity ω at $t = 8$. 30 equally spaced contour lines between $\omega = -12$ and $\omega = 12$. Thin shear layer corresponding to $\rho = \pi/100$. Re=$2000/2\pi$. Marginal resolution study. Left: P2/Q2 with a 150×150 mesh; Right: P3/Q3 with a 64×64 mesh.

A Discontinuous Galerkin Method in Moving Domains

I. Lomtev, R.M. Kirby and G.E. Karniadakis*

Division of Applied Mathematics, Brown University

Abstract. We present a matrix-free discontinuous Galerkin method for simulating compressible viscous flows in two- and three-dimensional moving domains in an Arbitrary Lagrangian Eulerian (ALE) framework. Spatial discretization is based on standard structured and unstructured grids but using an orthogonal spectral hierarchical basis. The method is third-order accurate in time, and converges exponentially fast in space for smooth solutions. We also report on open issues related to quadrature crimes and over-integration.

1 Introduction

In the current work we develop algorithms for the *compressible* Navier-Stokes equations in moving domains with high-order spectral/*hp* element discretizations. We use a *discontinuous Galerkin* approach that allows the use of an *orthogonal* polynomial basis of different order in each element. In particular, we develop a discontinuous Galerkin formulation for *both* the advective as well as the diffusive components of the Navier-Stokes equations. The discontinuous basis is orthogonal, hierarchical, and maintains a tensor-product property even for non-separable domains [1], [2]. Moreover, in the proposed method the conservativity property is maintained automatically in the element-wise sense by the discontinuous Galerkin formulation, while monotonicity is controlled by varying the order of the spectral expansion and by performing h-refinement around discontinuities.

We have followed the Arbitrary Lagrangian Eulerian (ALE) framework as in previous works, e.g. [3–5], but with an important difference on computing the grid velocity. Specifically, we developed a modified version of the force-directed method [6] to compute the grid velocity via incomplete iteration. We then update the location of the vertices of the elements using the known grid velocity. In addition to the ALE treatment, the proposed method is new both in the formulation (e.g. construction of inviscid and viscous fluxes, use of characteristic variables, no need for limiters) as well as in the discretization as it uses *polymorphic* (i.e. various shapes) subdomains. We will demonstrate this flexibility in the context of simulating viscous flows that require accurate boundary layer resolution.

* Author to whom all correspondence should be addressed

2 Numerical Formulation

We consider the non-dimensional compressible Navier-Stokes equations, which we write in compact form in an Eulerian reference frame as

$$\boldsymbol{U}_t + \nabla \cdot \mathbf{F} = Re_\infty^{-1} \nabla \cdot \mathbf{F}^\nu \quad \text{in } \Omega \tag{1}$$

where $\mathbf{F}$ and $\mathbf{F}^\nu$ correspond to inviscid and viscous flux contributions, respectively, and Re_∞ is the reference Reynolds number. Here the vector $\boldsymbol{U} = [\rho, \rho u_1, \rho u_2, \rho u_3, \rho e]^t$ with $\mathbf{u} = (u_1, u_2, u_3)$ the local fluid velocity, ρ the fluid density, and e the total energy. Splitting the Navier-Stokes operator in this form allows for a separate treatment of the inviscid and viscous contributions, which, in general, exhibit different mathematical properties.

In the following, we will solve the Navier-Stokes equations in a time-dependent domain $\Omega(t)$ by discretizing on a grid whose points may be moving with velocity $\mathbf{U}^g$, which is, in general, *different* than the local fluid velocity. This is the so-called Arbitrary Lagrangian Eulerian (or ALE) formulation which reduces to the familiar Eulerian and Lagrangian framework by setting $\mathbf{U}^g = 0$ and $\mathbf{U}^g = \mathbf{u}$, respectively [3,4]. In this context, we will review briefly the discontinuous Galerkin formulation employed in the proposed method. In the proposed formulation, no flux limiters are necessary as the entropy condition is satisfied in the L_2-norm as has been shown theoretically in [7].

2.1 Discontinuous Galerkin for Advection

Using the Reynolds transport theorem we can write the Euler equations in the ALE framework following the formulation proposed in [4] as

$$\boldsymbol{U}_t + G_{i,i} = -U_i^g{}_{,i} \boldsymbol{U}\,, \tag{2}$$

where the ALE flux term is defined as

$$G_i = (u_i - U_i^g)\boldsymbol{U} + p[0, \delta_{1i}, \delta_{2i}, \delta_{3i}, u_i], \quad i = 1, 2, 3.$$

We can recover the *Euler flux* $\mathbf{F}$ (see equation 1) by simply setting $\mathbf{U}^g = 0$, and in general we have that $G_i = F_i - U_i^g \boldsymbol{U}$. Now if we write the ALE Euler equations in terms of the *Euler flux* then the source term on the right-hand-side of equation (2) is eliminated and we obtain:

$$\boldsymbol{U}_t + F_{i,i} - U_i^g \boldsymbol{U}_{,i} = 0, \tag{3}$$

which can then be recasted in the standard quasi-linear form

$$\boldsymbol{U}_t + [A_i - U_i^g I]\boldsymbol{U}_{,i} = 0,$$

where $A_i = \partial F_i / \partial U$ $(i = 1, 2, 3)$ is the flux Jacobian and I is the unit matrix. In this form it is straightforward to obtain the corresponding characteristic variables since the ALE Jacobian matrix can be written

$$A_i^{ALE} \equiv [A_i - U_i^g I] = R_i \cdot [A_i - U_i^g I] \cdot L_i\,,$$

where brackets denote matrix. Here the matrix Λ contains in the diagonal the eigenvalues of the original Euler Jacobian matrix A, and R and L are the right- and left-eigenvector matrices, respectively, containing the corresponding eigenvectors of A. Notice that the shifted eigenvalues of the ALE Jacobian matrix do not change the corresponding eigenvectors in the characteristic decomposition.

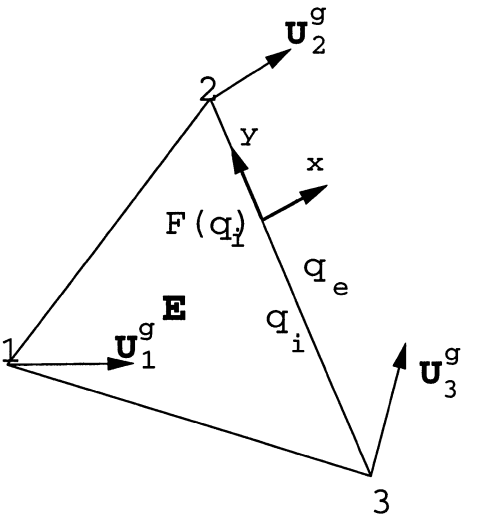


Fig. 1. Notation for a triangular element.

To explain the discontinuous Galerkin ALE formulation we consider the two-dimensional equation for advection of a conserved scalar q in a region $\Omega(t)$

$$\frac{\partial q}{\partial t} + \nabla \cdot \mathbf{F}(q) - \mathbf{U}^g \cdot \nabla q = 0.$$

In the discontinuous Galerkin framework, we test the equation above with discontinuous test functions v separately on each element (e) (see also [8,9]) to obtain

$$(v, \partial_t q)_e + (v, \nabla \cdot \mathbf{F}(q))_e - (v, \mathbf{U}^g \cdot \nabla q)_e$$

$$+ \int_{\partial T_e} v[\tilde{f}(q_i, q_e) - \mathbf{F}(q) - (q_{up} - q_i)\cdot \mathbf{U}^g] \cdot \hat{n} ds = 0. \tag{4}$$

Here $(\cdot, \cdot)$ denotes inner product evaluated over each element, and $\tilde{f}$ is a numerical boundary flux [8]; the notation is explained in figure 1; q_{up} is an upwind value and q_i is an interior value. Notice that this form is different

than the form used in the work of [10] and [11] where the time derivative is applied to the inner product, i.e.

$$\partial_t(v,q)_e + (v, \nabla\cdot\mathbf{F}(q))_e - (v, \mathbf{U^g}\cdot\nabla q)_e - (v, q\nabla\cdot\mathbf{U^g})_e$$

$$+ \int_{\partial T_e} v[\tilde{f}(q_i,q_e) - \mathbf{F}(q) - (q_{up}-q_i)\cdot\mathbf{U^g}]\cdot\hat{n}ds = 0. \tag{5}$$

Note that from Reynolds transport theorem we have that

$$\partial_t \int_{\Omega(t)} q d\Omega = \int_{\Omega(t)} (q_t + q\nabla\cdot\mathbf{U}^g) d\Omega,$$

where the partial time derivative on the right-hand-side is with respect to the moving ALE grid. The difference between the forms in equations 4 and 5 is that the different treatment of the time derivative introduces a term in the second case (equation (5)) that involves the divergence of the grid. While the two forms are equivalent in the continuous case, they are not necessarily equivalent in the discrete case.

To compute the boundary terms, we follow an upwind treatment based on *characteristics* similar to the work in [8], including here the term representing the grid motion. To this end, we need to linearize the ALE Jacobian *normal to the surface*, i.e. $[A - U_n^g I] = R[\Lambda - U_n^g I]L$, where U_n^g is the velocity of the grid in the $\mathbf{n}$ direction. The term $(q_{up} - q_i)$ expresses a jump in the variable at inflow edges of the element resulted from an upwind treatment. In the case of a system of conservation laws the numerical flux $\tilde{f}$ is computed from an approximate Riemann solver [8].

In this formulation, the space of test functions may contain formally discontinuous functions, and thus the corresponding discrete space contains polynomials within each "element" but zero outside the element. Here the "element" is, for example, an individual triangular region T_i in the computational mesh applied to the problem. Thus, the computational domain $\Omega = \bigcup_i T_i$, and T_i, T_j overlap only on edges. We employ a spectral polynomial basis as a trial basis which is orthogonal and has tensor-product form on triangles and quadrilaterals in two-dimensions as well as on tetrahedra, hexahedra, prisms and pyramids in three-dimensions (see [12] for details).

The diffusion part is not altered by the moving grids and we refer the reader to [13]; also details on the new grid velocity algorithm can be found in [14].

2.2 Quadrature Crimes and Over-Integration

Consider a space of polynomials of degree up to p used in a triangle. Then we *define* the quadrature order $q = p$, so that there are $q+2$ Gauss-Lobatto points in a-direction (across from the "collapsed" vertex) and $q+1$ Gauss-Radau

points in b-direction (the other two directions). In this case, the quadrature rule is exact for polynomials of degree $2q$ in the interior of the elements (in non-curvilinear geometries). All the boundary terms, i.e. boundary integrals and boundary fluxes are computed by the interpolation of the interior values to $q+1$ Gauss points on each edge. This is how we match the points of boundary flux computations between the adjacent elements. If the orders in the elements are different, then the maximum number of edge quadrature points should be taken for stability. In order to preserve conservativity the edge-fluxes need to be projected to the smaller polynomial space (between the two adjacent elements). We also note that on the edges the quadrature is exact for polynomials of degree $2q+1$. These are necessary conditions to guarantee the maximum possible accuracy of $p+1$ as proven in [15] for the linear case.

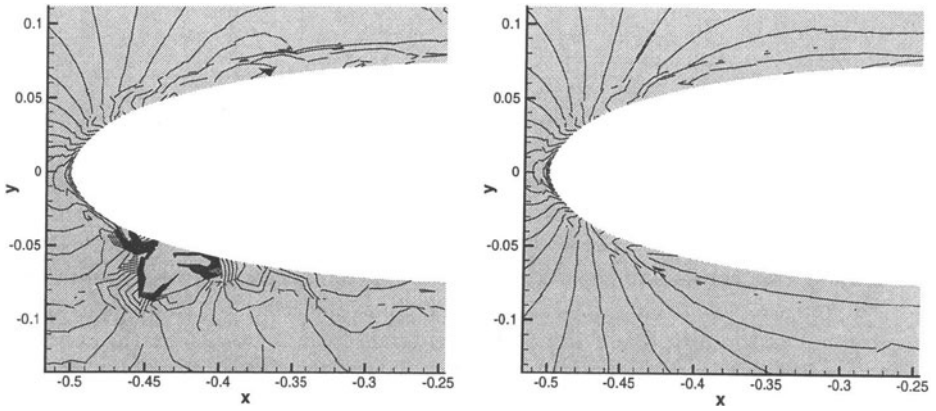


Fig. 2. Density contours for $Re = 10,000$ and $Ma = 0.2$. The simulation on the left was performed with $p = q = 3$ and on the right with $p = 3; q = 4$.

We report here on problems that may arise due to quadrature errors in solving the compressible Navier-Stokes equations using the discontinuous Galerkin method. In numerical experiments even with very low-order discretizations similar to the ones used in [16], we have found that *over-integration* in computing inner products in the weak formulation is important in obtaining asymptotically stable results, i.e. after long-time integration. This behavior usually occurs at relatively high Reynolds number, e.g. at flow past an airfoil at $Re = 10,000$. This is shown in figure 2 where we plot the case $p = q = 3$ on the left and the case ($p = 3$; $q = 4$) on the right. We see that the latter is stable, but the former develops very steep gradient very

close to the leading edge that renders the computation eventually unstable. If we simply increase both the the interpolation order and the quadrature order so that $p = q = 4$ the method still diverges, which reinforces further the aforementioned finding on over-integration. For a possible explanation of the exact nature of this behavior see [14].

3 Simulation of 3D Flapping Wing

In [14] we have demonstrated exponential convergence in space and third-order time accuracy. Here we include a 3D simulation example that shows the flexibility of the method in that it can sustain large grid distortions in three-dimensions without the need for h-remeshing. We consider the flow past a three-dimensional wing formed by a prismatic NACA 4420 airfoil placed at 20 degrees angle of attack. Two-dimensional subsonic and supersonic simulations were presented in [8]. In particular, we consider the wing moving according to

$$u = 0; \quad v = A cos(2\pi f t) H(z - z_0) 2(z - z_0)/(L_z/2); \quad w = 0$$

where z runs along the span of the airfoil, $z_0 = 2.5$ is the reference point, $L_z = 5$ is the spanwise length of the airfoil, $A = 0.5$ is the amplitude of the motion, and f is the frequency with $2\pi f = 1.57$; also $H(z)$ is the Heaviside function. The motion we simulate resembles in some general way the flapping motion characteristic of insect flight [17]. We have performed simulations at chord Reynolds number $Re = 680$ and Mach number $M = 0.3$. The discretization consists of 15,870 tetrahedra of $p = 3$ polynomial order and the time step was taken $\Delta t = 0.00025$. The origin of the reference frame is at the midpoint of the airfoil and the domain extends from $x = -2.5$ at the inflow to $x = 7.5$ at the outflow and from $y = -2.5$ to $y = 2.5$ at the sides. Here the non-dimensionalization is with respect to the chord length ($C = 2$ in our computations) and the freestream velocity ($U_\infty = 1.75$ in our computations).

We present a sequence of flow visualizations during one flapping cycle in figure 3. We use minima of density contours to capture the vortex tubes that are shed off the flapping wing. We see that there seems to be a clear lag between the motion of the flapping wing and the visualized vortex tubes. The flapping motion essentially re-arranges the vortex street resulting in a very different lift and drag force distribution (see [14] for details).

Finally, we conclude by commenting on the computational cost of the simulations. Both the two-dimensional and the three-dimensional simulations were run using an MPI-based parallel version of the method presented here with the partitioning based on a multi-level graph approach provided by the METIS software [18]. Specifically, the three-dimensional simulation was run for 33,000 time steps for a total of 50 CPU hours on 32 processors of the IBM SP2/P2SC system.

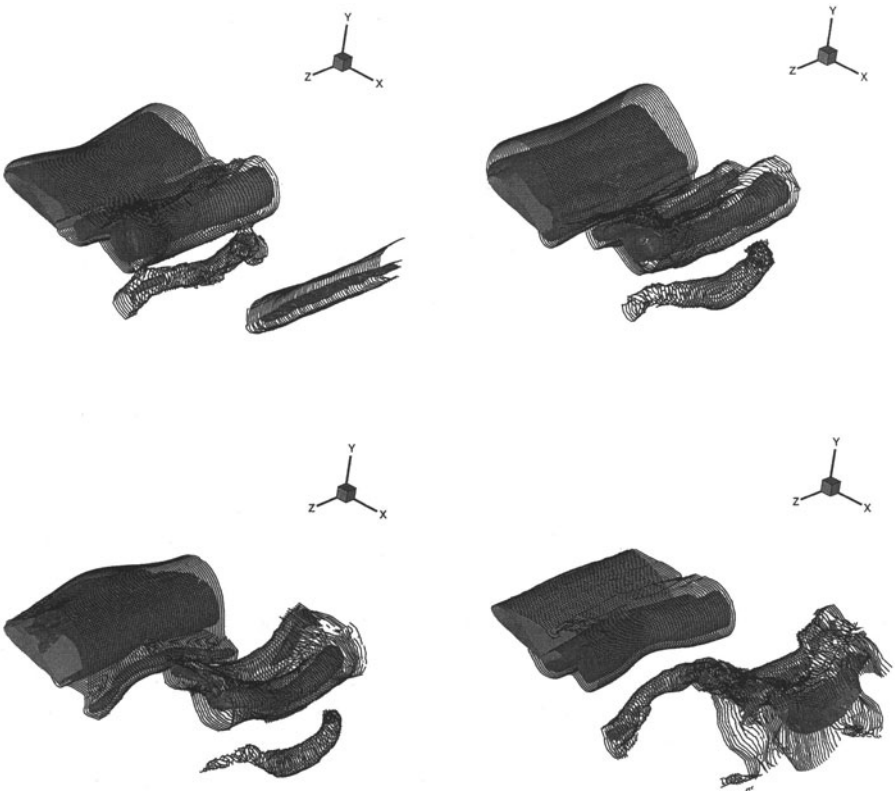


Fig. 3. Density contours from the 3D flapping wing simulation corresponding to time $t = 15.5$ (top left); 16.3 (top right); 17.5 (bottom left); 18.4 (bottom right). (The simulation with flapping started at $t = 10.36$ from a simulation with the stationary configuration and the flapping period is 4 in non-dimensional units.

Acknowledgments

This work was supported by AFOSR and DARPA and computations were done on the IBM SP2 at CFM and NPACI, and on the SGI O 2000 at NCSA.

References

1. M. Dubiner. Spectral methods on triangles and other domains. *J. Sci. Comp.*, 6:345, 1991.
2. S.J. Sherwin. Hierarchical *hp* finite elements in hybrid domains. *Finite Element in Analysis and Design*, 27:109–119, 1997.
3. T.J.R. Hughes, W.K. Liu., and T.K. Zimmerman. Lagrangian-Eulerian finite element formulation for incompressible viscous flows. *Comput. Methods Appl. Mech. Eng.*, 29:329, 1981.
4. C.S. Venkatasubban. A new finite element formulation for ALE Arbitrary Lagrangian Eulerian compressible fluid mechanics. *Int. J. Engng Sci.*, 33 (12):1743–1762, 1995.
5. R. Lohner and C. Yang. Improved ale mesh velocities for moving bodies. *Comm. Num. Meth. Eng. Phys.*, 12:599–608, 1996.
6. G. Di Battista, P. Eades, R. Tamassia, and I.G. Tollis. *Graph Drawing.* Prentice Hall, 1998.
7. G. Jiang and C.W. Shu. On a cell entropy inequality for discontinuous Galerkin methods. *Math. Comp.*, 62:531, 1994.
8. I. Lomtev, C. Quillen, and G.E. Karniadakis. Spectral/hp methods for viscous compressible flows on unstructured 2D meshes. *J. Comp. Phys.*, 144:325–357, 1998.
9. K.S. Bey, A. Patra, and J.T. Oden. hp version discontinuous Galerkin methods for hyperbolic conservation laws. *Comp. Meth. Appl. Mech. Eng.*, 133:259–286, 1996.
10. L.-W. Ho. *A Legendre spectral element method for simulation of incompressible unsteady free-surface flows.* PhD thesis, Massachustts Institute of Technology, 1989.
11. B. Koobus and C. Farhat. Second-order schemes that satisfy GCL for flow computations on dynamic grids. In *AIAA 98-0113, 36th AIAA Aerospace Sciences Meeting and Exhibit, Reno, NV*, January 12-15, 1998.
12. G.E. Karniadakis and S.J. Sherwin. *Spectral/hp Element Methods for CFD.* Oxford University Press, 1999.
13. I. Lomtev and G.E. Karniadakis. A discontinuous Galerkin method for the Navier-Stokes equations. *Int. J. Num. Meth. Fluids*, 29:587–603, 1999.
14. I. Lomtev, R.M. Kirby, and G.E. Karniadakis. A discontinuous Galerkin ALE method for viscous compressible flows in moving domains. *J. Comp. Phys.*, to appear, 1999.
15. B. Cockburn, S. Hou, and C.-W. Shu. The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: The multidimensional case. *J. Comp. Phys.*, 54:545, 1990.
16. F. Bassi and S. Rebay. A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations. *J. Comp. Phys.*, 131:267, 1997.

17. H. Liu and K. Kawachi. A numerical study of insect flight. *J. Comp. Phys.*, 146:124–156, 1998.
18. G. Karypis and V. Kumar. METIS: Unstructured graph partitioning and sparse matrix ordering system version 2.0. Technical report, Department of Computer Science, University of Minnesota, Minneapolis, MN 55455, 1995.

Discontinuous Galerkin for Hyperbolic Systems with Stiff Relaxation*

Robert B. Lowrie and Jim E. Morel

Los Alamos National Laboratory
Applied Theoretical and Computational Physics Division

Abstract. A Discontinuous Galerkin method is applied to hyperbolic systems that contain stiff relaxation terms. We demonstrate that when the relaxation time is unresolved, the method is accurate in the sense that it accurately represents the system's Chapman-Enskog approximation. Results are presented for the hyperbolic heat equation and coupled radiation-hydrodynamics.

1 Introduction

Many physical systems can be modeled by hyperbolic systems that contain relaxation terms, such as

$$\partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}) = \mathbf{s}(\mathbf{u}), \tag{1}$$

where $\mathbf{u}$, $\mathbf{f}(\mathbf{u})$, and $\mathbf{s}(\mathbf{u})$ are vectors of length p. Let τ represent the relaxation time contained in $\mathbf{s}(\mathbf{u})$. For small τ, we assume that (1) can be accurately represented by the reduced system

$$\partial_t \mathbf{v} + \partial_x \mathbf{g}(\mathbf{v}) = \tau \partial_x \left[\mathbf{D}(\mathbf{v}) \partial_x \mathbf{v}\right], \tag{2}$$

where $\mathbf{v}$ and $\mathbf{g}(\mathbf{v})$ are vectors of length q, with $q < p$, and $\mathbf{D}(\mathbf{v})$ is a $q \times q$ matrix. This reduced system is derived through a Chapman-Enskog expansion [1] and is sometimes referred to as the continuum or equilibrium-diffusion limit.

A simple example is the hyperbolic heat equation [2,3]:

$$\partial_t u + \partial_x v = 0, \tag{3a}$$

$$\partial_t v + \partial_x u = -v/\tau. \tag{3b}$$

For small τ, this system can be accurately represented by the heat equation, $\partial_t u = \tau \partial_{xx} u$. Although (3) appears simple, it presents a challenge for numerical methods. The difficulty is that for a given mesh and timestep, an accurate discretization of (1) may not asymptotically reduce to an accurate discretization of (2) as $\tau \to 0$.

* This work performed under the auspices of the U.S. Department of Energy by Los Alamos National Laboratory under Contract W-7405-ENG-36.

Methods for stiff systems have recently been an active area of research [2,4, and many others]. Such systems can be found, for example, in combustion, multi-phase flows, and rarefied gas dynamics. In addition, the equations that govern neutron and radiation transport are stiff, where Discontinuous Galerkin (DG) with linear elements has been used with good success [5,6]. In our experience, DG performs very well for other stiff systems and in this paper we summarize the results of two examples. More detailed results and discussion are presented in [8].

2 The Discontinuous Galerkin Method

In this section we give a brief outline of our implementation of DG for a single space dimension, with an emphasis on the source-term treatment. For more details, see [7,8]. Let the spatial mesh be made up of cells defined as the intervals $[x_{m-1/2}, x_{m+1/2}]$ with $\Delta x_m = x_{m+1/2} - x_{m-1/2}$. For each cell m, the interval $[x_{m-1/2}, x_{m+1/2}]$ is mapped to $\xi \in [-1, 1]$ and $\mathbf{u}$ is approximated as

$$\mathbf{u}|_m(\xi, t) = \sum_{j=0}^{k} \mathbf{U}_{m,j}(t)\phi_j(\xi).$$

The basis set $\{\phi\}$ are Lagrange-Legendre polynomials, with $\xi_0 = -1 < \xi_1 < \ldots < \xi_k < \xi_{k+1} = 1$ the Gauss-Lobatto integration points. This basis satisfies

$$\phi_i(\xi_j) = \delta_{i,j}, \tag{4}$$

so that $\mathbf{U}_{m,j}$ represents the value of the solution with respect to cell m at ξ_j.

Following the 'quadrature free' approach [7], DG(k) for Eq. (1) can be written as

$$\dot{\mathbf{U}} + \frac{1}{\Delta x_m}\mathbf{C}_B\mathbf{B} - \mathbf{C}_V\mathbf{V} = \mathbf{S}, \tag{5}$$

where the subscript m is henceforth dropped unless needed and

$$\mathbf{U} = (\mathbf{U}_0, \mathbf{U}_1, \ldots, \mathbf{U}_{k+1})^T, \qquad \mathbf{B} = (\mathbf{F}_{m-1/2}, \mathbf{F}_{m+1/2})^T,$$
$$\mathbf{S} = (\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_{k+1})^T, \qquad \mathbf{V} = (\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_{n_V})^T,$$
$$\mathbf{s}_j = \mathbf{s}(\mathbf{U}_j), \qquad \mathbf{f}_j = \mathbf{f}(\mathbf{u}(\xi_j^V, t)).$$

To account for the possible nonlinearity of $\mathbf{f}(\mathbf{u})$, its volume integral is approximated using n_V-quadrature points, at locations $\{\xi^V\}$, with $n_V \geq k+1$. The matrix $\mathbf{C}_B$ is $(k+1) \times 2$ and $\mathbf{C}_V$ is $(k+1) \times n_V$. Both of these matrices are independent of the particular element and can be computed once and stored.

The quantity $\mathbf{F}_{m+1/2} \equiv \mathbf{F}(\mathbf{U}_{m,k+1}, \mathbf{U}_{m+1,0})$ is any suitable flux function. At least in this study, the wave decomposition in the flux solver is based on

the 'frozen' flux $\mathbf{f}(\mathbf{u})$, as opposed to including any effects of the 'equilibrium' waves defined by $\mathbf{g}(\mathbf{v})$. This issue is discussed further in [8].

The time integration is carried out using a simple predictor-corrector approach. For the predictor, (5) is discretized as

$$\frac{\mathbf{U}^{n+1/2} - \mathbf{U}^n}{\Delta t/2} + \frac{1}{\Delta x_m}\mathbf{C}_B\mathbf{B}^n - \mathbf{C}_V\mathbf{V}^n = \mathbf{S}^{n+1/2},$$

where the superscript-n denotes the time-level. The corrector-step is

$$\frac{\mathbf{U}^{n+1} - \mathbf{U}^n}{\Delta t} + \frac{1}{\Delta x_m}\mathbf{C}_B\mathbf{B}^{n+1/2} - \mathbf{C}_V\mathbf{V}^{n+1/2} = \theta\mathbf{S}^{n+1} + (1-\theta)\mathbf{S}^n,$$

where $0 \leq \theta \leq 1$. We typically use $\theta = 1$ (fully implicit), which in general is only first-order in time. An alternative is $\theta = 1/2$ (Crank-Nicholson), which is nominally second-order when the relaxation time is resolved, but is well known to give oscillatory behavior if the relaxation time is unresolved. In practice, if the relaxation time is unresolved, a fully implicit treatment is sufficiently accurate since the remaining terms are treated with second-order accuracy [8].

Note that each step is *point implicit*, where by 'point' we mean a single $\mathbf{U}_j$. It is critical that the basis satisfy (4); otherwise, in general, the method would be implicit over all $(k+1)$-values of $\mathbf{U}_j$ within an element.

3 Hyperbolic Heat Equation

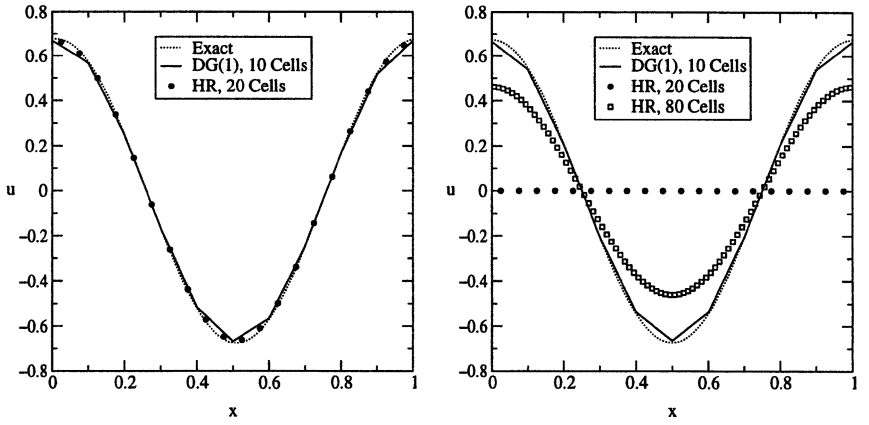
In this section we present results for (3) on the domain $x \in [0,1]$, with periodic boundary conditions, and the initial condition $u = v = \cos(2\pi x)$. For reference, we compare with a high-resolution (HR) finite-volume method that uses a central-difference slope reconstruction and the time-integrator presented in the previous section.

The DG results were run at a Courant number of 0.3 (stability limit is 1/3), while the HR results used a Courant number of 0.8 (stability limit is 1). The time-integrator used $\theta = 1/2$, although little difference was observed in the values of u with $\theta = 1$. No slope-limiting was applied in either method.

Figure 1 shows the results for two values of τ and the final time set to $0.01/\tau$. The exact total amount of damping is the same for both cases. The results show that DG(1) is fairly independent of τ, whereas for $\tau = 1 \times 10^{-5}$, the HR results are significantly over-damped, even for 80-mesh cells.

4 Radiation Hydrodynamics

The non-relativistic Euler equations of gas dynamics, coupled with a gray P_1-model of radiation transport, can be written in non-dimensional form as



(a) $\tau = 1 \times 10^{-2}$, $t = 1$ (b) $\tau = 1 \times 10^{-5}$, $t = 1000$

Fig. 1. Results for the hyperbolic heat equation. DG(1) and HR used the same time integrator.

[9,10]

$$\partial_t \rho + \partial_x(\rho v) = 0, \tag{6a}$$

$$\partial_t(\rho v) + \partial_x(\rho v^2 + p) = -\mathcal{P} S_F, \tag{6b}$$

$$\partial_t(\rho E) + \partial_x\left[(\rho E + p)v\right] = -\mathcal{P}\mathcal{C} S_E, \tag{6c}$$

$$\partial_t E_r + \mathcal{C}\partial_x F_r = \mathcal{C} S_E, \tag{6d}$$

$$\partial_t F_r + \tfrac{1}{3}\mathcal{C}\partial_x E_r = \mathcal{C} S_F, \tag{6e}$$

where ρ is the material density, v the velocity, E the total material specific energy, p the material pressure, E_r the radiation energy density, and F_r the radiation flux. The coupling terms are given by

$$S_E = \sigma_t(T^4 - E_r) + \sigma_t \tfrac{v}{\mathcal{C}}(F_r - \tfrac{4}{3}\tfrac{v}{\mathcal{C}} E_r), \tag{6f}$$

$$S_F = -\sigma_t(F_r - \tfrac{4}{3}\tfrac{v}{\mathcal{C}} E_r) + \sigma_t \tfrac{v}{\mathcal{C}}(T^4 - E_r), \tag{6g}$$

where σ_t is the flow-length scale over the photon mean-free-path and T is the temperature. There are two non-dimensional constants in this system:

$$\mathcal{C} = c/a_\infty, \qquad \mathcal{P} = a_R T_\infty^4 / \rho_\infty a_\infty^2,$$

where c is the lightspeed, a_R the radiation constant, and '∞' denotes reference conditions, with a_∞ the reference soundspeed.

For large σ_t, (6) can be approximated by

$$\partial_t \rho + \partial_x(\rho v) = 0, \tag{7a}$$

$$\partial_t(\rho v) + \partial_x(\rho v^2 + p^*) = 0, \tag{7b}$$

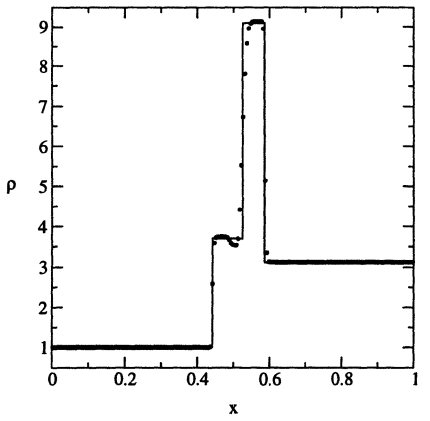
$$\partial_t(\rho E^*) + \partial_x\left[(\rho E^* + p^*)v\right] = \partial_x\left[\frac{\mathcal{P}c}{3\sigma_t}\partial_x T^4\right], \tag{7c}$$

where $p^* = p + \mathcal{P}T^4/3$ and $E^* = E + \mathcal{P}T^4/\rho$. This limit is often referred to as the *equilibrium-diffusion limit* [9,10].

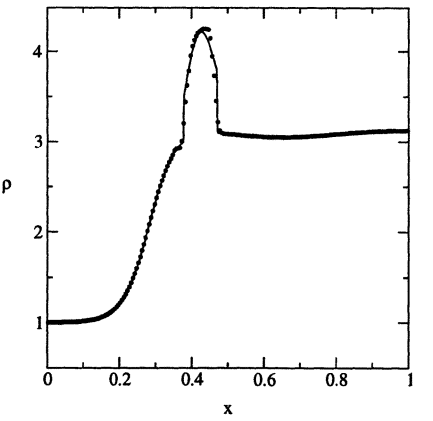
The results in Fig. 2 show the effects of σ_t on a Riemann problem. The slope limiter used for these results is described in [8]. For $\sigma_t = 0$, radiation and hydrodynamics decouple, and the exact hydrodynamic solution is two shocks moving to the left separated by a contact discontinuity. As $\sigma_t \to \infty$, the exact solution approaches a single shock moving to the left at Mach 10. Starting errors are apparent at $x = 0.4$ and $x = 0.7$ in Fig. 2c. We suspect these starting errors would be smaller if the Riemann solver included source-term effects. Nevertheless, these results are of reasonable quality for problems where standard finite-volume methods would be impractical.

References

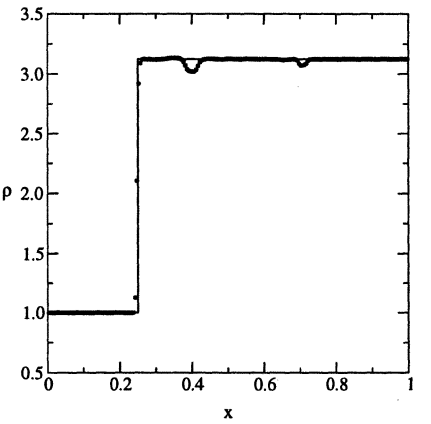
1. G.-Q. Chen, C. D. Levermore, and T.-P. Liu. Hyperbolic conservation laws with stiff relaxation terms and entropy. *Communications on Pure and Applied Mathematics*, 47:787–830, 1994.
2. M. Arora and P. L. Roe. Issues and strategies for hyperbolic problems with stiff source terms. In *Barriers and Challenges in Computational Fluid Dynamics*, volume 6 of *ICASE/LaRC Interdisciplinary Series in Science and Engineering*, pages 139–154. Kluwer, 1998.
3. J. Hittinger. Model problem for hyperbolic systems with relaxation. Preprint, 1999.
4. S. Jin. Runge-Kutta methods for hyperbolic conservation laws with stiff relaxation. *Journal of Computational Physics*, 122:51–67, 1995.
5. E. Larsen and J. Morel. Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes II. *Journal of Computational Physics*, 83 (1):212–236, 1989.
6. J. E. Morel, T. A. Wareing, and K. Smith. A linear-discontinuous spatial differencing scheme for S_n radiative transfer calculations. *Journal of Computational Physics*, 128:445–462, 1996.
7. H. Atkins and C. W. Shu. Quadrature-free implementation of discontinuous Galerkin method for hyperbolic-equations. *AIAA Journal*, 36(5):775–782, 1998.
8. R. Lowrie and J. Morel. Discontinuous Galerkin for stiff hyperbolic systems. In *Proceedings of the 14th AIAA Computational Fluid Dynamics Conference*, 1999. Paper 99-3307.
9. D. Mihalas and B. W. Mihalas. *Foundations of Radiation Hydrodynamics*. Oxford University Press, 1984.
10. R. Lowrie, J. Morel, and J. Hittinger. Coupling radiation and hydrodynamics. *Astrophysical Journal*, 521(1), 1999.



(a) $\sigma_t = 0$



(b) $\sigma_t = 100$



(c) $\sigma_t = 1 \times 10^8$

Fig. 2. DG(1) results for a Riemann problem in radiation hydrodynamics, $t = 0.05$, Courant number 0.3, $\theta = 1$, $\gamma = 5/3$, $\mathcal{P} = 44.930910$, $\mathcal{C} = 100$. The initial condition is $(\rho, u, p) = (1.0, 0.0, 0.6)$ for $x < 0.75$ and $(\rho, u, p) = (3.122689, -6.797632, 2.874136)$ for $x > 0.75$, with E_r and F_r set to their equilibrium values. Symbols are the cell averages of the 200-cell solution. In each plot, the solid line is the following: a) exact solution, b) 2000-cell solution (mesh converged to plotting scale), c) exact solution for $\sigma_t \to \infty$.

Finite Element Output Bounds for Parabolic Equations: Application to Heat Conduction Problems

Luc Machiels

Department of Mechanical Engineering, M.I.T., 77 Mass. Ave., Cambridge, MA, 02139, USA

Abstract. We present a Neumann–subproblem *a posteriori* finite element procedure for the efficient calculation of rigorous, constant–free, sharp lower and upper estimators for linear functional outputs of parabolic equations discretized by a discontinuous Galerkin method in time. We first formulate the bound procedure; we then provide illustrative numerical examples for problems of unsteady heat conduction.

1 Introduction

Classical finite element design problems are often solved iteratively based on repeated appeal to the numerical solution procedure for different design parameters. Therefore, the finite element solution method for the partial differential equations must be sufficiently inexpensive to permit numerous evaluations, yet sufficiently fine to demonstrably represent the true performance of the system. A recent series of papers [9–12] introduces a new *a posteriori* error control strategy which reconciles these two conflicting requirements. The method considerably generalizes earlier techniques in that we obtain quantitative constant–free bounds — contrary to earlier explicit techniques [3–5] — for the output of interest — contrary to earlier implicit techniques [1,2,6]. To date, the method has been successfully applied to a variety of problems; for a review, see [8].

We propose here an extension of the method to treat parabolic problems. The method, based on a discontinuous Galerkin (dG) finite element discretization in time [4], may be viewed as an implicit Aubin-Nitsche construction. Two finite element meshes are used in space: a global coarse mesh and a decoupled fine mesh. A unique fine mesh discretizes the time interval. A "classical" hybridization technique [6] permits to compute the estimators in terms of solutions of spatially *local* Neumann subproblems computed at each time step. The local subproblems are also decoupled in time. In this work, we only consider the evaluation of spatial discretization errors; the extension of the method to assess time discretization errors uses two meshes in time (a coarse mesh and a fine mesh).

2 Problem Statement

Given $T > 0$, and given two Hilbert spaces, X, of norm $\|\cdot\|_X$, and Y, of norm $\|\cdot\|$, chosen such that $Y \subset X \subset Y'$, where Y' denotes the dual of Y, we consider the following problem: given $f \in L^2(0,T;Y')$ and $u_0 \in Y$, find $u \in L^2(0,T;Y) \cap C^0([0,T];X)$ such that $du/dt \in L^2(0,T;Y')$ and

$$\left\langle \frac{du}{dt}(t), v \right\rangle + a(t;u(t),v) = \langle f(t), v\rangle \quad \forall v \in Y. \tag{1}$$

Here $u(0) = u_0$, $\langle\cdot,\cdot\rangle$ denotes the duality pairing between Y' and Y, and, $\forall t \in [0,T]$, $a(t;v,w) : Y \times Y \to \mathbb{R}$ is a bilinear form satisfying:

1. $\exists M > 0$ such that $|a(t;v,w)| \leqslant M\|v\|\|w\|$, a.e. $t \in [0,T], \forall v,w \in Y$,
2. $\exists \alpha > 0$ such that $a(t;v,v) \geqslant \alpha\|v\|^2$, a.e. $t \in [0,T], \forall v \in Y$.

We now define the mesh $0 = t_0 < t_1 < \cdots < t_n < t_{n+1} < \cdots < t_N = T$, the mesh diameter $\tau = \max(t_{n+1} - t_n)$, and the intervals $I_n =]t_n, t_{n+1}[$, for $n = 0,1,\ldots,N-1$. We then introduce the spaces, $\mathbb{P}_q(I_n;Y) = \{v : I_n \to Y \,|\, v(t) = \sum_{s=0}^{q} v_s t^s, v_s \in Y\}$, and the spaces $V_q(Y) = \{v \in L^2(0,T;Y) \,|\, v_{|I_n} \in \mathbb{P}_q(I_n;Y)$, for all $n = 0,1,\ldots,N-1\}$. We also define $v^>(t_n) = \lim_{s\to 0+} v(t_n + s)$ and $v^<(t_n) = \lim_{s\to 0-} v(t_n + s)$, and the jumps $[v]_n = v^>(t_n) - v^<(t_n)$. The dG(q)-method for Eq. (1) defines an approximate solution $u_\tau \in V_q(Y)$ iteratively for $n = 0,1,\ldots,N-1$ by setting $[u_\tau]_0 = u_\tau^>(0) - u_0$ and

$$\int_{I_n} \left[\left\langle \frac{du_\tau(t)}{dt}, v(t) \right\rangle + a(t;u_\tau(t),v(t)) \right] dt = \int_{I_n} \langle f(t), v(t)\rangle \, dt - ([u_\tau]_n, v(t_n))_X, \quad \forall v \in \mathbb{P}_q(I_n;Y).$$

Finally, we define a linear functional, the output of interest $S : L^2(0,T;Y) \to \mathbb{R}$, as $S(v) = \int_0^T \langle \ell(t), v(t)\rangle \, dt$, where $\ell \in L^2(0,T;Y')$. In this work, we assume that the mesh diameter is sufficiently small such that u_τ is a good approximation in the sense that $s \approx s_\tau = S(u_\tau)$ with reasonable certainty. Here $s = S(u)$ denotes the exact output.

3 Error Bound Formulation

We assume that we are given two finite element subspaces: Y_H, the coarse space, and Y_h, the fine space; we require $Y_H \subset Y_h \subset Y$. (We also assume $u_0 \in Y_H \subset Y_h$.) The associated coarse–space and fine–space approximations, $u_{\tau,H}$ and $u_{\tau,h}$, exhibit complementary advantages and disadvantages. The fine space solution, $u_{\tau,h} \in V_q(Y_h)$, yields a very good approximation, $S(u_{\tau,h})$, of the exact output s; nevertheless, the computational effort required to obtain $u_{\tau,h}$ will typically be prohibitive. In contrast, the coarse-space solution,

$u_{\tau,H} \in V_q(Y_H)$, can be obtained with relatively modest computational effort; nevertheless the fidelity of the corresponding approximate output, $S(u_{\tau,H})$, is no longer assured.

Prior to the description of the bound procedure, we define $a^s(t;v,w) = \frac{1}{2}(a(t;v,w) + a(t;w,v))$, the symmetric part of $a(t;v,w)$, for all $t \in [0,T]$. Finally, we introduce the "broken" spaces, $\hat{Y}_H \subset \hat{Y}_h \subset \hat{Y}$, such that $Y = \{v \in \hat{Y} \,|\, B(v,q) = 0, \forall q \in Z\}$, $Y_h = \{v \in \hat{Y}_h \,|\, B(v,q) = 0, \forall q \in Z\}$, and $Y_H = \{v \in \hat{Y}_H \,|\, B(v,q) = 0, \forall q \in Z_H\}$. Here $Z_H \subset Z$ are additional Hilbert spaces and $B : \hat{Y} \times Z \to \mathbb{R}$ is a bilinear form. (See [7–12] for examples of particular instantiations of the broken spaces and of the form B).

The bound procedure proceeds in five steps:

1. Compute $u_{\tau,H} \in V_q(Y_H)$ as the solution of the primal problem

$$\int_{I_n} \left[\left\langle \frac{du_{\tau,H}(t)}{dt}, v(t) \right\rangle + a(t; u_{\tau,H}(t), v(t)) \right] dt = \int_{I_n} \langle f(t), v(t) \rangle \, dt - ([u_{\tau,H}]_n, v(t_n))_X, \quad \forall v \in \mathbb{P}_q(I_n; Y_H),$$

for all $n = 0, 1, \ldots, N-1$. We denote the corresponding primal residuals by $\mathcal{R}^{\mathrm{pr},n}(v)$, for all $v \in \mathbb{P}_q(I_n; \hat{Y})$ and for all $n = 0, 1, \ldots, N-1$.

2. Compute $\psi_{\tau,H} \in V_q(Y_H)$ as the solution of the dual problem

$$\int_{I_n} \left[- \left\langle v(t), \frac{d\psi_{\tau,H}(t)}{dt} \right\rangle + a(t; v(t), \psi_{\tau,H}) \right] dt = - \int_{I_n} \langle \ell(t), v(t) \rangle \, dt + ([\psi_{\tau,H}]_{n+1}, v(t_{n+1}))_X, \quad \forall v \in \mathbb{P}_q(I_n; Y_H),$$

for all $n = 0, 1, \ldots, N-1$, and where we take $\psi^{>}_{\tau,H}(t_N) = 0$. We denote the corresponding dual residuals by $\mathcal{R}^{\mathrm{du},n}(v)$, for all $v \in \mathbb{P}_q(I_n; \hat{Y})$ and for all $n = 0, 1, \ldots, N-1$.

3. Compute the hybrid fluxes, $\sigma^{\mathrm{pr}}_{\tau,H} \in V_q(Z_H)$ and $\sigma^{\mathrm{du}}_{\tau,H} \in V_q(Z_H)$, which satisfy the equations

$$\int_{I_n} B(v(t), \sigma^{\mathrm{pr}}_H(t)) \, dt = \mathcal{R}^{\mathrm{pr},n}(v), \quad \forall v \in \mathbb{P}_q(I_n, \hat{Y}_H),$$

$$\int_{I_n} B(v(t), \sigma^{\mathrm{du}}_H(t)) \, dt = \mathcal{R}^{\mathrm{du},n}(v), \quad \forall v \in \mathbb{P}_q(I_n, \hat{Y}_H),$$

for all $n = 0, 1, \ldots, N-1$.

4. Compute the "reconstructed errors", $\hat{e}^{\mathrm{pr}} \in V_q(\hat{Y}_h)$ and $\hat{e}^{\mathrm{du}} \in V_q(\hat{Y}_h)$,

$$2 \int_{I_n} a^s(t; \hat{e}^{\mathrm{pr}}(t), v(t)) \, dt = \mathcal{R}^{\mathrm{pr},n}(v) - \int_{I_n} B(v(t), \sigma^{\mathrm{pr}}_H(t)) \, dt,$$

$$2 \int_{I_n} a^s(t; \hat{e}^{\mathrm{du}}(t), v(t)) \, dt = \mathcal{R}^{\mathrm{du},n}(v) - \int_{I_n} B(v(t), \sigma^{\mathrm{du}}_H(t)) \, dt,$$

for all $v \in \mathbb{P}_q(I_n, \hat{Y}_h)$ and for all $n = 0, 1, \ldots, N-1$.

5. Evaluate the lower and upper bound approximations:

$$s_H^{\pm} = S(u_{\tau,H}) \pm \int_0^T \kappa(t) a^s(t; \hat{e}^{\pm}(t), \hat{e}^{\pm}(t))\, dt,$$

where $\kappa(t)$ is a positive piecewise constant function, $\kappa(t) = \kappa^n$ for $t \in I_n$, with $\kappa^n > 0$, and $\hat{e}^{\pm}(t) = \hat{e}^{\mathrm{pr}}(t) \mp \hat{e}^{\mathrm{du}}(t)/\kappa(t)$.

Since $\hat{e}^{\mathrm{pr}}$ and $\hat{e}^{\mathrm{du}}$ do not depend on $\kappa(t)$, we can find, following [12], the function $\kappa(t)$ which minimizes the bound gap $\Delta_H = \frac{1}{2}(s_H^+ - s_H^-)$. Note also that the bound gap permit local (elemental) decomposition suitable for adaptive subsequent refinements [12].

Finally, it can be shown [7] that the bounds satisfy the following error expression

$$\begin{aligned} s_H^{\pm} = S(u_{\tau,h}) &\pm \int_0^T \kappa(t) a^s(t; \hat{e}^{\pm}(t) - e(t), \hat{e}^{\pm}(t) - e(t))\, dt, \\ &\pm \frac{1}{2} \sum_{n=0}^{N-1} \kappa^n (e^{>}(t_n) - e^{<}(t_n), e^{>}(t_n) - e^{<}(t_n))_X \\ &\mp \frac{1}{2} \sum_{n=0}^{N-2} (\kappa^{n+1} - \kappa^n)(e^{<}(t_{n+1}), e^{<}(t_{n+1}))_X \pm \frac{1}{2} \kappa^{N-1} (e^{<}(T), e^{<}(T))_X. \end{aligned}$$

If $\kappa^n = \kappa^{n+1}$, for $n = 0, \ldots, N-2$, then $s_H^+ \geq s_h$ and $s_H^- \leq s_h$, for all H, which is the desired bounding property. If we choose $\kappa(t)$ such that $\kappa^n \neq \kappa^{n+1}$ — a choice which arguably yields sharper bounds — the indefinite term $\mp(\kappa^{n+1} - \kappa^n)(e^{<}(t_{n+1}), e^{<}(t_{n+1}))_X$ remains. The essential point is that, for the class of problem we will consider, this indefinite term involves a weak norm ($\|\cdot\|_X$) which converges faster to zero than the definite term $\int_0^T \kappa(t) a^s(t; \hat{e}^{\pm}(t) - e(t), \hat{e}^{\pm}(t) - e(t))\, dt$. Therefore, for H sufficiently small ($H < H^*$), the definite term dominates the deviations of $s_H^{\pm}$ from s_h. We conclude that s_H^- and s_H^+ approach s_h from below and above, respectively.

4 Numerical Examples

We have obtained numerical results for the equation

$$\frac{\partial u}{\partial t} - \nabla \cdot (\nu \nabla u) = f \quad \text{in } (0,T) \times \Omega,$$

with $u = 0$ on $(0,T) \times \partial\Omega$ and $u|_{t=0} = u_0$ on Ω, in a simple domain $\Omega =]0,1[\times]0,1[$, and for $\mathbb{P}_1$–in–time and $\mathbb{P}_2$–in–space discretizations. The initial condition is $u_0 = 0$, the boundary conditions are homogeneous Dirichlet $u|_{\partial\Omega} = 0$, $\nu(\mathbf{x}) = 1$, $f(\mathbf{x}, t) = \sin(\pi t)$, $T = 2$, and the time step is taken uniform, $\tau = 0.02$. The output functional is given by $S(u) = \int_0^T \int_A u(\mathbf{x}, t)\, dA$,

where $A =]0.25, 0.75[\times]0.25, 0.75[$. For this problem, we have observed optimal convergence rate of the bound gap $\Delta_H \approx O(H^4)$ [7].

We next present results for an unsteady heat conduction problem in a composite material. The domain of the problem is represented in Fig. 1(a). The material is characterized by two different thermal conductivity: $\nu = 0.01$ inside the rectangles, and $\nu = 1$ in the remainder of the domain. The boundary conditions are homogeneous Neumann, $\frac{\partial u}{\partial n} = 0$, on the left and right sides of the domain, homogeneous Dirichlet, $u = 0$, on the top boundary, and we impose a time varying heat flux on the bottom boundary Γ_b, $\frac{\partial u}{\partial n} = \sin(\pi t) + 1$. The initial condition is $u_0 = 0$, $T = 2$, and the time step is taken uniform, $\tau = 0.04$. The isocontours of u are represented in Fig. 1(b) at $t \approx 1/2$ and Fig. 1(c) at $t \approx 3/2$. The output of interest is the mean temperature on Γ_b, $S(u) = \int_0^T \int_{\Gamma_b} u(\mathbf{x}, t)\, ds\, dt$.

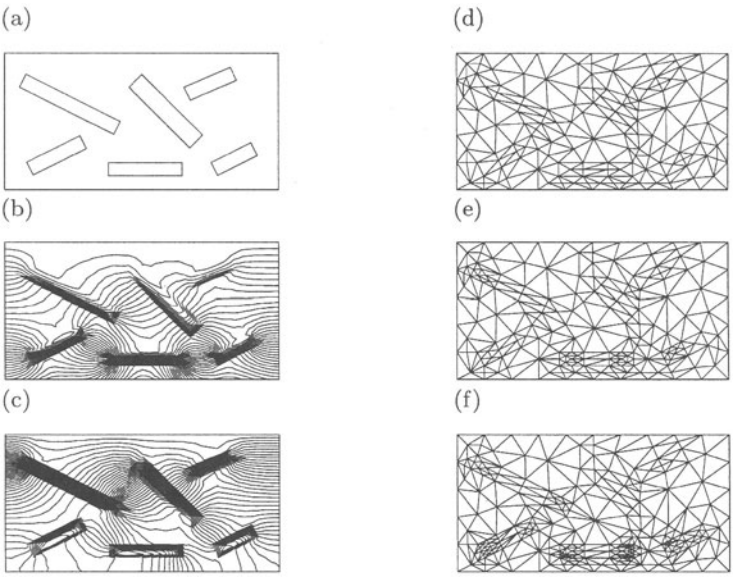


Fig. 1. (a) Domain. (b) Isolines of the solution $u_{\tau,H}$ at time $t \approx 1/2$. (c) Isolines of the solution $u_{\tau,H}$ at time $t \approx 3/2$. (d) Mesh $\mathcal{T}_H$. (e) Mesh $\mathcal{T}_{H'}$. (f) Mesh $\mathcal{T}_{H''}$.

In this example an automatic adaptive strategy is used which effectiveness is summarized in Table 1, in which $\mathcal{T}_H$, $\mathcal{T}_{H'}$, and $\mathcal{T}_{H''}$ denote the successive adapted meshes corresponding to Figures 1(d), (e) and (f), respectively.

This work was supported by NASA Grants NAG1-1978, NAG1-1587, and NAG4-105, DARPA and ONR Grant N00014-91-J-1-1889, and AFOSR Grant F49620-97-1-0052. L.M. was partially supported by Fulbright and BAEF Fel-

	$\mathcal{T}_H$	$\mathcal{T}_{H'}$	$\mathcal{T}_{H''}$
# of el.	286	340	462
s_H^-	12.85	12.92	13.14
s_H^+	14.65	13.80	13.47
Δ_H	0.9	0.44	0.165

Table 1. Adaptive refinement.

lowships. We would like to acknowledge our very fruitful collaboration with Prof. A. T. Patera and Prof. J. Peraire of M.I.T.

References

1. M. Ainsworth and J. T. Oden. A posteriori error estimation in finite element analysis. *Comp. Meth. in Appl. Mech. and Engrg.*, 142:1–88, 1997.
2. R. E. Bank and A. Weiser. Some a posteriori error estimators for elliptic partial differential equations. *Math. Comp.*, 44(170):283–301, 1985.
3. R. Becker and R. Rannacher. A feedback approach to error control in finite element method: Basic analysis and examples. *East-West J. Numer. Math.*, 4:237–264, 1996.
4. K. Eriksson and C. Johnson. Adaptive finite element methods for parabolic problems, I: a linear model problem. *SIAM J. Numer. Anal.*, 28(1):43–77, 1991.
5. D. Estep. A posteriori error bounds and global error control for approximation of ordinary differential equations. *SIAM J. Numer. Anal.*, 32(1):1–48, 1995.
6. P. Ladeveze and D. Leguillon. Error estimation procedures in the finite element method and applications. *SIAM J. Numer. Anal.*, 20:485–509, 1983.
7. L. Machiels. A posteriori finite element bounds for output functionals of discontinuous Galerkin discretizations of parabolic problems. Comp. Meth. in Appl. Mech. and Engrg., submitted.
8. L. Machiels, J. Peraire, and A. T. Patera. Output bound approximations for partial differential equations; application to the incompressible Navier–Stokes equations. In S. Biringen, editor, *Proceedings of the Istanbul Workshop on Industrial and Environmental Applications of Direct and Large Eddy Numerical Simulation.* Springer-Verlag, August 1998.
9. Y. Maday, A. T. Patera, and J. Peraire. A general formulation for a posteriori bounds for output functionals of partial differential equations; application to the eigenvalue problem. C.R. Acad. Sci. Paris, to appear.
10. M. Paraschivoiu and A. T. Patera. A hierarchical duality approach to bounds for the outputs of partial differential equation. *Comp. Meth. in Appl. Mech. and Engrg.*, 158:389–407, 1998.
11. M. Paraschivoiu, J. Peraire, and A. T. Patera. A posteriori finite element bounds for linear-functional outputs of elliptic partial differential equations. *Comp. Meth. in Appl. Mech. and Engrg.*, 150:289–312, 1997.
12. J. Peraire and A. T. Patera. Bounds for linear-functional outputs of coercive partial differential equations: local indictors and adaptive refinement. In P. Ladeveze and J. T. Oden, editors, *On New Advances in Adaptive Computational Methods in Mechanics*, 1997.

3D Unstructured Mesh ALE Hydrodynamics with the Upwind Discontinuous Galerkin Method *

Manoj K. Prasad, Jose L. Milovich, Aleksei I. Shestakov, David S. Kershaw, and Michael J. Shaw

Lawrence Livermore National Laboratory, Livermore, CA 94550, USA

Abstract. We describe a numerical scheme to solve 3D Arbitrary Lagrangian-Eulerian (ALE) hydrodynamics on an unstructured mesh using a discontinuous Galerkin method (DGM) and an explicit Runge-Kutta time discretization. Upwinding is achieved through Roe's linearized Riemann solver with the Harten-Hyman entropy fix. For stabilization, a 3D quadratic programming generalization of van Leer's 1D minmod slope limiter is used along with a Lapidus type artificial viscosity. This DGM scheme has been tested on a variety of hydrodynamic test problems and appears to be robust making it the basis for the integrated 3D inertial confinement fusion modeling code ICF3D. For efficient code development, we use C++ object oriented programming to easily separate the complexities of an unstructured mesh from the basic physics modules. ICF3D is fully parallelized using domain decomposition and the MPI message passing library. It is fully portable. It runs on uniprocessor workstations and massively parallel platforms with distributed and shared memory.

1 The ALE Hydrodynamics Equations

The motion of a compressible fluid is described by Euler's equations along with an equation of state (EOS). In an ALE code the computational mesh x_i, where $i = 1, 2, 3$ describes the 3D space, can move in time t:

$$x_i = x_i(x_i^0, t),\ x_i(t = 0) = x_i^0,\ \frac{\partial x_i}{\partial t} = V_i^g$$

where x_i^0 are the "Lagrangian" coordinates and V_i^g is an arbitrarily specified grid velocity. Euler's equations stated in conservation form, follow the time evolution of ρ (mass density), ρv_i (momentum density) and ρE (total energy density). In general, the fluid can be subjected to a body force per unit mass G_i (such as gravitational acceleration). In the Lagrangian frame the fluid equations are (in what follows summation over repeated indices is assumed)

$$\frac{\partial A_\alpha^g}{\partial t} + \frac{\partial F_{\alpha i}^g}{\partial x_i^0} = S_\alpha^g$$

* Work performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under contract number W-7405-ENG-48.

where A^g_α are the conservative state variables with corresponding fluxes $F^g_{\alpha i}$ and source terms S^g_α (α runs from 1 to 5):

$$A^g_\alpha \equiv J^g A_\alpha,\ F^g_{\alpha i} \equiv F_{\alpha j}\eta_{ji},\ S^g_\alpha \equiv J^g S_\alpha$$

$$A_\alpha \equiv \begin{pmatrix} \rho \\ \rho v_1 \\ \rho v_2 \\ \rho v_3 \\ \rho E \end{pmatrix},\ F_{\alpha j} \equiv \begin{pmatrix} \rho(v_j - V^g_j) \\ \rho v_1(v_j - V^g_j) + P\delta_{j1} \\ \rho v_2(v_j - V^g_j) + P\delta_{j2} \\ \rho v_3(v_j - V^g_j) + P\delta_{j3} \\ \rho E(v_j - V^g_j) + P v_j \end{pmatrix},\ S_\alpha \equiv \begin{pmatrix} 0 \\ \rho G_1 \\ \rho G_2 \\ \rho G_3 \\ \rho v_j G_j \end{pmatrix}$$

$$J^g_{ij} \equiv \frac{\partial x_j}{\partial x^0_i}, \eta_{ji} \equiv J^g J^{g\,-1}_{ji} = \frac{1}{2}\epsilon_{jkl}\epsilon_{imn}\frac{\partial x_k}{\partial x^0_m}\frac{\partial x_l}{\partial x^0_n}, J^g \equiv |J^g_{ij}| = \frac{1}{3}J_{ij}\eta_{ji}$$

where $E \equiv I + v_i v_i/2$ is the total energy per unit mass, $I \equiv I(\rho, P)$ is the internal energy per unit mass and P is the pressure. The EOS gives the explicit form for $I(\rho, P)$.

The choice $V^g_i = 0$ leads to an "Eulerian" code where the mesh is fixed in time while $V^g_i = v_i$ leads to a "Lagrangian" code where the mesh follows the fluid. Lagrangian codes have two basic advantages over the Eulerian version. First, material interfaces are exactly resolved. Secondly, the implementation of boundary conditions (specified normal velocity or pressure) is much simpler compared to the moving boundary problems inherent to the Eulerian codes.

2 The DGM Solution of the ALE Equations

We discretize the problem domain into an arbitrary and in general unstructured set of 3D linear finite elements (tetrahedrons, pyramids, prisms, and hexahedrons). Within each element K we use a piecewise (tri)linear approximation for the coordinates x_i, the body force per unit mass G_i, and the "primitive" variables $B_\alpha \equiv (\rho, v_1, v_2, v_3, P)$:

$$\{x_i(\mathbf{x}^0, t), G_i(\mathbf{x}^0, t), B_\alpha(\mathbf{x}^0, t)\} \approx \sum_{n=1}^{n_v}\{x_{in}(t), G_{in}(t), B_{\alpha n}(t)\}\phi_n(\xi)$$

where $\phi_n(\xi)$ are the 3D finite element (tri)linear basis functions which are equal to 1 at node n and 0 at all other nodes and $n_v = (4, 5, 6, 8)$ is the number of nodes in element K=(tets,pyramids,prisms,hexes). We are using isoparametric elements and ξ are the 3D isoparametric reference coordinates. It is important to note that because we are using a piecewise linear approximation for B_α there is no requirement of continuity across faces between neighboring elements, i.e. the name "discontinuous" finite elements. However we do require the coordinates x_i to be continuous across faces

To determine the vertex values $B_{\alpha n}(t)$ we use a DGM whereby we take moments of the ALE hydro equations with the n_v basis functions $\phi_n(\xi)$ in

each element K (the DGM equations):

$$\partial_t \int_K \phi_n(\xi) A_\alpha d^3x = \int_K \phi_n(\xi) S_\alpha d^3x + \int_K F_{\alpha i} \frac{\partial \phi_n(\xi)}{\partial x_i} d^3x - \int_{\partial K} N_i F_{\alpha i} \phi_n(\xi) d\Gamma$$

where $d^3x \equiv J^g d^3x^0$, we have integrated the flux term by parts, and ∂K denotes the surface of element K with outward unit length normal N_i. An essential aspect of our discretization is that it manifestly preserves the conservation properties of the continuum field equations since $\sum_{n=1}^{n_v} \phi_n(\xi) = 1$. All integrals on the right hand side of the DGM equation are computed numerically using Gaussian quadrature. The surface integral poses a problem as to what to use for $N_i F_{\alpha i}$ since the primitive variables are in general discontinuous across a face. We resolve this ambiguity by interpreting each point on the face as a 1D Riemann initial-value (shock tube) problem. We then solve this Riemann problem by Roe's characteristic decomposition along with Roe averaging generalized to arbitrary EOS:

$$N_i F_{\alpha i}^{Roe} \equiv \frac{1}{2}\{N_i F_{\alpha i}^- + N_i F_{\alpha i}^+ + [R^* sign(\Lambda^*) R^{*-1}]_{\alpha\beta}[N_i F_{\beta i}^- - N_i F_{\beta i}^+]\}$$

At each point on a face of an element we have two values of the normal component of the flux: $N_i F_{\alpha i}^+$ and $N_i F_{\alpha i}^-$ corresponding to the two sides of the face, $+$ denoting the side from which the outward pointing normal N_i emanates. The Roe average state (denoted with $*$) of the $+$ and $-$ states is defined through the following equations:

$$N_i F_{\alpha i}^- - N_i F_{\alpha i}^+ \equiv J_{\alpha\beta}^*(A_\beta^- - A_\beta^+),\ J^* \equiv R^* \Lambda^* R^{*-1},\ \Lambda_{\alpha\beta}^* \equiv \lambda_\alpha^* \delta_{\alpha\beta}$$

where $\lambda_\alpha(\mathbf{v}) \equiv \{\lambda_p, \lambda_m, \lambda_o, \lambda_o, \lambda_o\}, \lambda_o = N_i(v_i - V_i^g), \lambda_p = \lambda_o + c, \lambda_m = \lambda_o - c, \lambda_\alpha^* \equiv \lambda_\alpha(\mathbf{v}^*)$ and c is the adiabatic sound speed. At *boundary* faces we use an imaginary "ghost" state on the outside such that the Roe Riemann solver for the ghost and interior states will give the desired boundary condition (e.g. specified normal velocity, pressure). A well known defficiency of the Roe flux is its inability to properly identify an expansion fan containing a sonic point. To correct this we use the Harten-Hyman entropy fix that consists of modifying $|\lambda_\alpha^*|$ as follows:

$$|\lambda_\alpha^*| \rightarrow |\lambda_\alpha^*| + max(0, \epsilon - |\lambda_\alpha^*|),\ \epsilon \equiv max(0, \lambda_\alpha^* - \lambda_\alpha^-, \lambda_\alpha^+ - \lambda_\alpha^*),\ \lambda_\alpha^\pm \equiv \lambda_\alpha(\mathbf{v}^\pm)$$

At this point the DGM equations reduce to a system of n_v ordinary differential equations (ODE) in time for the n_v moments $M_{\alpha n} \equiv \int_K \phi_n(\xi) A_\alpha d^3x$ for which we use an explicit time-adapting second order Runge-Kutta integration. To implement this we require, first to compute the primitive state variables B_α from the moments M_α and second to determine Δt for stability of the numerical scheme. We compute $B_{\alpha n}$ from $M_{\alpha n}$ by first introducing auxiliary variables $\tilde{A}_\alpha$ which are linear in each element K and such that its n_v moments are equal to $M_{\alpha n}$. In each cell, a linear representation of the primitive variables B_α, is obtained by using a first order Taylor series expansion of

B_α in $\tilde{A}_\alpha$ around the average values $< \tilde{A}_\alpha >=< A_\alpha > \equiv \int_K A_\alpha d^3x / \int_K d^3x$. For numerical stability of the time integration we require a Courant-type time step control: Δt must be smaller than the time it takes for a wave originating on a face of an element to cross it. We heuristically implement this requirement as follows: $\Delta t \leq CFL\ t_{Courant}$ where $CFL \approx .3$ [1] and

$$t_{Courant} \equiv \min_K \frac{\int_K d^3x}{\int_{\partial K} \max(|\lambda_p^*|, |\lambda_m^*|, |\lambda_o^*|) d\Gamma}$$

3 3D Shock Stabilization

In second order schemes the values of the primitive variables $B_{\alpha n}$ may develop local maxima and minima behind discontinuities due to dispersive truncation errors. These violations of physical stability constraints can usually be controlled by a slope limiting technique which modifies or "stabilize" the nodal values $B_{\alpha n}$. To this end we have generalized VanLeer's 1D minmod slope limiter to an unstructured 3D mesh through a quadratic programming formulation. The central idea is to require each nodal value $B_{\alpha n}$, within an element, to be bounded by the minimum and maximum of the average values $< B_\alpha >$ of all elements surrounding the node n. In general this will not be true. To satisfy this requirement, we replace $B_{\alpha n} with B'_{\alpha n}$ obtained from a least squares formulation *subject to the constraint:* $< B'_\alpha >=< B_\alpha >$ so that conservation is not affected. In addition, in order to keep our scheme as second order accurate as possible, we construct within each element a *hybrid* primitive state variable:

$$B_{\alpha n}^{hybrid} \equiv (1-s)(rB'_{\alpha n} + (1-r)B_{\alpha n}^{Godunov}) + sB_{\alpha n}, \quad B_{\alpha n}^{Godunov} \equiv < B_\alpha >$$

where r measures the strength of the shock using pressure ratios ($0 \leq r \leq 1$, $r \approx 0$ near a very strong shock), and s measures the adiabaticity of the solution by examining the viscous shock heating rate: $(\Delta I - P\Delta\rho/\rho^2)/\Delta t$ ($0 \leq s \leq 1$, $s \approx 1$ for small entropy production rates) [3]. In cartesian geometry a simple way to monitor shock regions is to compute where $< -\nabla_i v_i >$ is positive (implying compression) using $\int_K -\nabla_i v_i d^3x = -\int_{\partial K} N_i v_i^* d\Gamma$ where v_i^* is the Roe average velocity on the face of element K. Once $B_{\alpha n}^{hybrid}$ are obtained, the "stabilized" moments $M_{\alpha n}^{hybrid}$ are constructed by the inverse of the algorithm used to derive the primitive variables from the moments. $M_{\alpha n}^{hybrid}$ are then used in the next Runge-Kutta iteration.

While the *hybrid* stabilization works well for a variety of problems involving shocks there are cases where it is insufficient. This has led us to implement a Lapidus type artificial viscosity, i.e. adding a source term of the type $\nabla(D^L \nabla A_\alpha)$ to the hydro equations. We have empirically found the following implementation of the Lapidus flux correction to work well - in the vicinity of shocks add the following source term to the right hand side of the

DGM equations for *interior* faces:

$$\frac{D^L}{l}[< A_\alpha^- >_n - < A_\alpha^+ >_n] \int_{\partial K} \phi_n(\xi) d\Gamma, \quad < A_\alpha^\pm >_n \equiv \frac{\int_{K\pm} \phi_n(\xi) A_\alpha^\pm d^3x}{\int_{K\pm} \phi_n(\xi) d^3x},$$

where D^L is an artificial diffusion coefficient that vanishes in the continuum limit, $\pm$ refers to the two elements at either side of the face, and l is some length scale across the face (e.g. distance between centers of elements + and −). We *define* the Lapidus diffusion coefficient through:

$$\frac{D^L}{l} \equiv \kappa \lambda_K^* \quad , \quad \lambda_K^* \equiv \frac{\int_{\partial K} \max(|\lambda_p^*|, |\lambda_m^*|, |\lambda_o^*|) d\Gamma}{\int_{\partial K} d\Gamma}$$

where κ is a dimensionless adjustable parameter around 0.3 (for stability of the explicit scheme) and λ_K^* is a characteristic wave speed on the face.

4 3D ALE Grid Velocity

The grid velocity V_i^g at node n is arbitrary in an ALE code. If $V_i^g = 0$ everywhere in the mesh one has an Eulerian code. A "Lagrangian" code would require $V_i^g = v_i$. However, v_i is discontinuous across faces while we require V_i^g to be continuous. We therefore formulate an *almost* "Lagrangian" code by using a "least squares" estimate of the fluid velocity at node n for V_i^g:

$$\text{Minimize} \sum_{\{f_n\}} [N_{if_n}^g V_{in}^g - Y_{f_n}]^2, \quad Y_{f_n} \equiv N_{if_n}^g V_{if_n}^g,$$

where the sum is over the set $\{f_n\}$ of all faces with vertex n. Y_{f_n} is determined by requiring the first component of the Roe flux, the mass flux, to vanish, i.e. $N_i F_{1i}^{Roe}[Y_{f_n}] = 0$ for interior faces. For boundary faces Y_{f_n} is set to either the prescribed normal component of the velocity or is determined from the specified boundary pressure through the vanishing of the mass flux. While this procedure determines V_{in}^g fully, there are instances when constraining the grid velocity may prove beneficial in avoiding mesh tangling. Therefore, we have implemented a set of either $n_c = 1, 2, 3$ linear constraints on the components of V_{in}^g. If $n_c = 3$, V_{in}^g is completely determined (e.g. a center symmetry node may be required not to move).

5 Hydrodynamic Test Problems and the ICF3D Code

Details of the DGM algorithm, as described above, can be found in [2] and [3]. The algorithm has been implemented in C++ using an object oriented (OO) approach. The OO design allows us to untangle the complexities of the unstructured mesh data structure from the basic physics algorithm modules.

Thus cells (tets,hexes..) and faces belong in separate classes and calculations such as flux integrals become virtual functions using pointers to access neccessary data. This design enables efficient code development.

We have tested the hydro code on a suite of problems relevant to inertial confinement fusion (ICF). Some of these problems are cylindrically or spherically symmetric. We have therefore extended our hydro code, to 3D cylindrical and spherical geometries [2], and used it in situations where symmetry is critically important. These problems have exact analytical solutions that can be checked against the computed ones. Our comparisons have shown extremely good agreement. For example, it is shown in [3] that the computation of linear growth rates for 2D and 3D Rayleigh-Taylor instability falls within 1% of the analytical value. We now discuss two further examples.

The Sedov point explosion problem has a self-similar analytical solution as described in [7]. The initial conditions are a constant density ρ_0 ideal γ law cold gas (pressure $P = 0$) with an instantaneous release of energy E_0 at some point so that $\rho E = E_0 \delta^3(\boldsymbol{x})$. An infinite strength expanding shock is generated with 3D spherical radius $r_s = k(E_0 t^2/\rho_0)^{1/5}$ and velocity $U_s = dr_s/dt = (2/5)r_s/t$ at time t where k is a dimensionless constant (for $\gamma = 5/3$: $k = 1.151667$). Behind the shock the density $\rho \rightarrow \rho_0(\gamma+1)/(\gamma-1)$ and pressure $P \rightarrow 2\rho_0 U_s^2/(\gamma+1)$ as $r \rightarrow r_s$. We simulate this problem using 3D cylindrical geometry (r, θ, z) with 45 x 1 x 45 mesh of hexahedrons covering $0 \leq r, z \leq 1.125$ and $0 \leq \theta \leq 2\pi$. We choose $\rho_0 = 1$, $E_0 = .4935889$, and run the code in (almost) Lagrangian mode up to time $t = 1$ with boundary conditions of vanishing normal velocity. In Fig.1 shows the mesh (for $\theta = 0$) and normalized density ρ/ρ_s and pressure P/P_s versus r (for $z = \theta = 0$) at $t = 1$. The mesh shows good spherical symmetry though there is a tendency for the mesh to tangle near the center where the density is extremely low. The density and pressure are in very good agreement with the analytical solution [7].

Our final example is that of an unstructured mesh computation of a spherical implosion in 3D cartesian (x, y, z) geomtery. We model an icosahedral wedge domain bounded by: a sphere of radius $r = 1$, 2 azimuthal planes $(\phi = \pm\pi/5)$, and a plane through the origin and $(\theta, \phi) = (cos^{-1}1/\sqrt{5}), \pm\pi/5)$. The unstructured mesh is generated using LaGrit [8] with 50 radial cells resulting in 28,208 tetrahedra, 5791 nodes, and 58,455 faces. We run the code in multiprocessor mode using domain decomposition by METIS [9]. The initial conditions are: ideal $\gamma = 5/3$ law gas with density $\rho_0 = 1$, pressure and velocity equal to zero. The code is run in (almost) Lagrangian mode with boundary conditions: boundary pressure $P^{bndry} = 4/3$ on the outer radial surface, $v_\theta = v_\phi = o$ on the planar boundary faces, and $v_i = 0$ at the origin. This problem has no known analytical solution so we run a comparison (structured mesh) spherical geometry (r, θ, ϕ) problem using 200 x 1 x 1 hexes covering: $0 \leq r \leq 1$, $0 \leq \theta\pi/2$, and $0 \leq \phi \leq \pi/4$. Fig.2 shows a side-on view of the density contours for the unstructured mesh calculation at time $t = .6$ along with density plots versus spherical radius r for the comparison

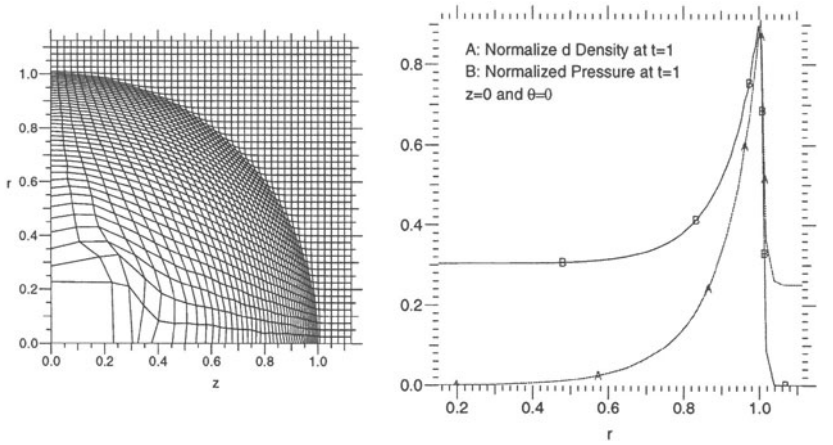


Fig. 1. Sedov point explosion problem: Plots of mesh (r versus z for θ=0), normalized density ρ/ρ_s and pressure P/P_s versus r for z=θ=0 at time $t = 1$

spherical geometry run at various times and with various resolutions to verify convergence of the answer. We see that the unstructured mesh calculation in cartesian geometry is in very good agreement with the comparison run. At time $t \approx .57$ the incoming shock reflects off the origin and at time $t = .6$ the reflected shock is at (spherical) radius $r = .09$ with the outer boundary at $r = .56$. The shock position, maximum and minimum density, and outer radius are all in excellent agreement. Fig.2 clearly shows the code's ability to maintain spherical symmetry, despite running on the asymmetric tetrahedral grid.

The success in simulating a wide span of test problems has made the hydro code the basis of the fully integrated three-dimensional ICF modeling code [4], ICF3D. Besides multiple material hydro, the physics modules of ICF3D include diffusive radiation and heat conduction transport, laser ray tracing, and realistic EOS.

The DGM hydro algorithm is ideally suited for parallelization. Indeed all physics modules of ICF3D have been parallelized [6] using domain decomposition and the MPI message passing library.

6 Open Problems

The *hybrid* 3D shock stabilization we employ is applied to each of the primitive variables, in particular each of the components of the 3D velocity vector. We have noticed that for problems with symmetry the relation between the cartesian velocity components (e.g. for cylindrical symmetry the azimuthal and axial velocity components vanish) may be destroyed by the hybrid stabilization. It is our conjecture, that understanding hybrid stabilization as the

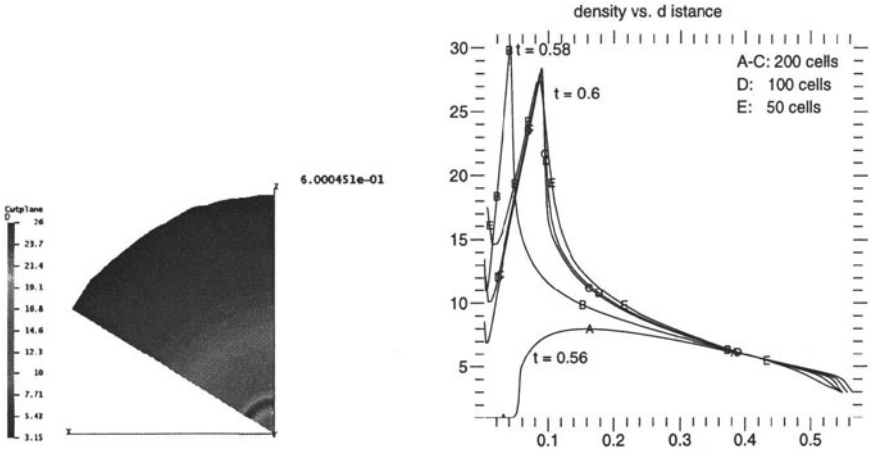


Fig. 2. Unstructured mesh implosion problem: Plots of side-on view of density contours for the unstructured mesh computation at time t=.6, density versus radius plots for the (structured mesh) spherical geometry comparison run with various resolutions and times.

discretization of a continuum "artificial viscosity" operator could shed some light in devising a stabilization scheme that would maintain the symmetry requirements. Until we better understand this issue, we have instead used our hydro code, in 3D cylindrical and spherical geometries [2], for situations where cylindrical or spherical symmetry is critically important.

We close with some remarks on the computational costs of the DGM algorithm described here. In comparison to traditional structured mesh finite volume methods, our scheme involves significantly (almost an order of magnitude) more number crunching and memory use. This cost should be balanced with the unstructured mesh ALE capability and ease of parallelization through domain decomposition. The DGM scheme described here can be cost effective for problems where accuracy on irregular meshes is desired with as few grid cells for a given resolution.

References

1. B. Cockburn and C. W. Shu, Math. Comput. **52**, 411 (1989).
2. D. S. Kershaw, M. K. Prasad and M. J. Shaw, 3D Unstructured Mesh ALE Hydrodynamics with the Upwind Discontinuous Finite Element Method. Lawrence Livermore National Laboratory report UCRL-JC-122104 (1995).
3. D. S. Kershaw, M. K. Prasad, M. J. Shaw and J. L. Milovich: 3D Unstructured Mesh ALE Hydrodynamics with the Upwind Discontinuous Finite Element Method. Comput. Methods Appl. Mech. and Engrg, **158**, 81-116 (1998).
4. A. I. Shestakov, M. K. Prasad, J. L. Milovich, N. A. Gentile, J. F. Painter and G. Furnish: The Radiation-Hydrodynamic ICF3D Code. to appear in Comput. Methods Appl. Mech. and Engrg, (1999).

5. A. I. Shestakov, J. L. Milovich and M. K. Prasad: ICF3D Results on Hydrodynamic Test Problems. Lawrence Livermore National Laboratory internal report (1997).
6. A. I. Shestakov, J. L. Milovich and D. S. Kershaw: Parallelization of a 3D Unstructured-grid, Laser Fusion Design Code. SIAM NEWS, News journal of the Society for Industrial and Applied Mathematics, **32**, 3, p. 6, April 1999.
7. L. D. Landau and E. M. Lifschitz: Fluid Mechanics, Pergamon Press, New York (1959).
8. Lagrit: http://www.t12.lanl.gov/~ lagrit.
9. METIS: http://www-users.cs.umn.edu/~ karypis/metis.

Some Remarks on the Accuracy of a Discontinuous Galerkin Method

P. Rasetarinera[1], M.Y. Hussaini[1], and F.Q. Hu[2]

[1] Program in Computational Science and Engineering,
The Florida State University, Tallahassee, FL 32306-4120
[2] Department of Mathematics and Statistics,
Old Dominion University, Norfolk, VA 23681

1 Introduction

The discontinuous Galerkin methods have recently found increasing applications in computational fluid dynamics because of their robustness and other practical features. The key feature that distinguishes the discontinuous spectral Galerkin method from its traditional counterpart is that the basis functions in each element are independent of the basis functions in the contiguous elements. The method is thus compact, and it easily accommodates any boundary conditions and complex geometry.

A Fourier analysis of the semi-discrete discontinuous Galerkin method applied to wave propagation problems is provided in the work of Hu, Hussaini and Rasetarinera [1], where background literature on the subject is cited. It was shown therein that the dispersion relation and the dissipation rate of the method depends on the formula employed to evaluate the flux at the interface between adjacent elements. In the case of the scalar advection equation specifically, an upwind formula for the interface flux was found to produce larger dissipation error relative to the dispersion error. For the centered flux, the dissipation rate is exactly zero, but the range of wave numbers for which the discrete dispersion relation accurately approximates the exact one is relatively small.

The present work examines further the accuracy of the upwind and centered schemes for wave propagation problems. Numerical results are presented to support the analysis.

2 Error Analysis

For the propagation of a plane wave $e^{i(kx-\omega t)}$, the approximate solution u_h of a scalar advection equation obtained by a N^{th} order discontinuous Galerkin scheme on a uniform mesh of size h is a superposition of N traveling waves

$$u_h(x,t) = \sum_{l=0}^{N-1} \hat{C}_l e^{i(knh-\omega_l t)} v(\xi) \; ; \; x = nh + h\xi, \; \xi \in [0,1] \qquad (1)$$

where $\hat{\mathbf{C}}_l e^{i(knh-\omega_l t)}$ represents the vector of expansion coefficients of the l^{th} wave in the n^{th} element and $v(\xi) = (v_0(\xi), \cdots, v_{N-1}(\xi))$ are the local basis functions [1]. In (1), the numerical frequency ω_0 of the first wave approximates the exact frequency ω. This first wave is called the physical mode.

The approximation error in an element in the L_2 norm is

$$||u_{exact} - u_h(x,t)||_2 = ||e^{i(kh\xi-\omega t)} - \sum_{l=0}^{N-1} e^{-i\omega_l t}\hat{\mathbf{C}}_l v(\xi)||_2. \tag{2}$$

This error can be decomposed into three parts. By the triangle inequality, the right hand side of (2) is bounded by

$$||u_{exact} - u_h(x,t)||_2 \leq ||e^{ikh\xi} - \hat{\mathbf{C}}_0 v(\xi)||_2 + ||(e^{-i\omega_0 t} - e^{-i\omega t})\hat{\mathbf{C}}_0 v(\xi)||_2$$
$$+ ||\sum_{l=1}^{N-1} e^{-i\omega_l t}\hat{\mathbf{C}}_l v(\xi)||_2. \tag{3}$$

The first term of (3) represents the approximation error associated with the initial condition. Specifically, this is the error due to the approximation of the initial condition in the physical mode. The second term represents the evolution error in the physical mode. The third term is the error due to the non-physical modes.

First we examine the evolution error, where the quantity of interest is $|\omega - \omega_0|$. For the centered scheme, it represents the phase error. For the upwind scheme it combines both dispersion and dissipation errors. Figure 1 shows the behavior of this error for wave numbers ranging from 0 to 2π. The convergence rate of the evolution error for the upwind scheme is $\mathcal{O}((kh)^{2N})$ but it is non-uniform for the centered scheme. For wave numbers between 0 and π, even-order centered schemes exhibit slower convergence rates than the same order upwind schemes. The convergence is faster for the higher wave numbers. On the other hand, odd-order centered schemes converge faster for low wave numbers and slower for high wave numbers.

To quantify the resolution of the discontinuous Galerkin method for an error tolerance ϵ, let k_{max} be the highest wave number such that

$$|\omega - \omega_0| < \epsilon. \tag{4}$$

Figure 2 shows the resolution of the scheme in terms of the number of points per wave length (computed as $2\pi N/(k_{max}h)$) for the numerical flux

$$\mathbf{F}(u_1, u_2) = \frac{1}{2}(u_1 + u_2 + \alpha(u_1 - u_2)), \quad 0 \leq \alpha \leq 1.$$

Figure 2 indicates that the fourth order centered scheme is optimal among schemes of order less than seven for $0.001 \leq \epsilon \leq 0.0025$. For a given wave number, it requires fewer points or degrees of freedom to satisfy (4) than the fifth or sixth order schemes.

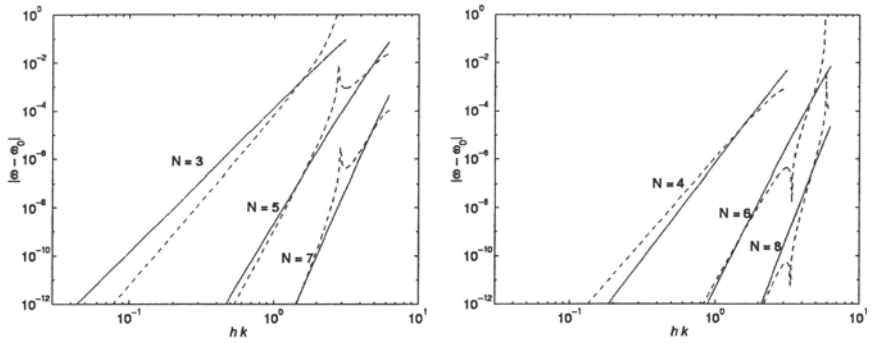


Fig. 1. Convergence of the physical mode. Even-order methods ($N = 4, 6, 8$) (left), odd order methods ($N = 3, 5, 7$) (right). - - - centered flux, — upwind flux.

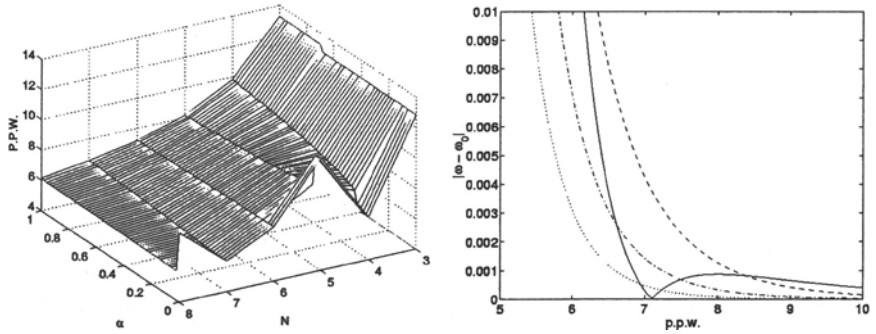


Fig. 2. Resolution in term of points per wave length (p.p.w.) for an error tolerance of 0.001. — 4^{th} order centered, - - - $5^t h$ order upwind, – . –6^{th} order upwind, . . . 7^{th} order centered

Next, we consider the approximation error of the initial condition. The behavior of the error is presented in Figure 3. For the upwind scheme, the convergence rate of this projection error is $\mathcal{O}((kh)^N)$. For the centered scheme, the projection error follows the same non-uniform trend as the evolution error. Figure 4 shows that the projection error dominates the convergence rate of both the upwind and the centered schemes at low wave numbers. For high wave numbers, the evolution error dominates. The reasons are excessive damping in the case of the upwind scheme and accumulation of the phase error in the case of the centered scheme.

Finally, we look at the error due to the non-physical modes. The behavior of the error due to the non-physical modes for the fourth order scheme at $t = 20\pi/\omega$ is displayed in Figure 4. For the centered scheme there is no damping at any wave number but a small upwinding is sufficient to damp the non-

physical modes at low wave numbers. This damping decreases significantly as the wave number increases.

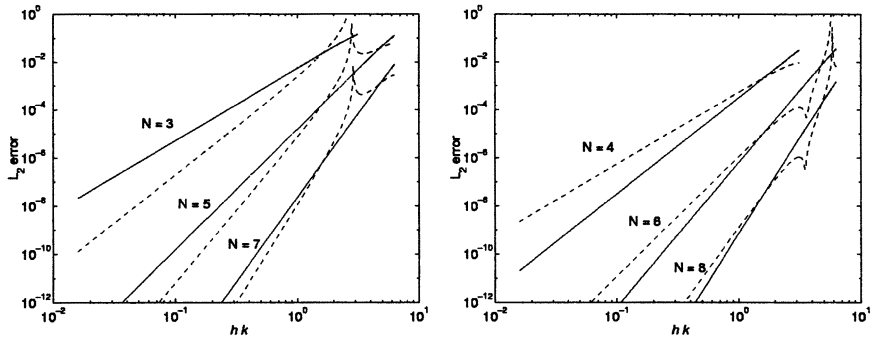


Fig. 3. L^2 errors of the initial projection onto the accurate mode. Even order methods ($N = 4, 6, 8$) (left), odd order methods ($N = 3, 5, 7$) (right).- - - centered flux, — upwind flux

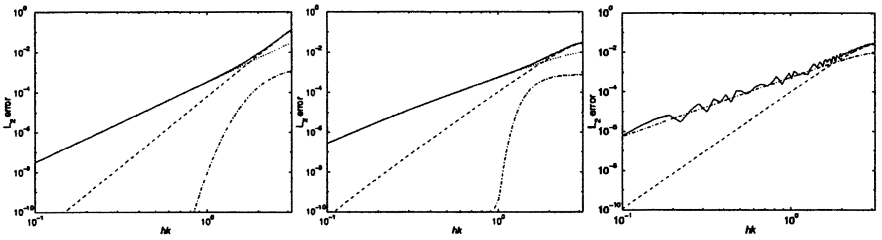


Fig. 4. L^2 errors at $t = 20\pi/\omega$ for the upwind (left), slightly upwind ($\alpha = 0.1$) (center) and the centered (right) fourth order schemes. — global error, - - - physical mode error, – . – non-physical modes error, ... projection error

3 Numerical results

We now present numerical results to validate the previous analysis. The spherical wave equation

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial r} + \frac{u}{r} = 0$$

is solved with the initial condition $u = 0$ over the domain $5 \leq r \leq 450$, which is discretized with 100 elements. A boundary condition $u = sin(\pi t/4)$ is applied at $r = 5$ and a third order Runge-Kutta scheme is used for the time discretization.

Given the conclusions of the previous section, we solve this problem with fourth order schemes at a resolution of 7.19 points per wave length. Figure 5

shows three solutions obtained at $t = 300$ near the wave front for $\alpha = 0, 0.1$ and 1.0. Note that the upwind solution is heavily damped. The solutions obtained with the centered scheme and with the slightly upwind ($\alpha = 0.1$) scheme show small dispersion and dissipation errors. At the wave front of all three solutions, we can observe the Gibbs phenomenon resulting from the discontinuity in the first derivative. Note that the oscillations are more pronounced with the upwind scheme. Ahead of the wave front, we observe non-physical modes from the centered scheme which are damped when using the upwind schemes ($\alpha = 0.1 \text{or} 1.0$).

The conclusion on the efficiency of the fourth order centered scheme is borne out in Figure 6 where the fourth order centered scheme is compared to the sixth order upwind and centered schemes. The same resolution is used for all three schemes.

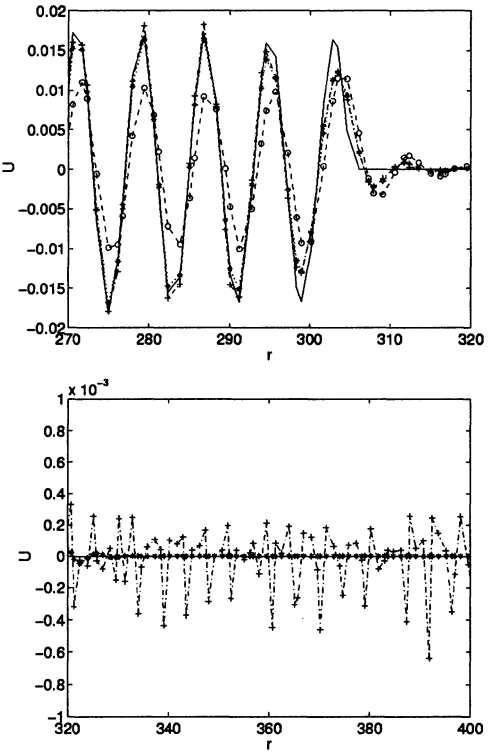


Fig. 5. Fourth order scheme, comparison between exact and approximate solution at $t = 300$. Wave front (top), non-physical modes (bottom). (+) centered, (o) upwind with $\alpha = 1.0$, (*) upwind with $\alpha = 0.1$, — exact

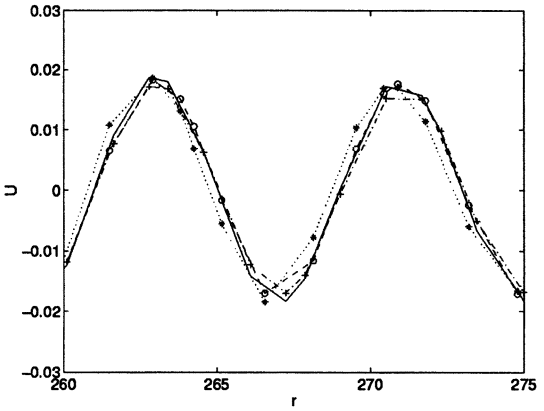


Fig. 6. Comparison between the fourth order centered scheme (+), and the sixth order centered (*) and fully upwind (o) schemes. — exact solution

4 Conclusions

In this work, we have shown that the centered fourth order scheme has an optimal resolution for an error tolerance in the approximate interval $[0.001, 0.0025]$ on the dispersion relation. For high wave numbers, the convergence of the discontinuous Galerkin method is governed by the evolution error. For low wave numbers, the dominant error comes from the projection of the initial condition. For upwind schemes the convergence rate of the evolution error is $\mathcal{O}((kh)^{2N})$ while the convergence rate of the projection error is $\mathcal{O}((kh)^N)$. For centered schemes the convergence rate is non-uniform. The convergence of even-order centered schemes is slower than the same order upwind schemes for wave numbers between 0 and π but it is faster for higher wave numbers. On the other hand, odd-order centered schemes converge faster for low wave numbers and slower for high wave numbers.

References

1. Hu, F.Q., Hussaini, M. Y., Rasetarinera, P.: An analysis of the discontinuous Galerkin method for wave propagation problems. Journal of Computational Physics **151** (1999) 921–946

Coupling Continuous and Discontinuous Techniques: An Adaptive Approach

Mirko Sardella

Dip. di Matematica, Politecnico, Corso Duca degli Abruzzi 24, 10129 Torino, Italy.

Abstract. In [S1], a method for the numerical approximation of singularly perturbed convection diffusion problems was introduced. In this note, we will show an a posteriori error estimate for this method.

1 Introduction

Recently, space-discontinuous techniques such as the Local Discontinuous Galerkin method (see, *e.g.*, [CS]) or the finite volume method ([CMmC]) have been extensively used in the approximation of convection-dominated flows. This is a consequence of their properties of conservation and flexibility. On the other hand, the large number of degrees of freedom makes these methods inefficient for diffusion dominated flows. Thus, in the context of convection-diffusion problems, it seems quite natural to construct a scheme exploiting the good features of these techniques in the convective terms, while taking advantage of the good properties of continuous finite elements in the discretization of the diffusive phenomena.

In [S1], we introduced a new scheme based on a coupling of finite elements and finite volumes: we consider the usual Galerkin discretization of the problem, but we modify the convective part according to the finite volume node-centered approach. The resulting scheme preserves conservation, fulfills consistency and realizes a stability property similar to that achieved with the SUPG technique. In this note, we present an ***a posteriori*** error estimate for such a method together with a numerical test problem.

2 The boundary value problem and the method

Let Ω be an open and bounded subset of $\mathbb{R}^2$ with polygonal boundary $\partial\Omega$. We consider the following singularly perturbed convection-diffusion problem: find u such that

$$\begin{cases} -\nu\Delta u + \nabla\cdot(\boldsymbol{a}u) + cu = f & \text{in } \Omega\ , \\ u = 0 & \text{on } \Gamma_D\ , \\ \dfrac{\partial u}{\partial \boldsymbol{n}} = 0 & \text{on } \Gamma_N\ , \end{cases} \tag{1}$$

where Γ_D and Γ_N are two open and disjoint subsets of $\partial\Omega$ such that $\bar{\Gamma}_D \cup \bar{\Gamma}_N = \partial\Omega$ and $\boldsymbol{n}$ is the outward normal vector to $\partial\Omega$. Moreover, we assume:

1. $\nu \in \mathbb{R}$, $\boldsymbol{a} \in \left(W^{1,\infty}(\Omega)\right)^2$, $c \in L^\infty(\Omega)$ and $f \in L^2(\Omega)$.
2. $\nu > 0$, $\boldsymbol{a} \cdot \boldsymbol{n} \geq 0$ a.e. on Γ_N and there exists a constant κ, satisfying $\kappa > 0$ if $\Gamma_D = \emptyset$ and $\kappa \geq 0$ if $\Gamma_D \neq \emptyset$, such that $\frac{1}{2}\nabla \cdot \boldsymbol{a} + c \geq \kappa$ a.e. in Ω.

We indicate the inner product in $L^2(\Omega)$ by $(\cdot,\cdot)$, while $\|\cdot\|_s$ and $|\cdot|_s$ stand for the $H^s(\Omega)$-norm and -seminorm, respectively. Existence and uniqueness of the weak solution of (1) follow from the Lax-Milgram Lemma.

For the discretization, we assume that for some $h_0 > 0$ there exists a family $\{\mathcal{T}_h\}_{h \in (0,h_0]}$ of admissible triangulations of Ω satisfying the following conditions:

1. $\mathcal{T}_h$ is a quasi-uniform triangulation, for each $0 < h \leq h_0$,
2. Any $P \in \bar{\Gamma}_D \cap \bar{\Gamma}_N$ is a node of $\mathcal{T}_h$.

We fix a numbering $(P_i)_{i \in I}$ of the nodes of $\mathcal{T}_h$. For all $i \in I$, we denote by γ_i the set of triangles $T \in \mathcal{T}_h$ containing the vertex P_i, and by ω_i the set of indices $j \in I$ such that $j \neq i$ and P_j is a vertex of a triangle $T \in \gamma_i$. If we fix an interior point Z_T in every $T \in \mathcal{T}_h$, we can construct the *Finite Volume* dual mesh $\mathcal{V}_h$ by associating to each vertex P_i a control volume $V_i \in \mathcal{V}_h$ as follows. Join each point Z_T $(T \in \gamma_i)$ to the points $Z_{T'}$ of the triangles $T' \in \gamma_i$ having a side in common with T. If $P_i \in \Omega$, the polygonal domain surrounding P_i obtained in this way is the control volume V_i. If $P_i \in \partial\Omega$, then we complete the contour by connecting the middle point of each boundary side of any $T \in \gamma_i$ with the points P_i and Z_T. Let P_i, P_j be two vertices of a triangle $T \in \mathcal{T}_h$ and let $V_i, V_j \in \mathcal{V}_h$ the associated control volumes. We set $l_{ij} := \partial V_i \cap \partial V_j$ and $l_i^b := \partial V_i \cap \partial\Omega$ (if not empty); note that both l_{ij} and l_i^b are segments or unions of segments. Moreover, we denote by $\boldsymbol{n}_i$ the outward normal to ∂V_i and by $\boldsymbol{n}_{ij}$ its restriction to l_{ij}. Then we make the following assumptions on the dual mesh $\mathcal{V}_h$:

1. If $|l|$ is the usual Lebesgue measure of l, and h_T is the diameter of T, we assume: $|l_{ij}| \sim h_T \sim |l_i^b|$.
2. If $\boldsymbol{v}_{ij} := P_j - P_i$ and $\boldsymbol{e}_{ij} := \frac{\boldsymbol{v}_{ij}}{\|\boldsymbol{v}_{ij}\|}$, the angle $\sigma_{ij} := \widehat{\boldsymbol{e}_{ij}\boldsymbol{n}_{ij}}$ satisfies: $|\sigma_{ij}| \leq \frac{\theta_0}{3}$, where θ_0 is the smallest angle in the triangulation.

We define $X_h := \{v \in C^0(\bar{\Omega})\colon\ v \in \mathcal{P}_1(T)$ for every $T \in \mathcal{T}_h\}$ and $W_h := \{w \in L^2(\Omega)\colon\ w \in \mathcal{P}_0(V)$ for every $V \in \mathcal{V}_h\}$, where $\mathcal{P}_k(D)$ is the space of polynomial functions of degree less than or equal to k on the domain D. Moreover, we introduce the projection operator $\mathcal{S}_h : C^0(\bar{\Omega}) \longrightarrow W_h$ s.t. $\psi \mapsto \mathcal{S}_h(\psi) := \sum_{V \in \mathcal{V}_h} \psi(P_V)\chi_V$, where P_V is the node belonging to $V \in \mathcal{V}_h$, and χ_V is the characteristic function of the set V.

We are now ready to introduce the method. For each $u_h \in X_h$ and each l_{ij}, we define

$$u_{ij}^h := \Big(\frac{1}{2} + \lambda_{ij}(\boldsymbol{a}_h \cdot \boldsymbol{n}_{ij})\Big)u_h(P_i) + \Big(\frac{1}{2} - \lambda_{ij}(\boldsymbol{a}_h \cdot \boldsymbol{n}_{ij})\Big)u_h(P_j)\ . \qquad (2)$$

Here, a_h is the constant approximation of a on l_{ij} defined as follows: $\int_{l_{ij}} a_h \cdot n_{ij} = \int_{l_{ij}} a \cdot n_{ij}$. In addition, λ_{ij} is a "weight" parameter satisfying: $\lambda_{ij} = \lambda_{ji}$, and $\lambda_{ij} \geq 0$. Note that (2) actually defines a sort of "averaged upwind" approximation; indeed u_{ij}^h can be rewritten as:

$$u_{ij}^h = \frac{u_h(P_i) + u_h(P_j)}{2} + \lambda_{ij}(a_h \cdot n_{ij})|P_i - P_j| \frac{u_h(P_i) - u_h(P_j)}{|P_i - P_j|} \ .$$

Next, set

$$b_h(u_h, v_h) := \sum_{i \in I} v_h(P_i) \Big(\sum_{j \in \omega_i} \int_{l_{ij}} (a_h \cdot n_{ij})\, u_{ij}^h \, dx + \int_{l_i^b} (a_h \cdot n_{ij})\, u_h \, dx \Big) \ ,$$

and $c_h(u_h, v_h) := (\tilde{c}_h\, S_h(u_h), S_h(v_h))$, where $\tilde{c}_h$ is the piecewise constant approximation of c on each volume V_i defined by $\tilde{c}_h := \frac{1}{|V_i|} \int_{V_i} c\, dx$.

We consider the discrete problem: find $u_h \in X_h$ such that

$$\nu(\nabla u_h, \nabla v_h) + b_h(u_h, v_h) + c_h(u_h, v_h) = (f, S_h(v_h)) \ , \quad \forall v_h \in X_h \ . \tag{3}$$

In [S1] we proved the existence and uniqueness of the approximate solution u_h, togheter with its property of conservation and an a priori error estimate. In particular, if we set $\lambda_T := \min_{l_{ij} \cap T \neq \emptyset} \lambda_{ij}$, we proved the following Theorem. [1].

Theorem 1. *(Stability) There exists a constant $\beta > 0$ and a vector field a_h^* constant on each $T \in \mathcal{T}_h$, such that, if we define*

$$\|v\|_{1,h}^2 := \nu \|\nabla v\|_0^2 + \beta \sum_{T \in \mathcal{T}_h} h_T \lambda_T \|a_h^* \cdot \nabla v\|_{0,T}^2 + \kappa \|v\|_0^2 \ , \quad \forall v \in H_0^1(\Omega) \ ,$$

then, if $\kappa > 0$, the following estimate holds:

$$\|u_h\|_{1,h}^2 \preceq \frac{1}{\kappa} \|f\|_0^2 \ .$$

Moreover, if $\Gamma_D \neq \emptyset$ (which is the case if $\kappa = 0$), then

$$\|u_h\|_{1,h}^2 \preceq \frac{1}{\nu} \|f\|_0^2 \ .$$

The constant β depends on θ_0, but it is independent of the mesh size h. The vector field a_h^* satisfies $\|a - a_h^*\|_{\infty,T} = O(h)$, for each $T \in \mathcal{T}_h$; moreover, if a is piecewise constant, one can choose $a_h^* = a$.

[1] Given two functions N_1 and N_2 we use the notations $N_1 \preceq N_2$ if there exists a strictly positive constant C such that $N_1 \leq C N_2$.

3 A Posteriori Error Estimate

We present an *a posteriori* error estimate for (3). The error estimators yield global upper and local lower bounds of the error measured in the energy norm: $|||u|||^2 := \nu|u|_1^2 + \|u\|_0^2$ or $|||u|||^2 := \nu|u|_1^2$ in the case $\|c\|_\infty = 0$. For the sake of brevity, we omit the details of the proofs, which are contained in [S2]. Moreover, for the sake of simplicity, we assume a to be piecewise constant.

We need some preliminary notations. For any $T \in \mathcal{T}_h$, we denote by $\mathcal{E}(T)$ the set of the edges of T which does not intersect Γ_D. For any $l \in \mathcal{E}(T)$, we denote by $[\nabla\phi \cdot n]_l$ the jump of $\nabla\phi \cdot n$ across l in the direction n (it is independent of the orientation of n). Moreover, we define $k_1 := \max(1, \|c\|_\infty)$, and $k_2 := \|a\|_\infty \nu^{-1/2}$. Finally, we set $\alpha_T := \min(1, h_T\nu^{-1/2})$ and $\alpha_l := \min(1, |l|\nu^{-1/2})$ or $\alpha_T := h_T\nu^{-1/2}$ and $\alpha_l := |l|\nu^{-1/2}$, in the case $\|c\|_\infty = 0$.

Theorem 2. *Let f_h be an arbitrary approximation of f. Set:*

$$\eta_{R,T} := \|f_h - \nabla\cdot(au_h) - cu_h\|_{0,T} \ ,$$

$$\eta_{R,l} := \|[-\nu\nabla u_h \cdot n_l]\|_{0,l} \ ,$$

$$\eta_{1,S}^2 := \sum_{l_{ij}\cap T\neq\emptyset} \Big(\int_{l_{ij}} (a\cdot n_{ij})\,\nabla u_h\cdot(P_m - P)\,dP\Big)^2 \ ,$$

$$\eta_{2,S}^2 := \sum_{l_{ij}\cap T\neq\emptyset} \Big(|l_{ij}|^{-1/2}\int_{l_{ij}} \lambda_{ij}(a\cdot n_{ij})^2\,\nabla u_h\cdot(P_i - P_j)\,dP\Big)^2 \ ,$$

$$\eta_{3,S}^2 := (h_T\alpha_T)^{-2} \sum_{V_i\cap T\neq\emptyset} \Big(\int_{V_i\cap T} c\,\nabla u_h\cdot(P - P_i)\,dP\Big)^2 \ ,$$

$$\eta_{S,T}^2 := \eta_{1,S}^2 + \eta_{2,S}^2 + \eta_{3,S}^2 \ ,$$

where, for any l_{ij}, we denote by P_m the middle point between P_i and P_j. Then:

$$\Big\{\sum_{T\in\mathcal{T}_h} \frac{\alpha_T^2}{(k_1+k_2\alpha_T)^2}\eta_{R,T}^2 + \sum_{l\in\mathcal{E}(T)} \frac{\alpha_l}{\nu^{1/2}(k_1+k_2\alpha_l)^2}\eta_{R,l}^2\Big\}^{1/2} \preceq$$

$$|||u-u_h||| + \Big\{\sum_{T\in\mathcal{T}_h} \frac{\alpha_T^2}{(k_1+k_2\alpha_T)^2}\|f-f_h\|_{0,T}^2\Big\}^{1/2} \ ,$$

$$|||u-u_h||| \preceq \Big\{\sum_{T\in\mathcal{T}_h} \alpha_T^2(\eta_{R,T}^2 + \eta_{S,T}^2) + \sum_{l\in\mathcal{E}(T)} \alpha_l\eta_{R,l}^2\Big\}^{1/2}$$

$$+ \Big\{\sum_{T\in\mathcal{T}_h} \alpha_T^2\|f-f_h\|_{0,T}^2\Big\}^{1/2} \ .$$

Remark 3. If we compare the above estimates with the ones derived by Verfürth (see, *e.g.*, [V]), then we observe the presence of an additional term, namely $\eta_{S,T}$, in the upper bound which is not present in the lower bound. This "unsimmetry" in the two sides of the estimates is due to the lack of the usual Galerkin orthogonality property for our scheme. □

Remark 4. The above estimates should be compared also with the a posteriori error estimates presented in Angermann's works [A1,A2]. The author introduces a method for the discretization of (1) for which the Galerkin orthogonality property fails, too. He constructs upper and lower bounds for the error based on the partition of unity approach of [BR]. The error, however, is measured in a norm only implicitly defined by the variational problem. Hence, it is not straightforward to adapt these estimates to mesh refinement algorithms. □

Finally, we show how the above error estimates can be utilized to obtain a refined mesh. To this end, we follow the strategy of equidistribution of the error indicator presented in [MPR]. We require that the true relative error is bounded in terms of a prescribed tollerance TOL as follows:

$$\frac{|||u-u_h|||}{|||u_h|||} \preceq TOL \ , \tag{4}$$

The simplest way to satisfy (4) is to require for each $T \in \mathcal{T}_h$ that

$$\alpha_T^2(\eta_{R,T}^2+\eta_{S,T}^2)+\alpha_T^2\|f-f_h\|_{0,T}^2+\sum_{l\in\mathcal{E}(T)}\alpha_l\eta_{R,l}^2 \le \frac{TOL^2}{Nele}|||u_h|||^2 \ , \tag{5}$$

where $Nele$ is the number of elements in the triangulation. Given $T \in \mathcal{T}_h$, if (5) is not satisfied then T must be refined: the interior point Z_T is added to the primary mesh. As an example of application, we consider the following problem, studied in [J,MPR].

Problem 5. We consider (1) with $\Omega = (0,1)^2$, $\nu := 10^{-3}$, $\boldsymbol{a} := (2,1)$ and $f \equiv 0$; on the boundary, we impose the inhomogeneous Dirichlet condition $u(0,y) = 1$ for $0 < y \le 1$, $u(x,1) = 1$ for $0 \le x < 1$, $u(x,y) = 0$ if $x = 1$ or $y = 0$. □

In Fig. 1 we show the final resulting mesh, obtained by setting $TOL = 0.3$, and the contour levels of the obtained solution. As we can see, the approximate solution presents a well resolved outflow boundary layer in the upper right corner. Even if the adapted mesh strictly follows the contour level, we can observe a residual diffusion in the resolution of the inner layer. Finally, we observe that the method works well also on triangulations less regular than the ones for which our theorems hold.

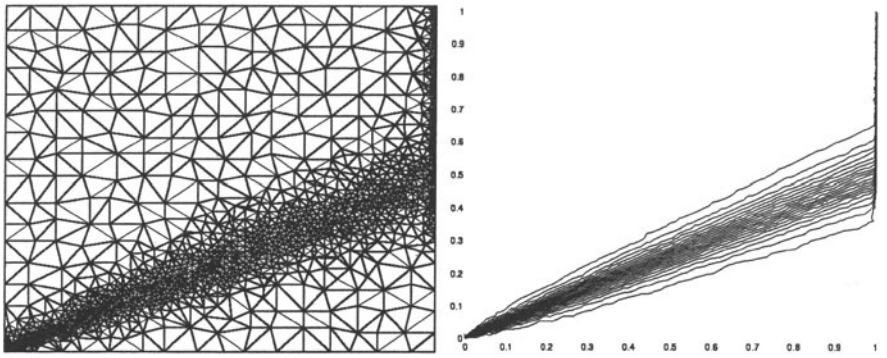


Fig. 1. The refined mesh (2485 points) and the related solution

References

[A1] Angermann, L.: Balanced a posteriori error estimates for finite volume type discretizations of convection diffusion problems. Computing **55** (1995) 4, 305–323

[A2] Angermann, L.: Error estimates for the finite element solution of an elliptic singularly perturbed problem. IMA J. Numer. Anal **15** (1995), 161–196

[BR] Babuška, I., Rheinboldt, W. C.: Error estimates for adaptive finite element computation. SIAM J. Numer. Anal. **14** (1978) 4, 736–754

[CMmC] Cai, Z., Mandel, J., McCormick, S.: The finite volume element method for diffusion equations on general triangulations. SIAM J. Num. Anal. **28** (1991) 2, 392–402

[CS] Cockburn, B., Shu, C.-W.: The local discontinuous Galerkin method for time dependent convection- diffusion systems. SIAM J. Num. Anal. **35** (1998) 6, 2440–2463

[J] Johnson, C.: Adaptive finite element methods for diffusion and convection problems. Comput. Meth. Appl. Mech. Engrg. **82** (1990) 301-322

[MPR] Medina, J., Picasso, M., Rappaz, J.: Error estimates and adaptive finite elements for nonlinear diffusion-convection problems. Math. Mod. and Meth. in Appl. Sci. **6** (1996), 689-712

[S1] Sardella, M.: A coupled finite element-finite volume method for convection-diffusion problems. To appear in IMA J. Numer. Anal.

[S2] Sardella, M.: Ph-D Thesis. In preparation.

[V] Verfürth, R.: A posteriori error estimators for convection diffusion equations. Num. Math. **80** (1998) 4, 641-663

A Discontinuous Galerkin Method for the Shallow Water Equations with Source Terms

Dirk Schwanenberg, Jürgen Köngeter

Institute of Hydraulic Engineering and Water Resources Management, Aachen University of Technology, 52056 Aachen, Germany

Abstract. The authors present a numerical solution for the shallow water equations based on the Runge Kutta Discontinuous Galerkin method. Modeling sink and source terms introduces restrictions to the space discretization and a modification of the slope limiter. Hydraulic test problems and a real-world application show the good performance of the scheme.

1 Introduction

A wide variety of physical phenomena are governed by the shallow water equations (SWE). Some examples are tides in oceans, the breaking of waves in shallow beaches, open-channel flow problems such as roll waves, flood waves in rivers and surges. The SWE are a set of nonlinear hyperbolic equations and approximate the depth-averaged free-surface gravity flow problem of an incompressible fluid.

The aim of this work is to present a numerical solution for the SWE based on the Runge Kutta Discontinuous Galerkin (RKDG) method which takes benefit of its main advantages, e.g. mass and momentum conservation, the handling of complex geometries, compactness and adaptivity.

2 Governing Equations

The SWE for one-dimensional transient open channel flows, written in conservative form, are

$$U_t + F_x = S, \quad \text{in } (0,L)\times(0,T), \tag{1}$$

$$\text{with } U = \begin{pmatrix} h \\ q \end{pmatrix}, \quad F = \begin{pmatrix} q \\ \frac{q^2}{h} + \frac{gh^2}{2} \end{pmatrix} \tag{2}$$

in which h = depth of flow, $q = uh$ = discharge, u = velocity and g = acceleration due to gravity. The right-hand side of the system represents sinks and sources of the momentum arising from the bed slope and friction losses,

$$S = \begin{pmatrix} 0 \\ gh(S_0 - S_f) \end{pmatrix}, \quad \text{with } S_f = \frac{n^2 q|q|}{h^{10/3}}, \quad S_0 = -z_x, \tag{3}$$

where z_x = bed slope is the spatial partial derivative of the bottom elevation z. The friction losses S_f are assumed to be given by Manning's formula where n = Manning's roughness coefficient.

3 The RKDG method

In this section we give a review of the one-dimensional RKDG method for the SWE using Legendre polynomials as local basis functions. Detail about the scheme and the extension to the two-dimensional case can be found in Cockburn (1998).

3.1 The Discontinuous Galerkin (DG) space discretization

For each partition of the interval $(0,L)$, $\{x_{j+1/2}\}_{j=1,\ldots,N}$, we set $I_j = (x_{j-1/2}, x_{j+1/2})$. We seek an approximation $U_h = (h_h, q_h)^T$ to U such that for each time $t \in [0,T]$, U_h belongs to the finite dimensional space $P^k(I)$ in I of degree at most k. After multiplying (1) by the test function v_h and integration over I_j we integrate the flux term by parts to obtain the weak formulation

$$\begin{aligned} &\forall\, j=1,\ldots,N\,,\ \ \forall U_h \in P^k(I_j)\,,\ \ \forall v_h \in P^k(I_k):\\ &\int_{I_j} \partial_t U_h(x,t)\, v_h(x)\, dx - \int_{I_j} F(U_h(x,t))\, \partial_x v_h(x)\, dx + \\ &H(U_h)_{j+1/2}(t)\, v_h(x_{j+1/2}^-) - H(U_h)_{j-1/2}(t)\, v_h(x_{j-1/2}^+) \\ &= \int_{I_j} S(U_h(x,t))\, v_h(x)\, dx\,. \end{aligned} \tag{4}$$

Note that the function U_h is discontinuous at the points $x_{j+1/2}$ so that we have to replace the nonlinear flux F by a numerical flux H that depends on the two values of U_h at the points $x_{j+1/2}$.

By choosing the Legendre polynomials P_m as local basis functions we obtain a diagonal mass matrix under consideration of their L^2-orthogonality. The approximate solution U_h is then defined by

$$U_h(x,t) = \sum_{m=0}^{k} U_j^m\, \varphi_m(x), \quad \text{with} \quad \varphi_m(x) = P_m\left(\frac{2(x-x_j)}{\Delta_j}\right). \tag{5}$$

As in the standard Galerkin method we choose the local basis function $\varphi_m(x)$ as test function v_h.

While the modeling of the friction losses is straightforward, the appearance of the bed slope introduces restrictions to the choice of U_h. The system is in a balanced state if the water elevation, the sum from bottom elevation z and fluid depth h, is constant and if the discharge q is zero. To consider this balanced state in the discretization we have to approximate the water depth of an order equal or higher than the bottom elevation. By using a linear representation for the bottom elevation, i.e. a constant bed slope, we have to choose a local basis function of degree $k \geq 1$.

Note that if we use numerical integration for the integrals of F and S it should be exact for the above given balanced state.

3.2 The numerical flux

To complete the discretization in space it remains to choose the numerical flux H. The DG scheme is monotone if $H(U_L,U_R)$ is a Lipschitz, consistent, monotone flux (Cockburn, 1998). For the shallow water equations Toro (1992) presented a suitable HLL-type flux based on the suggested approximations of Harten et. al. (1983),

$$H^{HLL}(U_L,U_R)=\begin{cases} F_L & \text{if} \quad 0\le S_L \\ \dfrac{S_R F_L - S_L F_R + S_L S_R (U_R - U_L)}{S_R - S_L} & \text{if} \quad S_L \le 0 \le S_R \\ F_R & \text{if} \quad 0 \ge S_R \end{cases} \tag{6}$$

The wave speeds are chosen under assumption of two-rarefaction waves,

$$S_L = \min(u_L - \sqrt{gh_L}, u^* - \sqrt{gh^*}), \tag{7}$$

$$S_R = \min(u_R + \sqrt{gh_R}, u^* + \sqrt{gh^*}), \tag{8}$$

$$\text{with} \quad \sqrt{gh^*} = \frac{1}{2}(\sqrt{gh_L} + \sqrt{gh_R}) - \frac{1}{4}(u_R - u_L), \tag{9}$$

$$u^* = \frac{1}{2}(u_L + u_R) + \sqrt{gh_L} - \sqrt{gh_R}\,. \tag{10}$$

Numerical experience with the HLL flux and a Roe flux with entropy fix does not show any significant impact on the computed solutions for an element degree k ≥ 1. Note that the above given wave speeds are obtained under an assumption of a wet bed, i.e. a non-zero flow depth h, on both sides of the computational domain. The speeds for a dry bed on one side can be found in Fraccarollo and Toro (1995).

3.3 The TVD Runge Kutta time discretization

After discretizing in space by the DG method it remains a system of ODEs for the degrees of freedom that can be rewritten in the form

$$\frac{d}{dt}U_h = L_h(U_h), \quad \text{in } (0,T). \tag{11}$$

To maintain the TVD property of the scheme the time discretization is done by a high order TVD Runge Kutta scheme. It should be at least of an order $k+1$, e.g. in Gottlieb and Shu (1998). The Euler forward step is the optimal first order method.

3.4 The TVBM slope limiter

To obtain high order stability for the space discretization $k \ge 1$ a slope limiter is applied on every computational result of the Runge Kutta method. The limiting is performed in the local characteristic variables U_c, V_c, defining V as the unlimited variable U. In this work we present the limiting procedure for $k = 1$. Limiters for higher order approximations $k > 1$ of U_h and the multidimensional case can be found in Cockburn (1998).

As we mentioned before, the system is in a balanced state if the fluid elevation is constant with zero discharge. If the bed slope is non-zero the limiting should lead to a constant fluid elevation instead of a constant fluid depth. Assuming a constant bed slope z_x the variables can be modified by

$$\tilde{V}_j^1 = V_j^1 + \begin{pmatrix} \frac{1}{2}\Delta_j z_x \\ 0 \end{pmatrix}, \quad \Delta\tilde{V}_{j+1/2} = \overline{V}_{j+1} - \overline{V}_j + \begin{pmatrix} z_{j+1} - z_j \\ 0 \end{pmatrix}, \tag{12}$$

obtaining a local elevation instead of a fluid depth. Note that the mean of the local elevation and the depth are equal in each element so that no influence results on the limiting in the characteristic field. Now we apply a scalar limiter on all components of the vectors in their characteristic fields:

$$\tilde{U}_{c,j}^1 = m(\tilde{V}_{c,j}^1, \Delta\tilde{V}_{c,j-1/2}, \Delta\tilde{V}_{c,j+1/2}) \tag{13}$$

$$\text{with} \quad m(a_1,\ldots,a_v) = \begin{cases} s\min_{1\le n\le v}|a_n|, \text{ if } s = sign(a_1) = \ldots = sign(a_v), \\ 0, \text{ otherwise.} \end{cases} \tag{14}$$

To obtain more than first order accuracy at extremas we replace the *minmod* function m by the TVB corrected *minmod* function $\tilde{m}$ defined as

$$\tilde{m}(a_1,\ldots,a_v) = \begin{cases} a_1, \text{ if} |a_1| \le M(\Delta x)^2, \\ m(a_1,\ldots,a_m), \text{ otherwise} \end{cases}. \tag{15}$$

4 Computational Results

Focusing on dam-break problems the applicability of the SWE is demonstrated by several authors, e.g. in Franccarollo and Toro (1995), Stansby et. al. (1998). In this work we compare the numerical solution with the analytical solution. Furthermore, we give an example of a real-world application of the scheme.

4.1 One-dimensional dam-break

We consider a flat rectangular channel with zero friction and zero bed slope. A dam at position $x = 0.5$m divides the channel in an upstream and a downstream section. The initial conditions are given by

$$U(x,0) = \begin{cases} (h_{up},0)^T, & \text{if } x \le 0.5\text{m} \\ (h_{down},0)^T, & \text{if } x > 0.5\text{m} \end{cases}. \tag{16}$$

At time $t = 0$ the dam is suddenly removed, causing a bore travelling downstream and a rarefaction wave traveling upstream. The analytic solution of this problem is given in Stoker (1957). Figure 1 shows the performance of the RKDG schemes of piecewise constant, piecewise linear and piecewise quadratic elements for a depth

ratio of 0.5, h_{up} = 1m at time t = 0.1s. The space discretization has a resolution of 10 elements. For the k = 1,2 schemes we use the TVBM slope limiter with M = 50.

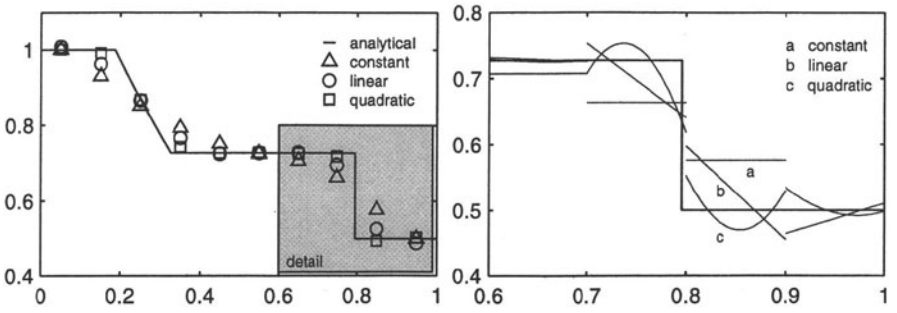


Figure 1. Mean depth [m] (left) and detail (right) for a one-dimensional dam-break obtained with k = 0,1,2, M = 50, Δj = 0.1m, t = 0.1s: Exact solution (solid line), piecewise constant solution (triangle), piecewise linear solution (circle), piecewise quadratic solution (square).

4.2 Circular dam-break

We consider a cylindrical dam with radius r = 11m separating the computational domain into an inner and external area with a fluid depth h_{in} = 10m, h_{ex} = 1m and zero discharge q_x, q_y. After removing the dam at time t = 0 the resulting flow shows the ability of the method to conserve high symmetries on unstructured triangle meshes. The flow changes from circular symmetry to axial symmetry during the reflection at the boundary that is assumed as solid wall.

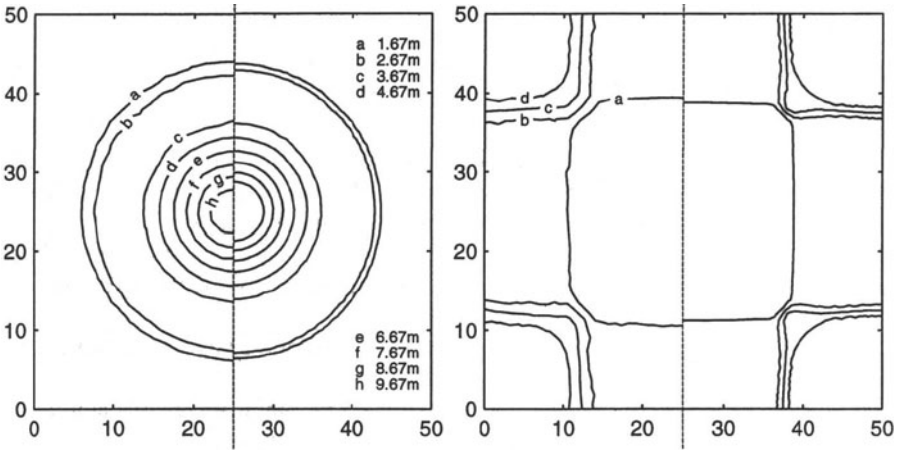


Figure 2. Flow depth [m] at t = 0.8s (left) and t = 4.5s (right) for circular dam-break obtained with k = 1, M = 50: low refined mesh with 1528 triangle elements (left area), high refined mesh with 5998 triangle elements (right area).

4.3 Real-world application

Furthermore, the two-dimensional SWE are applied to the study of a sudden weir opening at a creek near Aachen, Germany. The model includes friction losses and bottom elevation. Looking at the resulting flow in figure 3 we can approximately predict the flooding of the floodplain, the related water depth and velocity.

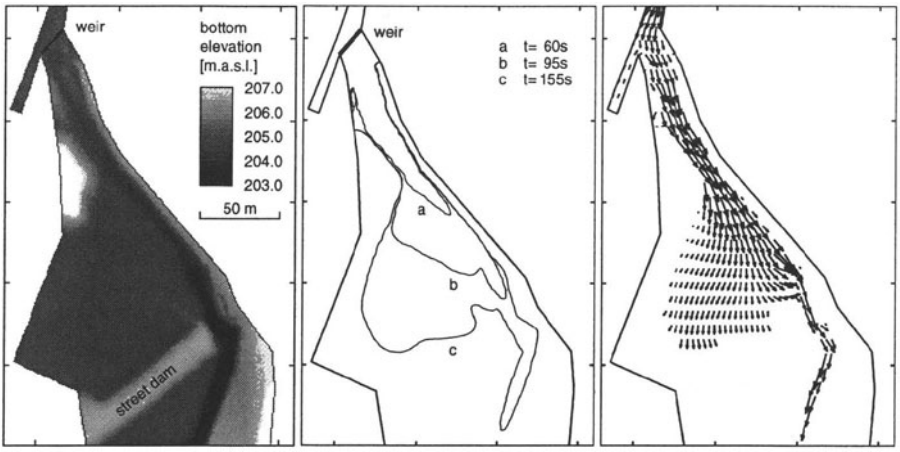


Figure 3. Flow after a sudden opening of a weir: bottom elevation (left), spreading of the water at $t = 60$s, $t = 95$s, $t = 155$s (middle), velocity at $t = 155$s (right).

Acknowledgements

Special thanks are due to Bernardo Cockburn (Minnesota, USA) for fruitful discussions.

References

Cockburn, B. (1998): Discontinuous Galerkin methods for convection dominated problems. RTO-ATV/VKI Special Course: Higher order discretisation methods in computational fluid dynamics. von Karman Institute, Belgium.

Fraccarollo, L., Toro, E.F. (1995): Experimental and numerical assessment of the shallow water model for two-dimensional dam-break type problems. J. Hydraulic Research, vol. 33, N. 6, pp. 843-863.

Gottlieb, S., Shu, C.-W. (1998): Total variation diminishing Runge-Kutta schemes. Mathematics in Computation, v67, pp. 73-85.

Stansby, P.K., Chegini, A., Barnes, T.C.D. (1998): The initial stages of dam-break flow. J. Fluid Mech. 374, Cambridge University Press, pp. 407-424.

Stoker, J. (1957): Water Waves. Interscience.

Toro, E.F. (1992): Riemann problems and the WAF method for solving the two-dimensional shallow water equations. Phil. Trans. R. Soc. Lond. 338, pp. 43-68

Dispersion Analysis of the Continuous and Discontinuous Galerkin Formulations

Spencer Sherwin

Aeronautics, Imperial College, Prince Consort Road, London, SW7 2BY, UK

Abstract. The dispersion relation of the semi-discrete continuous and discontinuous Galerkin formulations are analysed for the linear advection equation. In the context of an spectral/*hp* element discretisation on an equispaced mesh the problem can be reduced to a $P \times P$ eigenvalue problem where P is the polynomial order. The analytical dispersion relationships for polynomial order up to $P = 3$ and the numerical values for $P = 10$ are presented demonstrating similar dispersion properties but show that the discontinuous scheme is more diffusive.

1 Introduction

In this paper we derive the phase properties of the discontinuous and continuous *hp* element Galerkin formulation [1,2] of the linear advection equation. To gain a better insight into the phase properties of these schemes we analytically construct an $P \times P$ eigenvalue problem which completely describes the phase properties of both the continuous and discontinuous Galerkin schemes on an equi-spaced mesh. This analysis shows that the discontinuous Galerkin formulation has a comparable dispersion relationship as the continuous version although the discontinuous formulation show significant damping at higher frequencies.

2 Continuous and Discontinuous Galerkin Formulation

Considering the linear advection equation:

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0, \tag{1}$$

we assume an equispaced discretisation within which the solution is a approximated by $u(x,t) \simeq u^e(x,t) = \sum_{p=0}^{p=P} \phi_p(x)\hat{u}_p^e(t)$ within the e^{th} elemental region $[x_l^e, x_u^e]$. Taking the inner product with respect to the expansion basis $\phi_q(x)$ we obtain the elemental Galerkin approximation:

$$\left(\phi_q(x), \frac{\partial u^e}{\partial t}(x,t)\right)_e + \left(\phi_q(x), \frac{\partial u^e}{\partial x}(x,t)\right)_e \quad \forall\, q. \tag{2}$$

In the standard continuous Galerkin formulation $\phi_p(x)$ is typically defined in terms of an interior and boundary type decomposition so that a globally C^0 continuous can be constructed. Introducing the elemental matrices

$\boldsymbol{M}[q,p] = (\phi_q, \phi_p)$, $\boldsymbol{D}[q,p] = (\phi_q, \frac{\partial \phi_p}{\partial x})$ the global form of equation (2) can be represented in matrix form as:

$$\boldsymbol{Z}\underline{\boldsymbol{M}}\boldsymbol{Z}^T\hat{\boldsymbol{u}}_t + \boldsymbol{Z}\underline{\boldsymbol{D}}\boldsymbol{Z}^T\hat{\boldsymbol{u}} = 0, \tag{3}$$

where $\boldsymbol{Z}$ is the matrix operation of direct stiffness assembly. In equation (3) the use of the underlined matrix $\underline{\boldsymbol{M}}$ represents the extension of the local matrices $\boldsymbol{M}$ to a global system of block diagonal matrices. A similar extension has also been assumed for $\hat{\boldsymbol{u}} = \boldsymbol{Z}\underline{\hat{\boldsymbol{u}}}^e$ where $\hat{\boldsymbol{u}}^e$ is a local vector of expansion coefficients, i.e. $\underline{\boldsymbol{u}}^e[p] = \hat{u}^e_p$. See [3] for more details.

For the discontinuous Galerkin formulation we integrate the second term in equation (2) by parts to obtain:

$$\left(\phi_q, \frac{\partial u^e}{\partial t}\right)_e - \left(\frac{\partial \phi_q(x)}{\partial x}, u^e\right)_e + [\phi_q(x)u^e]^{x^e_u}_{x^e_l}. \tag{4}$$

To allow information to propagate from one elemental region to another the boundary flux is upwinded which is denoted as $u^e|_x = \tilde{u}^e|_x$. For the linear advection equation $\tilde{u}^e|_x$ is defined as:

$$\tilde{u}^e|_x = u^e_{x^e_u} \text{ if } x = x^e_u, \qquad \tilde{u}^e|_x = u^{e-1}_{x^{e-1}_u} \text{ if } x = x^e_l. \tag{5}$$

A numerically more convenient form is obtained by integrating the second term in equation (4) by parts again and substituting in the definition of $u^e(x,t)$ to arrive at the equation:

$$\sum_e \left[\left(\phi_q, \textstyle\sum_p \phi_p\right)_e \frac{\partial \hat{u}_p}{\partial t} + \left(\phi_q, \textstyle\sum_p \frac{\partial \phi_p}{\partial x}\right)_e \hat{u}_p + [\phi_q(\tilde{u}^e - u^e)]^{x^e_u}_{x^e_l} \right] = 0.$$

This equation can be represented in an elemental matrix form as:

$$\boldsymbol{M}\frac{\partial \boldsymbol{u}^e}{\partial t} + \boldsymbol{D}\boldsymbol{u}^e + \boldsymbol{F}\boldsymbol{u}^e + \boldsymbol{G}\boldsymbol{u}^{e-1} = 0 \tag{6}$$

where $\boldsymbol{M}, \boldsymbol{D}$ have the same definition as before and applying the upwinding condition (5) $\boldsymbol{F}[q,p] = \phi_q(x_l)\phi_p(x_l)$, $\boldsymbol{G}[q,p] = -\phi_q(x_l)\phi_p(x_u)$.

3 Phase Analysis

Considering an equispaced mesh of N_{el} elements within a periodic region $[x_a, x_b]$ the element matrices for both formulations become:

$$\begin{aligned} \boldsymbol{M}[q,p] &= \tfrac{h}{2}(\phi_q(\zeta), \phi_p(\zeta)), \quad & \boldsymbol{D}[q,p] &= (\phi_q(\zeta), \phi'_p(\zeta)), \\ \boldsymbol{F}[q,p] &= \phi_q(-1)\phi_p(-1), & \boldsymbol{G}[q,p] &= -\phi_q(-1)\phi_p(1) \end{aligned}$$

where $h/2$ is the Jacobian of the mapping from the region $[x^e_l, x^e_u]$ to $[-1,1]$ and $h = (x_b - x_a)/N_{el}$.

The global matrix system for the semi-discrete advection equation using the discontinuous formulations can be written as:

$$\begin{bmatrix} \boldsymbol{M} & 0 & \cdots & 0 \\ 0 & \boldsymbol{M} & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \boldsymbol{M} \end{bmatrix} \frac{\partial \underline{\hat{\boldsymbol{u}}}^e}{\partial t} + \begin{bmatrix} (\boldsymbol{D}+\boldsymbol{F}) & 0 & \cdots & \boldsymbol{G} \\ \boldsymbol{G} & (\boldsymbol{D}+\boldsymbol{F}) & \cdots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & 0 & \boldsymbol{G} & (\boldsymbol{D}+\boldsymbol{F}) \end{bmatrix} \underline{\hat{\boldsymbol{u}}}^e = 0 \quad (7)$$

In our formulation we assume an orthonormal basis of Legendre polynomials $L_p(\zeta)$ and so $\boldsymbol{M} = \boldsymbol{I}h/2$. We now seek a solution to the semi-discrete problem of the form

$$\underline{\hat{\boldsymbol{u}}}^e = e^{-i\omega t}[\boldsymbol{\alpha} e^{0i\theta}, \boldsymbol{\alpha} e^{i\theta}, \boldsymbol{\alpha} e^{i2\theta}, \cdots, \boldsymbol{\alpha} e^{i(N_{el}-1)\theta}]^T$$

where $i = \sqrt{-1}$ and $\boldsymbol{\alpha}$ is a vector of $P+1$ constants. For periodicity $e^{i\theta N_{el}} = 1$ which implies that $\theta_n = 2\pi n/N_{el}$ for $0 \leq n < N_{el}$. Substituting this expression into equation (7) we obtain the eigenvalue problem

$$\begin{gathered} [-(i\omega h)\boldsymbol{I} + \boldsymbol{A}]\,\boldsymbol{\alpha} = 0 \quad \text{where} \\ \boldsymbol{A}[q,p] = 2\left\{\boldsymbol{D}[q,p] + \boldsymbol{F}[q,p] + \boldsymbol{G}[q,p]e^{-i\theta}\right\}, \quad \boldsymbol{\alpha}[p] = \alpha_p \end{gathered} \quad (8)$$

For $P = 0$, which is the Godunov scheme, $\phi_0(\zeta) = 1/\sqrt{2}$ and so $\boldsymbol{D} = 0, \boldsymbol{F} = 1/2, \boldsymbol{G} = -1/2$ and the eigenvalue problem reduces to $[-i\omega h + 1 - e^{-i\theta}]\alpha_0 = 0$. Therefore $ih\omega = (1 - e^{-i\theta})$ which is identical to the upwinded first order finite difference scheme. When $P = 1$, $\phi_0(\zeta) = 1/\sqrt{2}, \phi_1(\zeta) = \sqrt{3/2}\zeta$ and

$$\boldsymbol{A} = \begin{bmatrix} 1 - e^{-i\theta} & \sqrt{3}(1 - e^{-i\theta}) \\ \sqrt{3}(e^{-i\theta} - 1) & 3(1 + e^{-i\theta}) \end{bmatrix}$$

which leads to the eigenvalues $ih\omega_{1,2} = 2 + e^{-i\theta} \pm \sqrt{e^{-2i\theta} + 10e^{-i\theta} - 2}$.

The analytic eigenvalues for the next two polynomial orders were obtained using Mathematica and found to be, for $P = 2$:

$$ih\omega_1 = 3 - e^{-i\theta} - \beta^{-1/3}(14 - e^{-i\theta} + 3e^{i\theta}) + \beta^{1/3}e^{-i\theta} \quad (9)$$

$$\begin{aligned} ih\omega_{2,3} = {} & 3 - e^{-i\theta} + \beta^{-1/3}(14 - e^{-i\theta} + 3e^{i\theta})(1 \pm i\sqrt{3})/2 \\ & -\beta^{1/3}e^{-i\theta}(1 \mp i\sqrt{3})/2 \end{aligned} \quad (10)$$

where

$$\begin{aligned} \alpha &= 3 - 166e^{i\theta} + 1872e^{2i\theta} - 18e^{3i\theta} + 9e^{4i\theta} \\ \beta &= -1 + 21e^{i\theta} - 75e^{2i\theta} + 3e^{3i\theta} + 2\sqrt{\alpha}e^{i\theta} \end{aligned}$$

and for $P = 3$ are

$$ih\omega_{1,2} = 4 + e^{-i\theta} - \frac{\sqrt{\gamma}}{2} \pm \frac{\sqrt{\delta^-}}{2}, \qquad ih\omega_{3,4} = 4 + e^{-i\theta} + \frac{\sqrt{\gamma}}{2} \pm \frac{\sqrt{\delta^+}}{2} \quad (11)$$

where

$$\alpha = 24 + 2796e^{i\theta} + 342225e^{2i\theta} + 6271880e^{3i\theta} + 326640e^{4i\theta} + 1056e^{5i\theta} - 96e^{6i\theta}$$
$$\beta = -8 + 924e^{-i\theta} + 291e^{-2i\theta} - 2e^{-3i\theta} + \sqrt{\alpha}e^{-3i\theta}$$
$$\gamma = -16 + 72e^{-i\theta} + 4e^{-2\theta} + 4(50/\beta)^{1/3}(-e^{-2i\theta} - 66e^{-i\theta} + 2) + 2(20\beta)^{1/3}$$
$$\delta^{\pm} = 144e^{-i\theta} - 32 + 8e^{-2i\theta} - 4(50/\beta)^{1/3}(-e^{-2i\theta} - 66e^{-i\theta} + 2) - 2(20\beta)^{1/3}$$
$$\pm (\gamma)^{-1/2}(16e^{-3i\theta} + 432e^{-2i\theta} + 1968e^{-i\theta} + 64).$$

In the above expressions the n^{th} root of a complex number is considered to be $z^{1/n} = |z|^{1/n}e^{\angle z/n}$.

For the continuous Galerkin method we also seek a solution of the form

$$\hat{\underline{u}}^e = e^{-i\omega t}[\alpha e^{0i\theta}, \alpha e^{i\theta}, \alpha e^{i2\theta}, \cdots, \alpha e^{i(N_{el}-1)\theta}]^T$$

however this time α is a vector of P constants due to the assembly operation involved with the continuous scheme. Assuming that the vertex degrees of freedom are defined for $p = 0$ and P in the definition of $\phi_p(\zeta)$ then substituting this expression into globally assembled matrix system (3) we obtain the eigenvalue problem

$$[-i\omega \boldsymbol{B} + \boldsymbol{A}]\,\boldsymbol{\alpha} = 0 \quad \text{where} \quad (12)$$
$$\boldsymbol{B}[q,p] = \boldsymbol{M}[q,p] + \boldsymbol{M}[q,P]e^{i\theta}\delta_{p0} + \boldsymbol{M}[P,p]e^{-i\theta}\delta_{q0} + \boldsymbol{M}[P,P]\delta_{p0}\delta_{q0}$$
$$\boldsymbol{A}[q,p] = \boldsymbol{D}[q,p] + \boldsymbol{D}[q,P]e^{i\theta}\delta_{p0} + \boldsymbol{D}[P,p]e^{-i\theta}\delta_{q0} + \boldsymbol{D}[P,P]\delta_{p0}\delta_{q0}$$
$$0 \le p, q < P$$

Using a continuous basis of the form: $\phi_0(\zeta) = \frac{1-\zeta}{2}$, $\phi_p(\zeta) = \frac{1-\zeta}{2}\frac{1+\zeta}{2}P_p^{1,1}(\zeta)(0 < p < P)$ and $\phi_P(\zeta) = \frac{1+\zeta}{2}$, where $P_p^{1,1}(\zeta)$ is the Jacobi polynomial, for $P = 1$ we find that

$$\boldsymbol{M} = \frac{h}{2}\begin{bmatrix} 2/3, 1/3 \\ 1/3, 2/3 \end{bmatrix}, \boldsymbol{D} = \begin{bmatrix} -1/2, 1/2 \\ -1/2, 1/2 \end{bmatrix}$$

and so $\boldsymbol{A} = \frac{e^{i\theta}-e^{-i\theta}}{2} = i\sin\theta$ and $\boldsymbol{B} = \frac{h}{3}(2 + \frac{e^{i\theta}+e^{-i\theta}}{2}) = \frac{h}{3}(2 + \cos\theta)$ which gives us the dispersion relations

$$i\omega\frac{h}{3}(2 + \cos\theta) + i\sin\theta = 0 \quad \Rightarrow \quad h\omega = \frac{-3\sin(\theta)}{2 + \cos\theta}.$$

which is identical to the fourth order compact finite difference scheme using a three-point stencil. For $P = 2$ we obtain:

$$h\omega_{1,2} = \frac{4\sin\theta \pm 2\sqrt{40\sin^4(\theta/2) + 9\sin^2\theta}}{\cos\theta - 3}$$

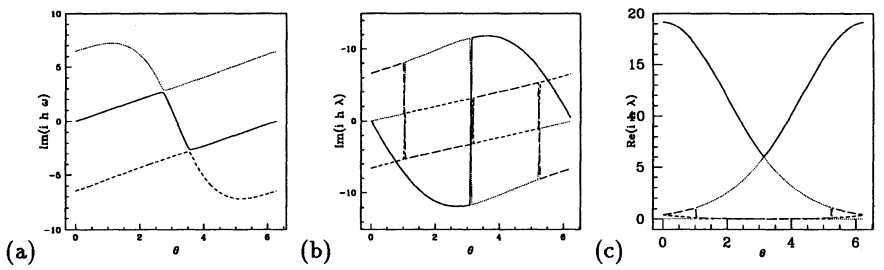


Fig. 1. Dispersion relations for (a) continuous scheme at $P = 3$ and (b,c) discontinuous scheme at $P = 3$.

and finally for $P = 3$ the eigenvalues are:

$$\alpha = 90\cos 2\theta + 5760\cos\theta + 6750$$
$$\beta = 540\sin 3\theta + 153900\sin\theta - 77760\sin 2\theta$$
$$\gamma = \sqrt{4\alpha^3 + \beta^2} - i\beta$$
$$ih\omega_1 = \frac{15e^{2\theta} - 15 - \alpha(2/\gamma)^{1/3}e^{i\theta} + e^{i\theta}(\gamma/2)^{1/3}}{3(1 + 8e^{i\theta} + e^{2i\theta})} \tag{13}$$
$$ih\omega_{2,3} = \frac{15e^{2\theta} - 15 + \alpha/(4\gamma)^{1/3}(1 \pm \sqrt{3}i)e^{i\theta} - (\gamma/16)^{1/3}(1 \mp \sqrt{3}i)e^{i\theta}}{3(1 + 8e^{i\theta} + e^{2i\theta})} \tag{14}$$

4 Discussion

Figure 1 illustrates the dispersion relation for the continuous and discontinuous formulations at a polynomial order of $P = 3$. The continuous formulation provides a purely imaginary phase solution (see eqn.(13)). Figures 1(b) and (c) show the imaginary and real components of the analytic phase relation for the discontinuous scheme given by equation (11). These plots completely define the phase relationship for any number of elements. In both imaginary component of the solution we see a linear growth with θ which represents the analytic dispersion relation for the linear advection equation. At very high frequencies the curves decay back to zero. The very dispersive modes of the discontinuous scheme are associated with a very fast damping as indicated by figure 1(c). Two notable features of the imaginary components shown in figure 1 are the branch jumping of the solutions and the multiple roots for each value of θ. The cause of the branch jumping is due to the restriction in argument of the trigonometric functions in the computer program between $-\pi < \theta < \pi$. When we evaluate the square root of a complex number of argument π a small perturbation, ϵ to the argument causes a jump from $\pi - \epsilon$ to $-\pi + \epsilon$ which leads to a change in sign of the complex component of the

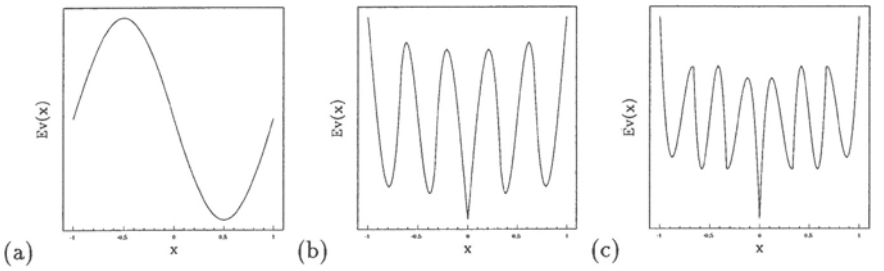


Fig. 2. Eigenvectors of discrete continuous problem $N_{el} = 6$, $P = 3$. (a)$\omega = 3.142 \approx \pi$ (b) $\omega = -16.08 \approx -5\pi$ and (c) $\omega = 21.60 \approx 7\pi$.

root. The multiple roots for a given θ are simply due to the ability of a high order expansion to resolve more than one eigenfunction within an element. This point is illustrated in figure 2 where we consider the eigenfunctions of the discrete problem for the continuous formulation when $N_{el} = 6, P = 3$. In this problem the first non-zero value of θ_n is $2\pi/6 \approx 1$ which produces three discrete values of $i\omega h$ (see figure 1(a)). For the analytic solution the three eigenvalues corresponding to the wavenumbers $\omega = \pi, -5\pi, 7\pi$ and should have eigenvalues of the form $e^{\omega x}$. As shown in figure 2 all branches produce a discernible approximation to these eigenfunctions.

In figure 3 we show a similar series of plot as shown in figure 1 for a polynomial order of $P = 10$. These plots were numerically evaluating from the systems given by equation (8) and (12). All results have been validated against a numerical eigenvalue evaluation of the complete semi-discrete system similar to equation (7) using LAPACK.

The ultimate aim of this investigation is to compare the relative advantages and disadvantages of the continuous and discontinuous formulations. Certainly from the implementation point of view the local elemental characteristic of the discontinuous scheme is very efficient. Both formulations also have comparable phase properties. However in the discontinuous formula-

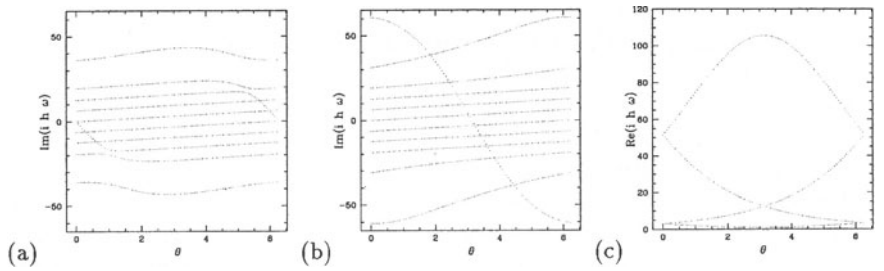


Fig. 3. Dispersion relations for (a) continuous scheme at $P = 10$ and (b,c) discontinuous scheme at $P = 10$.

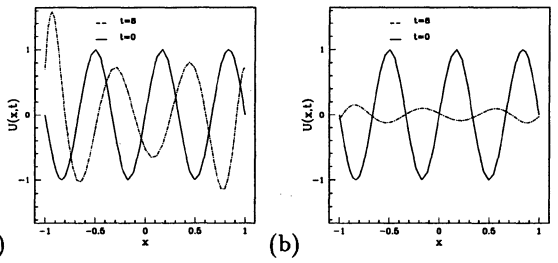


Fig. 4. Solution of the linear advection equation at $t = 0, 4$ for initial conditions of $\sin(3\pi x)$ with $N_{el} = 1, P = 10$. (a) continuous and (b) discontinuous.

tions the range of the dispersion relation is larger due to the greater number of degrees of freedom for a given polynomial order. Nevertheless at higher frequencies there is a significant damping which could be eliminated using a centred flux as investigated in [4]. Although the non-diffusive nature of the continuous scheme appears mathematically attractive it can lead to equally erroneous solution due to the poor phase properties of the higher frequencies. To illustrate this point in figure 4 we compare the two methods for $u(x, 0) = \sin(3\pi x)$ in $-1 < x < 1$ with $N_{el} = 1, P = 10$. The solid line shows the initial condition and the dotted line gives the solution at $t = 8$. From these figures we see the diffusive nature of the discontinuous scheme however we also note the the continuous scheme gives rise to a solution of magnitude greater than 1. Figure 2 demonstrates that the discrete eigenfunctions are not pure sinusoidal waves and so we would expect the projection of the initial conditions to involve more than one discrete eigenfunction. Although all of these frequencies are non-diffusive the poor phase property of the high frequency components can force the solution to become erroneous leading to the observed peak as the discrete eigenfunctions move in and out of phase.

The author would like to acknowledge Professor Mike Giles of the Oxford Computing Laboratory for insightful discussions.

References

B. Cockburn and C.W. Shu: TVB Runge-Kutta projection discontinuous Galerkin finite element methods for conservation laws II, General framework Math. Comp. **52** (1989) 411-435

I. Lomtev, C.W. Quillen and G. Karniadakis: Spectral/hp Methods for Viscous Compressible Flows on Unstructured 2D Meshes. J. Comp. Phys. Submitted (1997)

G.Em. Karniadakis and S.J. Sherwin: Spectral/*hp* Element Methods for CFD. Oxford University Press (1999)

F.Q.Hu, M.Y.Hussaini and P. Rasetarinera: An analysis of the discontinuous Galerkin method for wave propagation problems. Submitted to JCP (1999)

The Cell Discretization Algorithm; An Overiew

Howard Swann

San José State University
San José, CA 95192-0103, USA

Abstract. This non-conforming extension of the finite element method is illustrated with a model elliptic problem and other applications are sketched. New results concerning domain decomposition and the construction of a solenoidal basis for the Stokes equations are described.

1 Introduction

The cell discretization algorithm (CDA) is a non-conforming extension of the finite element method for approximating solutions for partial differential equations due to Greenstadt [5],[6] and Raviart and Thomas [8]. See also Dorr [4]. It is similar to the mortar method of Bernardi et al. [2].

A domain Ω is partitioned into *c*ells Ω_i and solutions are approximated by a linear combination of basis functions on each cell Ω_i. Another set of basis functions is defined for each interface Γ_{ij} between between cells Ω_i and Ω_j. These basis functions are used to enforce a form of weak continuity on approximations over the entire domain by requiring that the difference of the traces of approximations on the common boundaries of adjacent cells be orthogonal to increasing numbers of the interface basis set. These requirements, called *m*oment collocations, are expressed as a set of linear constraints on the coefficients to be used with the bases to define the approximation to the solution on each cell.

A weakly continuous basis that incorporates these moment collocation constraints is constructed so that Galerkin methods can be used to generate our approximate solutions. We have derived error estimates that establish convergence to the solutions for self-adjoint elliptic equations [3],[10], non-self-adjoint equations [13], parabolic equations [11] and hyperbolic equations [12]. The algorithm produces convergent approximations to the stationary and non-stationary Stokes equations [14],[16].

For this brief overview of the CDA, in section 2, we describe a simple example of the method that captures the essential ideas of the algorithm and shows the sort of error estimates that establish convergence of the approximations. Section 3 describes some recent results.

2 An elliptic problem

For our sample problem, we approximate $u \in H_0^1(\Omega)$ such that $-\Delta u + u = f$ where Ω is a disk and Ω_0 is its exterior. Partition Ω with a diameter into cells Ω_1 and Ω_2; the external boundary semicircles are Γ_{10} and Γ_{20}; Γ_{12} denotes the interface diameter between Ω_1 and Ω_2.

For each cell Ω_k, the solution u can be expressed in terms of a basis $\{B_i^k\}$ as $u(x) \mid_{\Omega_k} = \sum_{i=1}^m \beta_i^k B_i^k(x) + \mathcal{Q}_m(u)$, where $\sum_{i=1}^m \beta_i^k B_i^k(x)$ is u's $H^1(\Omega_k)$ projection on $\text{span}\{B_i^k\}_{i=1}^m$ and $\mathcal{Q}_m(u)$ is the projection on the complement. The norm of $\mathcal{Q}_m(u)$ is an important component of our error estimates.

We seek coefficients $\{b_i^k\}$ so that, on Ω_k, $\sum_{i=1}^m b_i^k B_i^k(x)$ is an approximation to the solution. This finite dimensional approximation space is denoted by $H[m]$.

For $v(x)$ defined for $x \in \Omega_i$, define $\gamma_{ij}(v)$ to be the trace map evaluating v on the boundary segment Γ_{ij}. There are constants C_{ij} such that

$$\| \gamma_{ij}(v) \|_{ij}^2 \equiv \int_{\Gamma_{ij}} [\gamma_{ij}(v)]^2 ds \leq C_{ij}^2 \int \int_{\Omega_i} | \nabla v |^2 + v^2 dx \equiv C_{ij}^2 \| v \|_{1,i}^2 \quad .$$

For the 'moment collocation' which gives our weak continuity across Γ_{ij}, we choose any basis $\{\omega_q^{ij}\}$ for $L_2(\Gamma_{ij})$. Thus for any $h \in L_2(\Gamma_{ij}), h = \sum_{q=1}^n d_q^{ij} \omega_q^{ij} + \mathcal{T}_n^{ij}(h)$, where $\sum_{q=1}^n d_q^{ij} \omega_q^{ij}$ is the L_2 projection of h on $\text{span}\{\omega_q^{ij}\}_{q=1}^n$ and $\mathcal{T}_n^{ij}(h)$ is the projection on the complement. For our example, weak continuity for approximation $u_{n,m}(x) \mid_{\Omega_k} \equiv \sum_{i=1}^m b_i^k B_i^k(x)$ is achieved by requiring that, on interface Γ_{12},

$$0 = \int_{\Gamma_{12}} [\gamma_{12}(u_{n,m}) - \gamma_{21}(u_{n,m})] \omega_q^{12} ds \equiv < [\gamma_{12}(u_{n,m}) - \gamma_{21}(u_{n,m})], \omega_q^{12} >_{12}$$

$$= \sum_{j=1}^m b_j^1 < \gamma_{12}(B_j^1), \omega_q^{12} >_{12} - b_j^2 < \gamma_{21}(B_j^2), \omega_q^{12} >_{12}$$

for $1 \leq q \leq n$, and, to achieve a weak match with the zero boundary values on Γ_{i0},

$$0 = < \gamma_{i0}(u_{n,m}) - 0, \omega_q^{i0} >_{i0} = \sum_{j=1}^m b_j^i < \gamma_{12}(B_j^i), \omega_q^{i0} >_{i0}$$

for $1 \leq q \leq$ n.

Independent of our basis, the space of functions that are weakly in $H_0^1(\Omega)$ in this manner is denoted $G_0[n]$. Our approximation space for the Cell method is $G_0[n][m] \equiv G_0[n] \cap H[m]$.

We note here that any polynomial implementation in $\mathbf{R}^K$ (with planar interior interfaces Γ_{ij}) includes a version of the $h-p$ finite element method as a special case [1]. For example, suppose our cells are in $\mathbf{R}^2$. If we use polynomials of degree less than or equal to p for the basis on each cell and choose

the first $p+1$ Legendre polynomials to be the collocation functions on each interface Γ_{ij}, our approximation is continuous throughout Ω, since the difference of the traces of the approximation on either side of any Γ_{ij}, if non-zero, is a polynomial of degree at most p, yet the difference must be orthogonal to the Legendre polynomial functions ω_k^{ij} for $k \leq p+1$. However, our computations show that it is not particularly advantageous to approximate with continuous functions using this method. We return to this later.

For self-adjoint elliptic problems, the moment collocation constraints are enforced by the use of Lagrange multipliers. For non self-adjoint problems and time dependent applications, we construct a basis $\{\mathcal{B}_i\}$ for $G_0[n][m]$ that automatically satisfies the collocation constraints in the following manner. If $\mathbf{b}$ is the vector of the coefficients used with the basis functions, we express the linear moment collocation constraints as $\mathbf{Mb} = \mathbf{0}$ where $\mathbf{M}$ is the collocation matrix. Thus the *n*ull space of $\mathbf{M}$ is the set of acceptable coefficient vectors $\mathbf{b}$. This is readily obtained from the $\mathbf{QR}$ factorization of $\mathbf{M}^T$:

$$\mathbf{M}^T = (\mathbf{Q}' \vdots \mathbf{Q}'') \begin{pmatrix} \mathbf{R} \\ \mathbf{O} \end{pmatrix} .$$

It then follows that the columns of $\mathbf{Q}''$ span the null space of $\mathbf{M}$. These columns are used to construct global basis $\{\mathcal{B}_i\}$. In the two cell case, express the i^{th} column as $(q_{1i}, q_{2i}, \ldots, q_{mi}, q_{(m+1)i}, \ldots, q_{2mi})$; a member of our new basis is $\mathcal{B}_i \equiv \sum_{j=1}^{m} q_{ji} B_j^1 + \sum_{j=m+1}^{2m} q_{ji} B_j^2$.

We return to our example. Convert the problem of solving $-\Delta u + u = f$ to the task of finding $u \in H_0^1(\Omega)$ such that for any $v \in H_0^1(\Omega)$,

$$(u, v)_1 = (f, v)_0 \ , \tag{1}$$

where $(\cdot, \cdot)_1$ denotes the $H^1(\Omega)$ inner product, and $(\cdot, \cdot)_0$ is the $L_2(\Omega)$ inner product. We extend $(\cdot, \cdot)_1$ to our space of discontinuous functions by summing $H^1(\Omega_k)$inner-products over k. We then use linear algebra to solve (1) in finite dimensional $G_0[n][m]$ to obtain an approximation denoted by $u_{n,m}$.

A typical error estimate is given in the following theorem [3],[10]; it establishes convergence for such approximations over general domains partitioned into N cells.

Theorem 1. *Suppose the solution u is in $H^2(\Omega)$. The normal derivative of u on Γ_{ij} is represented by $D_{\mathbf{n}_{ij}} u$; $\| \cdot \|_{ij}$ denotes the $L_2(\Gamma_{ij})$ norm. Let n_f be the largest number of faces Γ_{ij} of any of the N cells. C_T is the maximum of the the "trace constants" C_{ij}. (For squares or triangles of diameter h, $C_T < 3h^{-1/2}$.) Assume that the collocation functions $\{\omega_k^{ij}\}$ are $L_2(\Gamma_{ij})$- orthonormal. Suppose that $u_{n,m}$ denotes the approximation obtained by solving the linear system described above. Then*

$$\| u - u_{n,m} \|_1$$

$$\leq n_f C_T \sqrt{N} max\{ \| T_n^{ij}(D_{\mathbf{n}_{ij}} u) \|_{ij} \} + \sqrt{1 + 2(1/\mu) C_T^2 n_f} \| \mathcal{Q}_m(u) \|_1 \ .$$

A new symbol here is μ; it is the smallest eigenvalue for $\mathbf{M}\mathbf{C}^{-1}\mathbf{M}^T$ where $\mathbf{C}$ is the matrix of diagonal blocks $\mathbf{C}_k$, one for each Ω_k, with entries $(B_i^k, B_j^k)_{1,k}$. It is shown in [10] that $1/\mu$ is non-increasing as the number of basis functions m utilized on each Ω_i is increased. Hence the estimate establishes convergence as n becomes large and m becomes suitably larger since the norms of the complements of the projections become small.

We have obtained a number of results concerning μ. We show in [3] that $\mu \geq \min\{\mu_k\}$, where μ_k is the smallest eigenvalue for $\mathbf{M}_k\mathbf{C}_k^{-1}\mathbf{M}_k^T$; $\mathbf{M}_k$ is that portion of collocation matrix $\mathbf{M}$ that pertains to the cell Ω_k exclusively. Thus we can obtain an estimate for μ if we know the sorts of cells in the partition of Ω; we need not treat the entire system. Furthermore, for cells that are triangles, parallelograms, tetrahedra or parallelepipeds and are scaled to have diameter h, it is shown in [3] that $\mu_h \geq h\mu_1$, where μ_1 is the value when a largest side of Ω_i has unit diameter. Thus in this case, for estimates of $1/\mu$, it suffices to consider representative cells with diameter 1.

The norms of the terms $\| \mathcal{T}_n^{ij}\left(D_{\mathbf{n}_{ij}}u\right) \|_{ij}$ and $\| \mathcal{Q}_m(u) \|_1$ appearing in the theorem have been estimated when we use polynomial bases on cells partitioning Ω into triangles, parallelograms, tetrahedra or parallelepipeds (see [7]). These errors are expressed in terms of the degree p of the approximation on each cell, the degree q of the Legendre polynomial basis functions $\{\omega_k^{ij}\}$ used for collocation on the interfaces Γ_{ij}, and h, the largest of diameters of any cell. For example, applying estimates of Babuska et al. [1], the theorem gives the following error estimate for approximations (now denoted by $u_{q,p}$) of solutions for problems in $\mathbf{R}^2$, where the solution is in $H^k(\Omega)$: For constants C_1 and C_2 depending on the coefficients of a general elliptic operator and the angles between adjacent sides of cells,

$$\| u - u_{q,p} \|_1$$

$$\leq [C_1(h/2)^{\min(k-3,q)}q^{-(k-2)} + C_2\sqrt{1/\mu_1}h^{\min(k-2,p-1)}p^{-(k-1)}] \| u \|_{H^k} \quad .$$

When the solution is analytic, the q-dependency of the first term is of form $(.73(q+2))^{-(q+1.5)}$ and the p - dependency of the second term is of form $(.52p)^{-p}$ [3]. Typical experimental results for the analytic case are shown below. Fig. 1 (a) gives the error as a function of p, the degree of the approximation basis, for various values of q, where the cells are congruent squares. Fig. 1 (b) shows the errors when we fix q and vary p.

Recall that when $p = q$, the approximations are continuous, and we are implementing the finite element method. We see that enforcing continuity of the approximation is not all that necessary; the results are essentially as good when $q = p - 3$. Empirically, all experiments so far suggest setting $q = p-2$(for triangular cells) and $q = p-3$(for square cells) rather than $q = p$. This interesting result occurs in all the other experiments we have made with different kinds of partial differential equations. What seems to be happening is that although the first error term concerned with q is eliminated (in certain cases) when we enforce continuity by setting $q = p$, parameter $1/\mu_1$ becomes

very large (6,583 for a 10^{th} degree basis for squares), nullifying the apparent advantage obtained by eliminating the first error term [3]. On the other hand, when $p = 10$ and $q = 7, 1/\mu_1 = 567$.

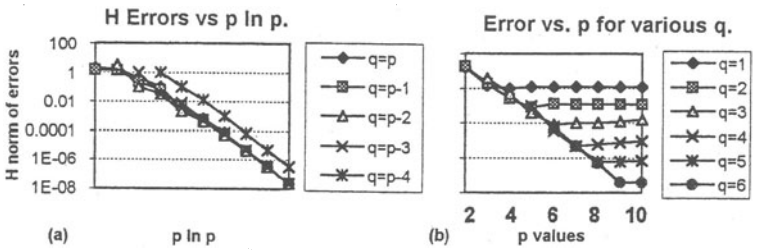


Fig. 1. Error calculations.

3 Recent Results

For the Stokes equations [14],[16], the necessary solenoidal condition div $\mathbf{u} = 0$ for solution vector $\mathbf{u}$ is imposed on each cell by requiring that div $\mathbf{u}_{n,m}$ be $L_2(\Omega_k)$ - orthogonal to basis functions $\{B_i^k : i = 1, \ldots, r\}$. This is enforced by adding additional rows to $\mathbf{M}$; the QR decomposition for $\mathbf{M}^T$ again gives us a global weakly continuous basis. When the basis functions are polynomials, we can enforce the solenoidal condition exactly, for, with an approximation $\mathbf{u}_{q,p}$ of degree p, div $\mathbf{u}_{q,p}$ is a polynomial of degree $p-1$; if this is orthogonal to basis functions containing div $\mathbf{u}_{q,p}$ in their span, the divergence is zero. We obtain approximations for both $\mathbf{u}$ and the pressure.

We have some new experimental results obtained while using cell discretization methods in an adaptive alternating Neumann - Dirichlet algorithm for domain decomposition [15]. The domain of an elliptic Dirichlet problem is partitioned into two subdomains. Solutions on the two subdomains can be patched together to form a solution to the original problem provided the solutions agree across the common interface of the subdomains and the normal derivatives there have equal absolute values and are opposite in sign. We use a method proposed by Rice et al. [9] and generate such a solution by alternating between imposing Neumann and Dirichlet conditions on the interface, with boundary data adapted from the results of the previous approximation. A-posteriori error estimates show that we have global convergence provided the computed interface errors are sufficiently small; the estimates also give a good indication of how to use the information obtained from a previous computation to best suggest new boundary values. We have yet to encounter an elliptic problem on a polygonal domain where this algorithm fails to converge quite rapidly, which suggests that it may be possible to prove general convergence for this method.

References

[1] Babuška, I., Suri, M.: The p and $h-p$ versions of the finite element method, an overview. Comput. Methods Appl. Mech. Engrg. **80**, (1990), no. 1-3, 5-26.

[2] Bernardi, C., Maday, Y., Patera, A. T.: Domain decomposition by the mortar element method, Asymptotic and numerical methods for partial differential equations with critical parameters (Beaune, 1992) 269-286, NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., 384, Kluwer Acad. Publ., Dordrecht, 1993.

[3] Cayco, M., Foster, L., Swann, H.: On the convergence rate of the cell discretization algorithm for solving elliptic problems, Math. Comp, **64**, 1397-1419 (1995).

[4] Dorr, M. R.: On the discretization of interdomain coupling in elliptic boundary-value problems. In T.F. Chan, R. Glowinski, J. Periaux and O. B. Widlund, editors. Domain Decomposition Methods, SIAM, 1989.

[5] Greenstadt, J.: Cell discretization, in Conference on Applications of Numerical Analysis, J.H. Morris, ed., Lecture Notes in Mathematics, 228, Springer-Verlag, New York, 1971, 70-82.

[6] Greenstadt, J.: The cell discretization algorithm for elliptic partial differential equations, SIAM J. Sci. Stat. Comput. Vol **3**, no. 3,(1982), 261-288.

[7] Hui, G., Swann, H.: On orthogonal polynomial bases for triangles and tetrahedra invariant under the symmetric group, Contemp. Math. vol. 218, (1998) 438-446.

[8] Raviart, P. A., Thomas, J. M.: Primal hybrid finite element methods for second order elliptic equations, Math. Comp., vol. **31** no. 138, (1977), 391-413.

[9] Rice, J. R., Vavalis, E. A., Yang, D.: Analysis of a nonoverlapping domain decomposition method for elliptic partial differential equations, J. Comp. Appl. Math. **87**, (1997), 11-19.

[10] Swann, H.: On the use of Lagrange multipliers in domain decomposition for solving elliptic problems, Math. Comp. **60**, No. 201, Jan, 1993, 49-78.

[11] Swann, H.: Error estimates using the cell discretization method for some parabolic problems, J. Comp. Appl. Math. **66**, (1996) 497-514

[12] Swann, H.: Error estimates using the cell discretization method for second-order hyperbolic equations, Numer. Methods Partial Differential Eq. **13**, (1997), 531-548.

[13] Swann, H.: Error estimates using the cell discretization method for steady-state convection-diffusion equations, J. Comput. Appl. Math, **82**, (1997), 389-405.

[14] Swann, H.: On approximating the solution of the stationary Stokes equations using the cell discretization algorithm, submitted, Numer. Methods Partial Differential Eq., 1998.

[15] Swann, H.: On using the cell discretization algorithm for mixed boundary value problems and domain decomposition, to appear, J. Comput. Appl. Math., Fall, 1999.

[16] Swann, H.: On approximating the solution of the non-stationary Stokes equations using the cell discretization algorithm, preliminary report, 1998, San Jose State Univ.

Accuracy, Resolution, and Computational Complexity of a Discontinuous Galerkin Finite Element Method

H. van der Ven* and J.J.W. van der Vegt†

*National Aerospace Laboratory NLR, P.O. Box 90502, 1006 BM Amsterdam, The Netherlands
†University of Twente, Faculty of Mathematical Sciences, P.O. Box 217, 7500 AE, Enschede, The Netherlands

Abstract. An analysis of the balance between the computational complexity, accuracy, and resolution requirements of a discontinuous Galerkin finite element method for the solution of the compressible Euler equations of gas dynamics is presented. The discontinuous Galerkin finite element method uses a very local discretization, which remains second order accurate on highly non-uniform meshes, but at the cost of an increase in computational complexity and memory use. The question of the balance between computational complexity and accuracy is addressed by studying the evolution of vortices in the wake of a wing. It is demonstrated that the discontinuous Galerkin finite element method on locally refined meshes can result in a significant reduction in computational cost.

1 Introduction

The accurate calculation of small scale flow structures presents a great challenge to computational fluid dynamics. Wake vortices, shocks, and the viscous sublayer in wall-bounded flows require a resolution which is orders of magnitude finer than in other regions of the flow. Efficient simulation of such structures is only feasible on highly non-uniform meshes, which are refined in the regions of interest. Accurate simulation of the flow structures on locally refined meshes is possible using Discontinuous Galerkin (DG) methods.

Discontinuous Galerkin finite element methods result in a very local discretization, which combines well with h-refinement because it maintains accuracy on non-smooth grids. The discontinuous Galerkin finite element method is, however, considerably more expensive, both in terms of computational complexity and memory usage, in comparison with the more commonly used finite volume methods. The key question to be addressed in this paper is whether for specific fluid dynamics problems, with vastly different length scales in two or more directions, the computational complexity of the DG method is more than compensated by its accuracy.

The balance between accuracy, resolution, and computational complexity of the DG finite element method is investigated by studying its efficiency in capturing the vortices in the wake of a wing. Numerical dissipation and

insufficient grid resolution cause serious problems in capturing vortical structures, and result in a smearing and decay of the vortical structures at some distance behind the wing. For many applications it is very important to be able to trace these vortical structures over a large distance downstream.

The outline of the paper is as follows. After a short description of the algorithm, the computational complexity of the method is analyzed. Subsequently its accuracy on highly non-uniform meshes is assessed for vortical flow. Finally, the balance between computational complexity and accuracy will be addressed.

2 Numerical Method

The numerical method used in the present investigation combines a discontinuous Galerkin discretization for the spatial discretization with a TVD-Runge Kutta time integration method and multigrid acceleration. This technique has received considerable theoretical interest during the last decade. Especially the work of Cockburn, Shu, et al. [1,2], significantly contributed to its theoretical development. In a series of papers van der Vegt and van der Ven [4–6] further developed the DG finite element method into a second order accurate numerical technique for the solution of the three-dimensional Euler equations of compressible gas dynamics on highly non-uniform hexahedral meshes. This method is used in the present investigation.

The most computationally intensive part of the method are the element face flux integrals. The straightforward computation of the face flux integrals requires four point Gauss quadrature rules for second order accuracy. Van der Vegt et al. [6] proposed an approximation to the flux integrals of the form

$$\int_S F(U_h) \cdot n\phi_m dx \approx F(\overline{U}_h^S) \cdot \int_S n\phi_m dx,$$

where S is a face, U_h is the state vector of the Euler equations, F is the flux function, n is the face normal, ϕ_m $(0 \leq m \leq 3)$ is the m-th basis function in the cell bounding S, and $\overline{U}_h^S$ is the face average. The volume fluxes are approximated likewise. Van der Vegt et al. [6] proved that second order accuracy is retained when the geometric terms are computed exactly. These approximations result in a number of flux calculations which is approximately equal to finite volume methods, and in a reduced computational complexity.

The algorithm is efficiently implemented in the program HEXADAP using a face based data structure, which allows full vectorization and parallelization of the code. The parallel performance of the code is further improved with a dynamic domain decomposition technique, which automatically redistributes the elements over the processors after grid adaptation [7,8]. The computational efficiency is further improved by local time stepping (with CFL=0.7) and a multigrid convergence acceleration algorithm, which uses a first order accurate scheme on the coarser grid levels.

	storage
flow field	$(3nm + n + 4)$R
geometry	84 R
topology	91 I
total	200 words

Table 1. Memory requirements per grid cell for the DG method. The following notations are used: n is the number of flow variables ($n = 5$), m the number of basis functions ($m = 4$), R refers to real variables (8 Bytes), I to integer variables (4 Bytes). Totals are in 8 Bytes words.

The DG finite element method results in an accurate discretization, but with increased memory use and a significantly larger computational complexity than finite volume methods. The DG method also solves equations for the three higher moments for all five variables of the Euler equations and stores these variables. This results in 20 degrees of freedom per grid cell, four times more than for finite volume methods.

3 Computational complexity

In this section the computational complexity of the DG method is analyzed and compared to a well-tuned finite volume Jameson algorithm implemented in the multi-block structured flow solver ENFLOW [3]. Since the Jameson algorithm is optimal in terms of computational complexity, the reader should be aware that this is the strictest comparison possible.

The memory requirements of the above DG method are tabulated in Table 1. The memory is split into three parts: flow field, geometry (grid points, mass matrix, element integrals, etc.) and topology. The latter is required since the hexahedron grids are unstructured. The block-structured finite volume flow solver ENFLOW requires approximately 20 words per cell. The second order DG flow solver HEXADAP on unstructured meshes has four times more degrees of freedom and requires 2.5 times more memory per degree of freedom than the block-structured finite volume flow solver.

The number of floating point operations per grid cell per fine grid iteration (including coarse grid corrections)of HEXADAP is 21 kflop, that is, 5 kflop per degree of freedom per fine grid iteration. In Table 2 the main components of the computation and their respective work load is shown. The main part is the Osher flux difference scheme, followed closely by the slope limiter. Note that the solution of the moment equations constitutes 20 % of the work load. Certain geometric contributions are recomputed at each stage in the Runge-Kutta scheme, amounting to 10% of the work load. Note that since the face flux computations constitute about 50% of the work load, a four point quadrature rule for the flux evaluation would increase the total work load by a factor of 2.5. Hence the above approximations to the flux integrals significantly reduce the computational complexity. The single pro-

	work load	average performance
Osher scheme	33 %	800
slope limiter	25 %	400
flow moments	20 %	1450
geometric contributions	10 %	400
left and right states	7 %	500
Runge Kutta	5 %	1500

Table 2. Distribution of work and average single processor vector performance (in Mflop/s) in the DG method

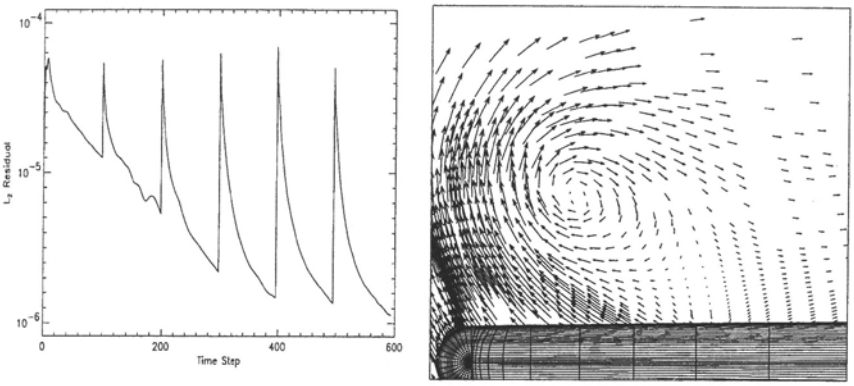
cessor vector performance is 600 Mflop/s (30% peak) on average on a NEC SX-4 and is mainly bounded by memory access.

The finite volume flow solver ENFLOW requires 2 kflop per grid cell per fine grid iteration. The unstructured DG method has, however, four times more degrees of freedom and is 2.5 times more computationally expensive, per degree of freedom, than ENFLOW. In the next section it will be shown that the DG method is accurate on highly non-uniform grids, which require significantly less elements than structured grids.

4 Results

The balance between computational complexity and accuracy of the DG finite element method discussed in this paper is investigated by calculating the flow field about a generic wing at a free stream Mach number $M_\infty = 0.84$ and angle of attack $\alpha = 3.06°$. The wake vortices of the wing are difficult to capture over a large distance, especially if the grid is not aligned with the vortex core. After calculating an initial solution on a grid of 130,000 cells, the grid is adapted five times until a grid with 250,000 cells is obtained. After each adaptation the mesh is repartitioned for parallel load balance, and the flow is advanced for 100 multigrid cycles, see Figure 1(a). Grid refinement is only performed in the wake and not over the wing. The grid is refined at the vortex using a vortex sensor based on vortex strength and the total pressure loss. Note that the derivatives of the velocity are directly available in each cell, since the DG method has 20 degrees of freedom per cell, including a fully resolved gradient of the state vector.

Figure 1(b) shows the vortex in a cross-section at $x = 3$ (the wing tip is located at $x = 1.4$ and the wing span is three). In Figure 2 the vortex, shown as streamlines, at a cross-section one and half wing span behind the wing tip on the one time refined mesh is compared with the results on the final refined mesh. It can be clearly seen that the vortex is better resolved and extends further downstream. Figure 3 shows the final adapted grid at $x = 3$, $x = 6$, and $x = 9$. Note that the initial mesh is not aligned with the vortex core, but the grid refinement accurately captures the vortex.



(a) Convergence history (b) Vortex at $x = 3$

Fig. 1. Multigrid convergence history of the L_2 residuals of the means and flow field of a generic wing ($M_\infty = 0.84$, $\alpha = 3.06°$). Only the residuals in the wake are measured.

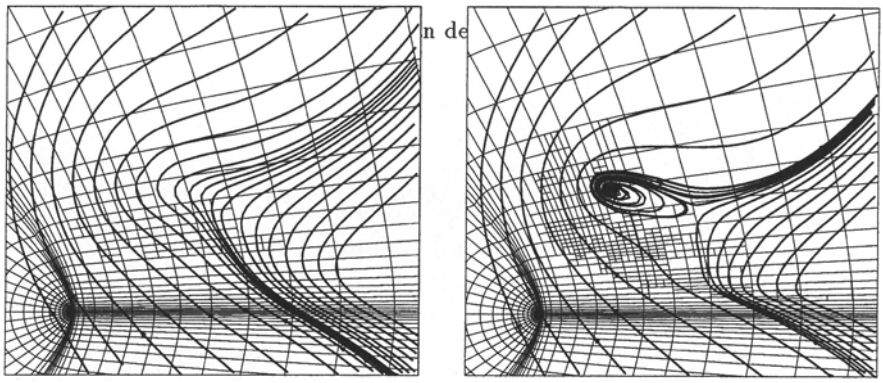
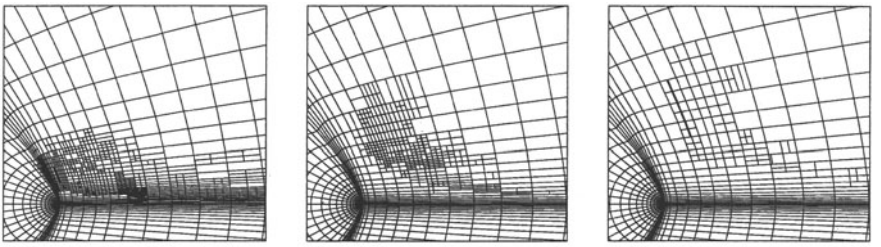


Fig. 2. Vortex comparison on one time refined grid (left) and final refined grid (right) at $x = 6$. The vortex is visualized using streamlines.

Locally at the vortex the initially structured grid is refined twice in all three directions. In case of uniform refinement in, say, a quarter of the mesh, a structured grid with the same resolution would require at least $\frac{1}{4} \cdot 64 = 16$ times more grid points than the adapted grid (and would be difficult to generate since the vortex position is unknown beforehand). Hence the higher computational complexity of the DG method using locally refined meshes is compensated by its accuracy on non-uniform meshes.

5 Conclusions

The DG method is efficient on highly non-uniform meshes and, combined with grid adaptation, is able to trace wake vortices over large distances. The computational complexity of the DG method per degree of freedom is



(a) mesh at $x = 3.0$ (b) mesh at $x = 6.0$ (c) mesh at $x = 9.0$

Fig. 3. Three cross-sections of the final refined non-uniform mesh behind a generic wing.

2.5 times the complexity of a finite volume block-structured method, both in flop count and memory use. This factor is similar to the difference between structured and unstructured (tetrahedra) finite volume schemes. The increased complexity of the unstructured DG method is compensated by its accuracy on non-uniform, locally refined, grids which require significantly less grid cells for the same resolution.

References

1. B. Cockburn, S. Hou and C.-W. Shu, The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: The multidimensional case, *Math. Comput.* **54**, 545 (1990).
2. B. Cockburn and C.-W. Shu, The Runge-Kutta discontinuous Galerkin finite element method for conservation laws V: Multidimensional systems, *J. Comput. Phys.* **141**, 199 (1998).
3. J.C. Kok, J.W. Boerstoel, A. Kassies and S.P. Spekreijse, A robust multi-block Navier-Stokes flow solver for industrial applications, in Proceedings 3rd ECCOMAS CFD Conference, September 1996, John Wiley and Sons, Chichester.
4. J.J.W. van der Vegt, Anisotropic grid refinement using an unstructured discontinuous Galerkin method for the three-dimensional Euler equations of gas dynamics, in *Proc. 12th AIAA CFD Conference, San Diego, California, 1995.* [AIAA Paper 95-1657-CP]
5. J.J.W. van der Vegt and H. van der Ven, Hexahedron based grid adaptation for future large eddy simulation, in *Proc. Progress and Challenges in CFD Methods and Algorithms, Seville, Spain, 1995.* [AGARD CP-578, p. 22-1]
6. J.J.W. van der Vegt and H. van der Ven, Discontinuous Galerkin finite element method with anisotropic local grid refinement for inviscid compressible flows, *J. Comput. Phys.* **141**, 46 (1998).
7. H. van der Ven and J.J.W. van der Vegt, Partitioning and parallel development of an unstructured, adaptive flow solver on the NEC SX-4, in *Proc. Parallel Computational Fluid Dynamics '97 Conference, Manchester, England, 1997.*
8. H. van der Ven and J.J.W. van der Vegt, Experiences with Advanced CFD Algorithms on NEC SX-4, in *Proc. Vector and Parallel Processing VECPAR '96*, edited by Palma and Dongarra, Lect. Notes in Computer Science, (Springer Verlag, 1997).

An ELLAM Scheme for Porous Medium Flows

Hong Wang

Department of Mathematics, University of South Carolina, Columbia, South Carolina 29208, USA

Abstract. An Eulerian-Lagrangian localized adjoint method (ELLAM) is presented for coupled systems of fluid flow processes occurring in porous media with point sources and sinks. The ELLAM scheme symmetrizes the governing transport equation, greatly eliminates non-physical oscillation and/or excessive numerical dispersion present in many large-scale simulators widely used in industrial applications. It can treat large mobility ratios, discontinuous permeabilities and porosities, anisotropic dispersion in tensor form, and point sources and sinks. It also conserves mass. Numerical experiments are presented. The relationship between ELLAM and discontinuous Galerkin methods (DGMs), and the possibility of developing a hybrid ELLAM-DGM scheme are discussed.

1 A Mathematical Model

Let $p(\mathbf{x},t)$ and $\mathbf{u}(\mathbf{x},t)$ be the pressure and Darcy velocity of a fluid mixture, and $c(\mathbf{x},t)$ be the concentration (volume fraction) of an invading fluid or concerned solute/solvent in the fluid mixture. The equation of mass conservation for the fluid mixture incorporated with the incompressibility condition, Darcy's law, and the equation of mass conservation for the concerned component yields a coupled system of partial differential equations (PDEs) that describes incompressible fluid flow processes in a porous medium reservoir Ω with point sources and sinks (injection and production wells) [1,6]

$$\begin{aligned} \nabla \cdot \mathbf{u} &= q, && \mathbf{x} \in \Omega, \quad t \in (0,T] \ , \\ \mathbf{u} &= -\frac{\mathbf{K}}{\mu(c)}(\nabla p - \rho g \nabla d), && \mathbf{x} \in \Omega, \quad t \in (0,T] \ , \end{aligned} \tag{1}$$

$$\phi \frac{\partial c}{\partial t} + \nabla \cdot (\mathbf{u}c - \mathbf{D}(\mathbf{u})\nabla c) = qc^*, \qquad \mathbf{x} \in \Omega, \quad t \in (0,T] \ . \tag{2}$$

In many cases, the thickness of the medium is significantly smaller than its length and width. Hence, it is reasonable to average the medium properties vertically and to assume $\Omega \subset \mathbb{R}^2$ with a nonuniform local elevation. $\mathbf{K}(\mathbf{x})$ is the permeability tensor of the medium, $\mu(c)$ is the concentration-dependent viscosity of the fluid mixture, which is determined by some mixing rule

$$\mu(c) = \mu_o[(1-c) + M^{\frac{1}{4}}c]^{-4} \tag{3}$$

where μ_o is the viscosity of the resident fluid and M is the mobility ratio. ρ is the density of the fluid mixture, g is the magnitude of gravitational

acceleration, $d(\mathbf{x})$ is the reservoir depth, $q(\mathbf{x},t)$ is external source and sink term that accounts for the effect of injection and production wells, $\phi(\mathbf{x})$ is the porosity of the medium (proportion of volume available to porous medium flows), c^* is a prescribed concentration at sources or is equal to c at sinks. $\mathbf{D}(\mathbf{u})$ is the diffusion-dispersion tensor that consists of molecular diffusion and (anisotropic velocity-dependent) mechanical dispersion

$$\mathbf{D}(\mathbf{u}) = \phi(\mathbf{x})d_m\,\mathbf{I} + d_t|\mathbf{u}|\,\mathbf{I} + \frac{d_l - d_t}{|\mathbf{u}|}\begin{pmatrix} u_x^2 & u_x u_y \\ u_x u_y & u_y^2 \end{pmatrix}, \tag{4}$$

where $\mathbf{u} = (u_x, u_y)$, d_m is the molecular diffusion coefficient, $\mathbf{I}$ is the identity tensor, and d_t and d_l are the transverse and longitudinal dispersivities, respectively.

System (1)–(2) needs to be closed by the initial and boundary conditions. In petroleum reservoir simulation the boundary $\partial\Omega$ is typically impermeable, leading to no-flow boundary conditions of the form [11]

$$\begin{aligned} \mathbf{u}\cdot\mathbf{n} &= 0, \quad (\mathbf{x},t) \in \partial\Omega \times [0,T]\ , \\ (\mathbf{D}(\mathbf{u},p)\nabla c)\cdot\mathbf{n} &= 0, \quad (\mathbf{x},t) \in \partial\Omega \times [0,T]\ . \end{aligned} \tag{5}$$

These conditions also arise in environmental modeling although other types of boundary conditions are possible [1]. For simplicity, we assume boundary conditions (5) and a rectangular domain $\Omega = (a_x, b_x) \times (a_y, b_y)$ [1,11].

Because diffusion or dispersion is often a small phenomenon relative to advection, Equation (2) is an advection-diffusion equation with advection being the dominant phenomenon. Additional features of (1)–(2) include the singularities of the solutions at point sources and sinks, discontinuous permeabilities and porosities, a large adverse mobility ratio in the flow processes that could cause viscous fingering phenomena, anisotropic dispersion in tensor form, as well as the enormous size of field-scale applications.

2 An ELLAM Scheme

We utilize a sequential decoupling and linearization technique for system (1) and (2), and a mixed finite element method to solve p and $\mathbf{u}$ from (1) with the value of c being taken at the current time step [6,7]. For simplicity, in this section we describe only an ELLAM scheme for (2) assuming that the pressure p and the Darcy velocity $\mathbf{u}$ in (2) are known.

Let $0 = t_0 < t_1 < \ldots < t_n < \ldots < t_{N-1} < t_N = T$ be a partition of the time interval $[0,T]$ with $\Delta t_n = t_n - t_{n-1}$. In the ELLAM formulation, we multiply (2) by space-time test functions w that are continuous and piecewise smooth, vanish outside the space-time strip $\overline{\Omega} \times (t_{n-1}, t_n]$, and are

discontinuous in time at time t_{n-1}. This yields a space-time weak formulation

$$\begin{aligned}
&\int_\Omega \phi(\mathbf{x})c(\mathbf{x},t_n)w(\mathbf{x},t_n)\,d\mathbf{x} + \int_{t_{n-1}}^{t_n}\int_\Omega (\nabla w\cdot \mathbf{D}(\mathbf{u})\nabla c)(\mathbf{y},\theta)\,d\mathbf{y}d\theta \\
&\quad - \int_{t_{n-1}}^{t_n}\int_\Omega c(\mathbf{y},\theta)\left[\phi(\mathbf{y})\frac{\partial w(\mathbf{y},\theta)}{\partial\theta} + \mathbf{u}(\mathbf{y},\theta)\cdot\nabla w(\mathbf{y},\theta)\right]d\mathbf{y}d\theta \qquad (6)\\
&\quad = \int_\Omega \phi(\mathbf{x})c(\mathbf{x},t_{n-1})w(\mathbf{x},t_{n-1}^+)\,d\mathbf{x} + \int_{t_{n-1}}^{t_n}\int_\Omega (c^*qw)(\mathbf{y},\theta)\,d\mathbf{y}d\theta\ ,
\end{aligned}$$

where $w(\mathbf{y},t_{n-1}^+) = \lim_{\theta\to t_{n-1},\theta>t_{n-1}} w(\mathbf{y},\theta)$ takes into account the fact that $w(\mathbf{x},t)$ is discontinuous in time at time t_{n-1}.

Careful analysis of various operator splittings in the ELLAM framework concludes that the test functions $w(\mathbf{y},\theta)$ in (6) should be chosen to satisfy the hyperbolic part of the adjoint equation of (2) [2]

$$\phi(\mathbf{y})\frac{\partial w(\mathbf{y},\theta)}{\partial\theta} + \mathbf{u}(\mathbf{y},\theta)\cdot\nabla w(\mathbf{y},\theta) = 0,\quad \mathbf{y}\in\overline{\Omega},\quad \theta\in[t_{n-1},t_n]\ . \qquad (7)$$

Equation (2) implies that the test functions $w(\mathbf{y},\theta)$ should be constant along the characteristics $\mathbf{y} = \mathbf{r}(\theta;\mathbf{x},t_n)$, defined by the differential equation

$$\begin{aligned}
\frac{d\mathbf{r}}{d\theta} &= \frac{\mathbf{u}(\mathbf{r},\theta)}{\phi(\mathbf{r})},\quad \theta\in[t_{n-1},t_n]\ ,\\
\mathbf{r}(\theta;\mathbf{x},t_n)\Big|_{\theta=t_n} &= \mathbf{x}\ .
\end{aligned} \qquad (8)$$

In the ELLAM scheme, we choose the test functions $w(\mathbf{x},t_n)$ to be piecewise-bilinear functions for $\mathbf{x}\in\overline{\Omega}$ at time t_n and define them by constant extension along the characteristics $\mathbf{r}(\theta;\mathbf{x},t_n)$ to the space-time strip $\overline{\Omega}\times(t_{n-1},t_n]$.

We enforce the Euler quadrature at time t_n to evaluate the source and sink term in (6). Note that for any $(\mathbf{y},\theta)\in\Omega\times[t_{n-1},t_n]$, there exists an $\mathbf{x}\in\Omega$ such that $\mathbf{y} = \mathbf{r}(\theta;\mathbf{x},t_n)$. Hence,

$$\begin{aligned}
&\int_{t_{n-1}}^{t_n}\int_\Omega c^*(\mathbf{y},\theta)q(\mathbf{y},\theta)w(\mathbf{y},\theta)\,d\mathbf{y}d\theta\\
&\quad = \int_\Omega\int_{t_{n-1}}^{t_n} c^*(\mathbf{r}(\theta;\mathbf{x},t_n),\theta)q(\mathbf{r}(\theta;\mathbf{x},t_n),\theta)w(\mathbf{x},t_n)\left|\frac{\partial(\mathbf{r},\theta)}{\partial(\mathbf{x},t_n)}\right|d\theta d\mathbf{y} \qquad (9)\\
&\quad = \Delta t_n\int_\Omega c^*(\mathbf{x},t_n)q(\mathbf{x},t_n)w(\mathbf{x},t_n)d\mathbf{x} + E_q(c^*,w)\ ,
\end{aligned}$$

where $\left|\frac{\partial(\mathbf{r},\theta)}{\partial(\mathbf{x},t_n)}\right| = 1+\mathcal{O}(t_n-\theta)$ is the Jacobian of the transformation, $E_q(c^*,w)$ is the local truncation error.

We can evaluate the diffusion-dispersion term similarly and obtain

$$\begin{aligned}
&\int_{t_{n-1}}^{t_n}\int_\Omega \nabla w(\mathbf{y},\theta)\cdot\mathbf{D}(\mathbf{u}(\mathbf{y},\theta))\nabla c(\mathbf{y},\theta)\,d\mathbf{y}d\theta\\
&\quad = \Delta t_n\int_\Omega \nabla w(\mathbf{x},t_n)\cdot\mathbf{D}(\mathbf{u}(\mathbf{x},t_n))\nabla c(\mathbf{x},t_n)\,d\mathbf{x} + E_{\mathbf{D}}(c,w)\ ,
\end{aligned} \qquad (10)$$

where $E_{\mathbf{D}}(c, w)$ is the local truncation error term.

Because of (7), the last term on the right-hand side of (6) vanishes if the characteristics defined by (8) are calculated exactly or is within the tolerance if they are approximated numerically [13]. In the ELLAM scheme, we substitute (9) and (10) into (6) and drop $E_q(c^*, w)$, $E_{\mathbf{D}}(c, w)$, and the last term on the left-hand side of (6). We define the trial functions $c(\mathbf{x}, t_n)$ to be piecewise bilinear functions on $\overline{\Omega}$ at time step t_n as in the standard finite element method. Note that the trial functions coincide with the test functions on $\overline{\Omega}$ at time level t_n. But the trial functions c are defined at time step t_n only while the test functions w are defined on the space-time strip $\overline{\Omega} \times (t_{n-1}, t_n]$ by constant extension along characteristics from $\overline{\Omega}$ at time t_n. This leads to the following ELLAM scheme

$$\begin{aligned} &\int_\Omega \phi(\mathbf{x}) c(\mathbf{x}, t_n) w(\mathbf{x}, t_n)\, d\mathbf{x} + \Delta t_n \int_\Omega (\nabla w \cdot \mathbf{D}(\mathbf{u}) \nabla c)(\mathbf{x}, t_n)\, d\mathbf{x} \\ &\quad = \int_\Omega \phi(\mathbf{x}) c(\mathbf{x}, t_{n-1}) w(\mathbf{x}, t_{n-1}^+)\, d\mathbf{x} + \Delta t_n \int_\Omega (c^* q w)(\mathbf{x}, t_n)\, d\mathbf{x} \ . \end{aligned} \tag{11}$$

Except for the first term on the right-hand side, all other terms in (11) are standard in finite element method and can be computed in a straightforward manner. In this term, the trial and test functions are actually defined at different time steps. Hence, the evaluation of this term is very challenging and raises various numerical difficulties [10]. We refer interested readers to [14] for detailed implementational issues. The ELLAM scheme (11) symmetrizes the transport PDE (2) and yields a 9-banded, symmetric and positive definite coefficient matrix. It generates accurate numerical solutions even if large time steps are used, and conserves mass [2].

3 Numerical Experiments

We previously carried out extensive numerical experiments in the context of linear transport PDEs with known analytical solutions. The numerical comparison shows that the ELLAM outperforms many widely used and well received methods [14,15]. In this paper, we apply the ELLAM scheme (11) to solve the system (1)–(2). Because no analytical solutions are available, we choose test problems with reported data and results in the literature [6,7]. We also compare the numerical solutions with those obtained using finer grids to verify their accuracy. The test problem is a standard five spot pattern displacement in reservoir simulation. Previously, the time steps used in these simulations range from a few days for the upwind finite difference method (UFDM) to one month for the modified method of characteristics (MMOC) [6,7]: The spatial domain $\Omega = (0, 1000) \times (0, 1000)$ ft^2, the porosity $\phi = 0.1$, the permeability coefficients (diagonal entries) are $k_x = k_y = 80$ md, the viscosity of oil $\mu_o = 1.0$ cp, the mobility ratio $M = 41$, the molecular diffusion is $D_m = \phi d_m = 0$, the longitudinal and transverse dispersions are

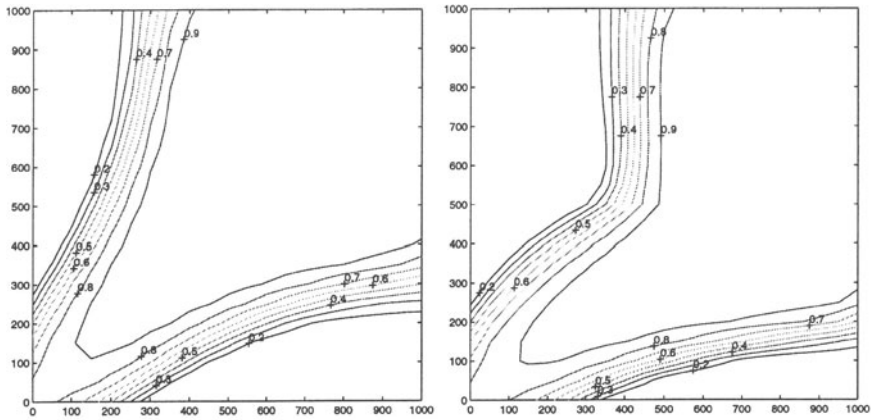
$D_l = \phi d_l = 5.0$ ft and $D_t = \phi d_t = 0.5$ ft, respectively. The injection well is located at the upper-right corner $(1000, 1000)$ with an injection rate of $q = 30$ ft^2/day and $c^* = 1.0$. The production well is located at the bottom-left corner $(0, 0)$ with a production rate of $q = 30$ ft^2/day. The initial concentration is $c_0(x, y) = 0$. We use a spatial grid of $\Delta x = \Delta y = 50$ ft, and a time step of $\Delta t = 360$ days (one year). The numerical result at 10 years is presented in Fig. 1(a). In Fig. 1(b), $k_x = k_y = 80$ md on $(0, 1000) \times (0, 500)$ and $k_x = k_y = 20$ md on $(0, 1000) \times (500, 1000)$ are used. These results show that with a much larger time step (one year compared to a few days with the UFDM and a month with the MMOC), the ELLAM scheme still generates accurate solutions with a significantly improved efficiency.

4 Connections to DGMs and Future Direction

Since it was proposed in early 1970s [9,12], the DGMs have shown their strength. Extensive studies have been carried out on DGMs [3–5,8]. The ELLAM scheme presented in this paper is related to DGMs in the following sense: The ELLAM formalism starts from a space-time framework, but eventually yields a numerical scheme defined only on $\overline{\Omega}$ at each time step t_n. DGMs use space-time trial and test functions that are discontinuous (say, in time at each time step t_n) to decouple the resulting discrete system in time. The ELLAM scheme uses test functions that are discontinuous in time at time step t_{n-1} to break the coupling of the discrete system in time, but still use trial functions that are defined uniquely at each time step t_n. Finally, both DGMs and ELLAM generate accurate numerical solutions with correct physical behavior. Because of the success the DGMs have demonstrated in the solution of nonlinear hyperbolic conservation laws (e.g. resolution of shock discontinuities) and the strength the ELLAM scheme has illustrated (e.g., generation of accurate solutions even if very large time steps are used), the development of a hybrid ELLAM-DGM scheme that uses discontinuous spatial approximations and possesses advantages of both DGMs and ELLAM schemes is under investigation.

References

1. Bear, J.: Hydraulics of Groundwater. McGraw-Hill, New York, 1979
2. Celia, M.A., Russell, T.F., Herrera, I., Ewing R.E.: An Eulerian-Lagrangian localized adjoint method for the advection-diffusion equation. Adv. Wat. Res. **13** (1990) 187–206
3. Chavent, G, Cockburn, B.: The local projection P^1-discontinuous Galerkin finite element method for scalar conservation laws. M^{2}AN **23** (1989) 565–592
4. Chen, Z., Cockburn, B., Jerome, J.W., Shu C.W.: Mixed-RKDG finite element methods for the 2-D hydrodynamic model for semiconductor device simulation. VLSI Designs **3** (1995) 1–14



(a) $k_x = k_y = 80$ md (b) piecewise permeability

Fig. 1. 10 year simulation with a time step of 1 year

5. Cockburn, B., Shu C.W.: TVB Runge-Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws II: general framework. Math. Comp. **52** (1989) 411–435
6. Ewing, R.E. (ed.): The Mathematics of Reservoir Simulation, Research Frontiers in Applied Mathematics. **1**, SIAM, Philadelphia, 1984
7. Ewing, R.E., Russell, T.F., Wheeler, M.F.: Simulation of miscible displacement using mixed methods and a modified method of characteristics. SPE 12241 (1983) 71–81
8. Falk, R., Richter, G.R.: Local error estimates for a finite element method for hyperbolic and convection-diffusion equations. SIAM J. Numer. Anal. **29** (1992) 730–754
9. LeSaint, P., Raviart, P.A.: On a finite element method for solving the neutron transport equation. DeBoor, C. (ed.), Mathematics Aspects of Finite Elements in Partial Differential Equations, Academic Press, (1974) 89–123
10. Morton, K.W., Priestley. A., Süli, E.: Stability of the Lagrangian-Galerkin method with nonexact integration. RAIRO M^2AN **22** (1988) 123–151
11. Peaceman, D.W.: Fundamentals of Numerical Reservoir Simulation. Elsevier, Amsterdam, 1977
12. Reed, W.H., Hill, T.R.: Triangular mesh methods for the neutron transport equation. Los Alamos Scientific Laboratory Report LA-UR-73-479, 1973
13. Wang, H.: A family of ELLAM schemes for advection-diffusion-reaction equations and their convergence analyses. Numer. Meth. PDEs **14** (1998) 739–780
14. Wang, H., Dahle, H.K., Ewing, R.E., Espedal, M.S., Sharpley, R.C., Man, S.: An Eulerian-Lagrangian localized adjoint method for advection-dispersion equations in two dimensions and its comparison to other schemes. SIAM J. Sci. Comp. (in press)
15. Wang, H., Ewing, R.E., Qin, G., Lyons, S.L., Al-Lawatia, M, Man, S: A family of Eulerian-Lagrangian localized adjoint methods for multi-dimensional advection-reaction equations. J. Comp. Phys. **152** (1999) 120–163

Application of the Discontinuous Galerkin Method to Maxwell's Equations Using Unstructured Polymorphic hp-Finite Elements

Tim Warburton

Oxford University Computing Laboratory, Oxford OX1 3QD, UK

Abstract. In this paper we demonstrate the efficiency of using the discontinuous Galerkin method for simulating electromagnetic scattering problems using Maxwell's equations. We show that it is possible to use unstructured *hp*-finite elements in mixed-element (polymorphic) grids. We include examples of scattering from a two-dimensional cylinder and preliminary results from a three-dimensional F15 geometry.

1 Introduction

High order finite volume methods, also known as discontinuous Galerkin methods, have been developed for a variety of problems starting in the 1970s for studying the neutron transport equation [13]. Subsequently they have been applied to solving hyperbolic conservation laws [4,5],[2], Euler and Navier-Stokes equations [10,1], amongst many other areas. These algorithms provide robust methods for coupling together piecewise, high-order polynomial approximations on unstructured finite element discretizations using flux jump functions in a variational setting. This paper extends the work of [17] to allow the use of flexible unstructured finite elements of arbitrary variable order.

We have developed a computer code, using these techniques, for solving the time-domain Maxwell's equations. We used this solver for calculations of electromagnetic wave scattering from complex geometries in two- and three-dimensions. Our implementation allows the use of unstructured triangles, quadrilaterals, tetrahedra, prisms and hexahedra for discretizing the volume surrounding the scatterer.

2 Elemental Description

We have adopted the elemental description of [7] which allows us to use unstructured triangles and quadrilaterals for discretizing a two-dimensional domain or tetrahedra, prisms and hexahedra for three-dimensional domains. The physical elements described as physical volumes which are mapped to a reference element, which are in turn mapped to a tensor product element $[-1,1]^{dim}$ where dim is the dimension of the space. In this framework it is

possible to integrate and differentiate a function, described by a set of values at quadrature points, with a computational cost of $O(Q^{dim+1})$ where Q is the order of the quadrature used.

The coordinate systems for the three-dimensional elements can be matched using fast algorithms [15]. The discontinuous Galerkin formulation (DGM) requires surface flux integrals to be evaluated. We calculate these quantities at a set of tensor-product Gauss nodes, arising naturally from the underlying tensor coordinates for the element. The Gauss distributed nodes will lie in at most two elements, reducing the communication cost in our parallel implementation of the algorithm. Since coordinate systems are chosen to match between elements, the Gauss nodes will also match.

In each element we use a basis consisting of a set of orthogonal polynomials which are non-regularly weighted Jacobi polynomials variously rediscovered by many, including [12,8,6,14,11,16]. The truncation error from projecting a function onto subsets of these bases decreases exponentially as the size of the subset is increased as long as the function is sufficiently smooth.

Each element has a useful purpose, for instance triangular elements allow very complex shapes to be meshed. Quadrilateral elements will support larger deformations under iso-parametric mapping when an element is used to represent a curved surface.

3 Formulation

The time-domain Maxwell equations for electric and magnetic fields can be written as:

$$\frac{\partial(\mu\mathbf{H})}{\partial t} = -\nabla \times \mathbf{E} - \rho\mathbf{H}$$
$$\frac{\partial(\epsilon\mathbf{E})}{\partial t} = \nabla \times \mathbf{H} - \sigma\mathbf{E}$$

where μ is the magnetic permeability, ϵ is the electric permittivity, ρ is the equivalent magnetic resistivity, σ is the electric conductivity, $\mathbf{H}$ is the magnetic field vector and $\mathbf{E}$ is the electric field vector.

We now define the state vectors:

$$\mathbf{s} = (\mu H_x, \mu H_y, \mu H_z, \epsilon E_x, \epsilon E_y, \epsilon E_z)^t = (B_x, B_y, B_z, D_x, D_y, D_z)^t$$
$$\mathbf{v} = (H_x, H_y, H_z, E_x, E_y, E_z)^t$$

then we seek to solve the variational statement of the problem: for each element find $\mathbf{s} \in P_h$ such that

$$(\phi, \frac{\partial \mathbf{s}}{\partial t}) = (\phi, \mathbf{A}\frac{\partial \mathbf{v}}{\partial x} + \mathbf{B}\frac{\partial \mathbf{v}}{\partial y} + \mathbf{C}\frac{\partial \mathbf{v}}{\partial z}) + (\phi, \mathbf{D}(\hat{\mathbf{v}} - \mathbf{v}^-)) \qquad (1)$$

where:

$$\mathbf{A} = \begin{pmatrix} 0&0&0&0&0&0 \\ 0&0&0&0&0&1 \\ 0&0&0&0&-1&0 \\ 0&0&0&0&0&0 \\ 0&0&-1&0&0&0 \\ 0&1&0&0&0&0 \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} 0&0&0&0&0&-1 \\ 0&0&0&0&0&0 \\ 0&0&0&1&0&0 \\ 0&0&1&0&0&0 \\ 0&0&0&0&0&0 \\ -1&0&0&0&0&0 \end{pmatrix} \quad \mathbf{C} = \begin{pmatrix} 0&0&0&0&1&0 \\ 0&0&0&-1&0&0 \\ 0&0&0&0&0&0 \\ 0&-1&0&0&0&0 \\ 1&0&0&0&0&0 \\ 0&0&0&0&0&0 \end{pmatrix}$$

$$\mathbf{D} = n_x\mathbf{A} + n_y\mathbf{B} + n_z\mathbf{C}$$

and n_x, n_y and n_z are the x, y and z components of the outward facing normals on the element faces, and $\hat{\mathbf{v}}$ is the state vector chosen by characteristic treatment. Here, "$-$" represents the trace of the variable from the elemental side of the boundary, and "$+$" represents the trace of the variable from the exterior side.

The flux vector $\mathbf{D}(\hat{\mathbf{v}} - \mathbf{v}^-)$ through a face is simply the difference of the state vector either side of the face projected onto the inward-bound characteristics, which are the positive eigenvectors of the the matrix $\mathbf{D}$:

$$\mathbf{D}(\hat{\mathbf{v}} - \mathbf{v}^-) = (\mathbf{S}_1 \cdot (\mathbf{v}^+ - \mathbf{v}^-))\mathbf{S}_1 + (\mathbf{S}_2 \cdot (\mathbf{v}^+ - \mathbf{v}^-))\mathbf{S}_2$$

$$\begin{aligned} \mathbf{S}_1 &= \frac{1}{\sqrt{2(n_x^2+n_z^2)}}(-n_xn_y, n_x^2+n_z^2, -n_yn_z, -n_z, 0, n_x)^t \quad \text{if } n_x^2+n_z^2 > 0 \\ &= \frac{1}{\sqrt{2(n_y^2+n_z^2)}}(n_y^2+n_z^2, -n_xn_y, -n_xn_z, 0, n_z, -n_y)^t \quad \text{otherwise} \\ \mathbf{S}_2 &= \frac{1}{\sqrt{2(n_x^2+n_z^2)}}(n_z, 0, -n_x, -n_xn_y, n_x^2+n_z^2, -n_yn_z)^t \quad \text{if } n_x^2+n_z^2 > 0 \\ &= \frac{1}{\sqrt{2(n_y^2+n_z^2)}}(0, -n_z, n_y, n_y^2+n_z^2, -n_xn_y, -n_xn_z)^t \quad \text{otherwise} \end{aligned}$$

3.1 Discrete Scheme for CEM

The elemental inner-product on the right hand side of equation (1) can be evaluated in two ways, either with direct polynomial manipulation or by using a collocation, tensor product approach.

The first involves straight matrix-matrix multiplication: a vector of polynomial coefficients for the state vector on an element, can be multiplied by a matrix to calculate the coefficients of the "x" derivative of the polynomial.

This is efficient for polynomial orders up to about $P = 6$, and can be easily implemented to take advantage of data locality and cache reuse.

The second method uses the tensor product nature of the underlying coordinates of the elements. In this approach, we calculate derivatives and integrals in a set of one-dimensional operations.

The first method has an asymptotic operation count of $O(N^{2dim})$ (here N is the expansion order) but a relatively small constant compared to the second method which has an operation count of $O(Q^{dim+1})$ (here Q is the quadrature order).

In the examples section we have used an explicit, second-order accurate, Adams-Bashforth time-integration scheme. The method is not limited to using this scheme. For instance, a fourth order low storage explicit Runge-Kutta scheme was used in [9].

3.2 Boundary Conditions

For the study of scattering by reflectors we decompose the state vector into the sum of a prescribed incident wave and a reflected wave. We write the state vector as:

$$\mathbf{s} = \mathbf{s}_{sc} + \mathbf{s}_{inc}$$

As we assume the incident wave $\mathbf{s}_{inc}$ satisfies Maxwell's equations, then by linearity, we solve the same equations for the scattered waves $\mathbf{s}_{sc}$.

At the far field we assume the scattered waves tend to zero. There, we use a flux boundary condition by specifying zero for the external state vector used in the flux difference. The upwind routine is used as previously described. We also use an absorbing layer which increases (ρ, σ) quadratically from zero at $4\sqrt{2}$ diameters from the center of the cylinder. This damps both the waves travelling outwards and their reflected components as they travel inwards.

We assume the reflector is a perfectly conducting body. The boundary conditions consistent with the equations are:

$$\mathbf{B}^+_{sc} = -\mathbf{n}(\mathbf{B}_{inc} \cdot \mathbf{n}) + (\mathbf{B}^-_{sc} - \mathbf{n}(\mathbf{B}^-_{sc} \cdot \mathbf{n}))$$
$$\mathbf{E}^+_{sc} = \mathbf{n}(\mathbf{E}^-_{sc} \cdot \mathbf{n}) - (\mathbf{E}_{inc} - \mathbf{n}(\mathbf{E}_{inc} \cdot \mathbf{n}))$$

3.3 Summary of Formulation

For clarity, we include the basic scheme (using the $O(N^{2dim})$ method) for straight sided tetrahedra, noting that the geometric factors resulting from the mapping to the reference element are constant.

First we define the following matrix operators on each element e:

$$\mathbf{D}^r_{nm} = (\phi_n, \partial_r \phi_m), \quad \mathbf{D}^s_{nm} = (\phi_n, \partial_s \phi_m), \quad \mathbf{D}^t_{nm} = (\phi_n, \partial_t \phi_m)$$

$$\mathbf{M}_{in} = \phi_n|_{\text{i'th Gauss face node}}, \quad \tilde{\mathbf{M}}_{ni} = (\phi_n, h_i^g)_{\partial e}$$

where the last inner-product is taken over the face containing the i'th Gaussian face node: h_i^g and (r, s, t) are the coordinates relative to the reference physical element.

We then assume that we are at time step $n+1$ and that we have the polynomial coefficients for the state vector at time step n which we call $\hat{\mathbf{v}}^n$. We calculate the time derivative with the following algorithm:

1 For each element:
  Calculate $\mathbf{d}_r = \mathbf{D}_r \hat{\mathbf{v}}^n$ and similarily $\mathbf{d}_s$ and $\mathbf{d}_t$.
  Use the chain rule to calculate $\mathbf{d}_x, \mathbf{d}_y, \mathbf{d}_z$
  Calculate $\mathbf{A}\mathbf{d}_x + \mathbf{B}\mathbf{d}_y + \mathbf{C}\mathbf{d}_z$
2 For each element:
  Calculate $\hat{\mathbf{v}}^n = \mathbf{M}\hat{\mathbf{v}}^n$ to evaluate the state vector
  at the Gauss nodes on each faces of each elements.
3 For each element:
  For each face:
    For each Gauss node:
      Calculate $\mathbf{f} = \mathbf{F}(\mathbf{v}^+ - \mathbf{v}^-)$
4 For each element:
  Calculate $\tilde{\mathbf{M}}\mathbf{f}$
5 For each element:
  Add the contributions from steps 1 and 4

We use the result from the above algorithm to advance the polynomial coefficient state vector to time step $n+1$.

4 Scattering from a 2D cylinder

We consider scattering of a plane wave by a perfectly conducting cylinder ($\rho = \sigma = 0$). We split the state vector into an incident field and a scattered field. Since the evolution equations are linear in the state vector we only solve for the scattered field. The incident wave is incorporated into the boundary condition at the cylinder. The incident plane wave and boundary/initial conditions have the form:

$$H_x^{inc} = cos(3(t-y)), \quad H_y^{inc} = 0, \quad E_z^{inc} = cos(3(t-y))$$

$$H_x^{sc}(x, y, 0) = 0, \quad H_y^{sc}(x, y, 0) = 0, \quad E_z^{sc}(x, y, 0) = 0$$

with far-field boundary conditions:

$$H_x^{sc} = 0, \quad H_y^{sc} = 0, \quad E_z^{sc} = 0$$

We have run this simulation in a domain of approximately 16 diameters, with a radial absorbing boundary layer (ABC) at the fringe of the domain. Figure 1 shows the numerical solution at $t = 30$ using the DGM approach. This is visually indistinguishable from the exact solution [3]. However in the test of L_2 and L_∞ error (excluding the ABC region) we see that the error saturates at a level of 10^{-4}. This saturation is probably due to small reflections from the ABC. In future work, we intend to investigate how to modify this and alternative ABCs to reduce back scattering from far-field boundaries.

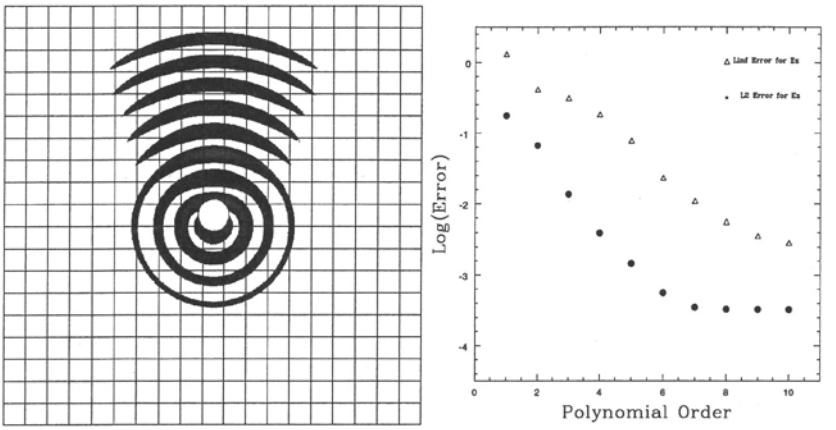


Fig. 1. Left: DGM solution with ABC and p=10 at t=30 for scattering of a plane wave incident on a perfectly conducting cylinder, Right:Convergence plot for L_2 and L_∞ error for the DGM approximation compared with the Mie solution at $t = 30$ as a function of polynomial order

5 Scattering from an F15 configuration

The previous example has been a formal way to test the accuracy of the methods outlined. We now consider a more substantial goal, which is simulation of electromagnetic scattering from an F15 airplane. The geometry presents intrinsic problems, since the wings and other features have sharp edges that cause singularities in the fields. The results shown in Figure 2 are an indication of work in progress, showing the calculated scattered magnetic fields (component aligned in the nose to tail direction) caused by an incident plane wave . As expected, the fields grow in an unbounded way at the sharp features on the body. These include the edges of the wings and three visible spots that are caused by slightly low resolution i.e. an uneven surface.

In this simulation we have used a fourth order expansion on a mesh of 123,000 tetrahedral elements (supplied by the "Grid Technology Thrust" project at Mississippi State University). The wall clock time per time step on 20 SP2-thin2 processors (computer time provided by G.Em. Karniadakis at Brown University) was 3.3 seconds. The computation achieved a peak Mflop count of 150 per processor, and averaged 95 Mflops (2 Gflop total).

In future work, we intend to compare the effectiveness of using *hp* adaption versus using filters at the sharp edge/features on the body in maintaining exponential convergence with increasing resolution.

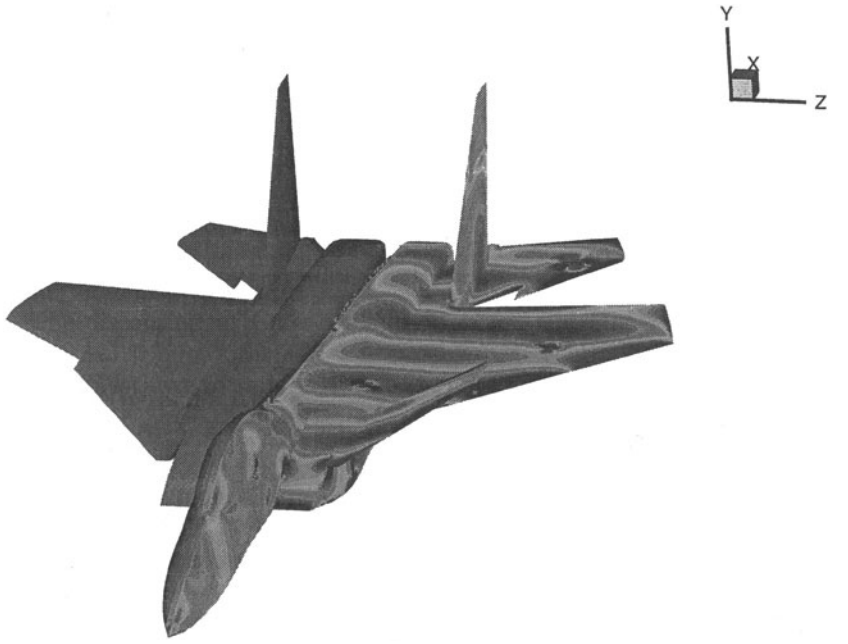


Fig. 2. Scattering from an F15, 123,000 tetrahedral elements, 4th order expansion. Surface and contours of the nose to tail component of scattered magnetic field.

References

1. Baumann C.E. and Oden J. T. The Discontinuous Galerkin Method Applied to CFD Problems. In *SIAM 45th Anniversary Meeting*, High order methods for compressible flow calculations, July 14-18 1997.

2. Bey K.S. and Oden J.T. *hp*-Version discontinuous Galerkin methods for hyperbolic conservation laws *Comp. Meth. Appl. Mech. Eng.*, 133:259-286, 1996
3. Bowman J.J., Senior T.B.A, and Uslenghi P.L.E. *Electromagnetic and Acoustic Scattering by Simple Shapes.* Hemisphere Publishing Corporation, 1987.
4. Cockburn B. and Shu C.-W. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws II : General framework . *Math. Comp.*, 52:411–435, 1989.
5. Cockburn B. and Shu C.-W. The Runge-Kutta discontinuous Galerkin method for conservation laws V: multidimensional systems. *J. of Comp. Phys.*, 141:199–224, 1998.
6. Dubiner M. Spectral methods on triangles and other domains. *J. Sci. Comp.*, 6:345, 1991.
7. Karniadakis G. E. and Sherwin S. J. *Spectral hpElement Methods for CFD.* Oxford University Press, 1999.
8. Koornwinder T. Two-variable analogues of the classical orthogonal polynomials. In Askey R.A., editor, *Theory and Applications of Special Functions.* Academic Press, 1975.
9. Kopriva D.A., Woodruff S.L. and Hussaini M.Y. *Discontinuous Spectral Element Approximation of Maxwell's Equation* International Symposium on Discontinous Galerkin Methods, Salve Regina University May 24-26, 1999
10. Lomtev I., Quillen C.B., and Karniadakis G. E. Spectral/hp methods for viscous compressible flows on unstructured 2d meshes. *J. Comp. Phys.*, in press, 1998.
11. Owens R.G. Spectral approximations on the triangle. *Proc. R. Sec. Lond. A*, 1997. Submitted.
12. Proriol J. Sur une famille de polynomes á deux variables orthogonaux dans un triangle. *C.R. Acad. Sci. Paris*, 257:2459–2461, 1957.
13. Reed W.H, and Hill T.R. Triangular mesh methods for the neutron transport equation Los Alamos Scientific Laboratory report LA-UR-73-479, Los Alamos, NM, 1973
14. Sherwin S.J. Hierarchical *hp* finite elements in hybrid domains. *Finite Elements in Analysis and Design*, 27:109–119, 1997.
15. Warburton T.C.E. *Spectral/hp Methods on Polymorphic Domains.* PhD thesis, Brown University, Division of Applied Mathematics, in progress.
16. Wingate B. A. and Taylor M. A. The natural function space for triangular and tetrahedral spectral elements. 1998. submitted to SIAM J. Num. Anal.
17. Yang B., Gottlieb D., and Hesthaven J.S. Spectral simulations of electromagnetic wave scattering. *J. Comput. Phys*, 1997.

A Space-Time Discontinuous Galerkin Method for Elastodynamic Analysis

Lin Yin[1,2,4], Amit Acharya[1,4], Nahil Sobh[1,2,4,5], Robert B. Haber[1,2,4,5], and Daniel A. Tortorelli[1,3,4,5]

[1] University of Illinois at Urbana-Champaign, Urbana IL 61801, USA
[2] Department of Theoretical & Applied Mechanics
[3] Department of Mechanical & Industrial Engineering
[4] Center for Simulation of Advanced Rockets
[5] Center for Process Simulation and Design

Abstract. We present a new space-time discontinuous Galerkin formulation for elastodynamics. The method allows for jumps in the field variables across inter-element boundaries with arbitrary orientation. The resulting method is locally conservative and admits a direct element-by-element solution procedure.

1 Introduction

1.1 Motivation and Previous Work

Hughes and Hulbert developed a time-discontinuous Galerkin procedure for second-order hyperbolic problems, such as elastodynamic analysis [2]. Johnson investigated the stability and convergence properties of the method [3]. The resulting method retains some of the attractive features of the fully-discontinuous methods for first-order problems, such as intrinsic stability and favorable convergence properties. However, other desirable properties, such as elementwise conservation and solution efficiency, are sacrificed. Wiberg and co-workers developed solution methods that mitigate the relative complexity of the time-discontinuous Galerkin method [6].

This paper presents a new space-time discontinuous Galerkin (DG) method for elastodynamic analysis, a problem that serves as a model for DG approximations of second-order hyperbolic problems in general. We seek a method that supports unstructured, fully-discontinuous space-time discretizations, that enforces element-wise conservation and that is compatible with direct element-by-element solution procedures. The key to the new formulation is the identification of the proper conditions across an arbitrary surface of discontinuity embedded in a space-time domain. We employ the notation of differential forms and the exterior calculus to streamline the presentation.

Although similar in some respects, the method presented here is distinct from the method introduced by Richter for the scalar wave equation [5]. The element-by-element solution strategy is similar to the one presented by Lowrie, Roe and van Leer [4] for hyperbolic conservation laws.

2 Formulation

2.1 Domain of Influence and Mesh Constraints

Consider a space-time domain, $\mathcal{D} \subset E^{d+1}$, in which d is the spatial dimension. The boundary of $\mathcal{D}$ is comprised of disjoint parts $\bar{\Gamma}_{\mathrm{D}}$, $\bar{\Gamma}_{\mathrm{N}}$, $\bar{\Gamma}_0$ and $\bar{\Gamma}^+$. These are the space-time trajectories of the Dirichlet boundary, the Neumann boundary, the initial time boundary and the terminal time boundary.

Our element-by-element solution procedure depends on the notion of *domains of influence* for dynamic data [1]. Let $\mathcal{E}_{\mathbf{z}}$ be some elastodynamic event that occurs at $\mathbf{z} \in \mathcal{D}$. The domain of influence of $\mathcal{E}_{\mathbf{z}}$, denoted $\mathcal{I}_{\mathbf{z}}$, is the interior of a cone with vertex $\mathbf{z}$ and whose boundaries are determined by the fastest wave speed in the material. The response outside $\mathcal{I}_{\mathbf{z}}$ is independent of $\mathcal{E}_{\mathbf{z}}$.

Let $\mathcal{Q} \subset \mathcal{D}$ be open, $\Gamma_{\mathrm{N}} = \partial\mathcal{Q} \cap \bar{\Gamma}_{\mathrm{N}}$, $\Gamma_{\mathrm{D}} = \partial\mathcal{Q} \cap \bar{\Gamma}_{\mathrm{D}}$, $\Gamma_0 = \partial\mathcal{Q} \cap \bar{\Gamma}_0$, $\Gamma_* = \Gamma_{\mathrm{N}} \cup \Gamma_{\mathrm{D}} \cup \Gamma_0$, and $\mathcal{B}(\mathbf{z}, r)$ be an open ball of radius r at $\mathbf{z}$. We partition $\partial\mathcal{Q}\backslash\Gamma_*$ into three disjoint parts: the local outflow boundary $\Gamma^+ := \{\mathbf{z} \in \partial\mathcal{Q}\backslash\Gamma_* : \mathcal{I}_{\mathbf{z}} \cap \mathcal{Q} = \emptyset\}$, the local inflow boundary $\Gamma^- := \{\mathbf{z} \in \partial\mathcal{Q}\backslash\Gamma_* : \lim_{r\to 0} (\mathcal{I}_{\mathbf{z}} \cap \mathcal{B}(\mathbf{z}, r)) \subset \mathcal{Q}\}$ and the remaining mixed boundary. Linear momentum flows in both directions across a mixed boundary; so adjacent elements that share a mixed boundary must be solved simultaneously. However, if a finite element mesh contains no mixed boundaries, then an element-by-element solution can be achieved.

Let ω be an s-form, $0 \leq s \leq d$, that is continuous on the interior of $\mathcal{Q}$, but that admits jumps across $\partial\mathcal{Q}$. In anticipation of the construction of various jump integrals, we define the s-form ω^- on Γ^- by

$$\omega^-(\mathbf{z}) := \lim_{\varepsilon\to 0+} \omega(\mathbf{z} + \varepsilon\mathbf{n}(\mathbf{z})) \;\; \forall\, \mathbf{z} \in \Gamma^-; \tag{1}$$

in which $\mathbf{n}$ is an outward normal vector with respect to $\mathcal{Q}$ on Γ^-.

2.2 The Elastodynamic Problem

The governing equations and boundary conditions for the linearized elastodynamic problem are

$$\mathrm{d}\mathbf{M} + \mathbf{b} = \mathbf{0} \text{ on } \mathcal{D} \tag{2}$$

$$(\mathbf{M} - \mathbf{g}) = \mathbf{0} \text{ on } \bar{\Gamma}_{\mathrm{N}} \tag{3}$$

$$\mathbf{u} - \mathbf{h} = \mathbf{0} \text{ on } \bar{\Gamma}_{\mathrm{D}} \tag{4}$$

$$\mathbf{u} - \mathbf{u}_0 = \mathbf{0} \text{ on } \Gamma_0 \tag{5}$$

$$\dot{\mathbf{u}} - \dot{\mathbf{u}}_0 = \mathbf{0} \text{ on } \Gamma_0, \tag{6}$$

in which $\mathbf{u}$ is the 0-form on $\mathcal{D}$ with vector coefficients determined by the displacement field, and $\mathbf{M} = \breve{\mathbf{M}}(\mathbf{du})$ is the d-form on $\mathcal{D}$ with vector coefficients that corresponds to the stress-momentum tensor [1]. The body-force vector $\mathbf{b}$

is a $(d+1)$-form on $\mathcal{D}$ with vector coefficients. The notation $\mathrm{d}(\cdot)$ denotes the exterior derivative on $\mathcal{D}$. The Neumann condition (3) involves the prescribed flux of linear momentum $\mathbf{g}$, a d-form with vector coefficients on $\bar{\Gamma}_{\mathrm{N}}$ that can include both traction and inertia components. The prescribed vector fields $\mathbf{h}$ on $\bar{\Gamma}_{\mathrm{D}}$ and $(\mathbf{u}_0, \dot{\mathbf{u}}_0)$ on Γ_0 specify the Dirichlet and initial conditions, respectively. For simplicity, we assume that the body force is non-singular on $\mathcal{D}$ (that is, we exclude impulse loads).

2.3 Localized Problem

We introduce a partition of $\mathcal{D}$ (corresponding to a finite element mesh) for which every subdomain $\mathcal{Q}$ is free of mixed-type boundaries. It is then possible to solve the problem one subdomain at a time, starting from the initial inflow boundary. We construct a broken solution space that is continuous and smooth within each subdomain, but that admits jumps across subdomain boundaries. Then the local problem on a typical subdomain $\mathcal{Q}$ is

$$\mathrm{d}\mathbf{M} + \mathbf{b} = \mathbf{0} \text{ on } \mathcal{Q} \tag{7}$$

$$[\mathbf{M}] = \mathbf{0} \text{ on } \Gamma^- \cup \Gamma_{\mathrm{N}} \cup \Gamma_0 \tag{8}$$

$$\mathrm{d}_\Gamma [\mathbf{u}] = \mathbf{0} \text{ on } \Gamma^- \cup \Gamma_{\mathrm{D}} \cup \Gamma_0 \tag{9}$$

$$\int_{\Gamma^- \cup \Gamma_{\mathrm{D}} \cup \Gamma_0} [\mathbf{u}] = \mathbf{0} \tag{10}$$

where $[\mathbf{M}]|_{\Gamma^-} := \mathbf{M} - \mathbf{M}^-$, $[\mathbf{M}]|_{\Gamma_{\mathrm{N}}} := \mathbf{M} - \mathbf{g}$, $[\mathbf{M}]|_{\Gamma_0} := \mathbf{M} - \rho \dot{\mathbf{u}}_0 \mathbf{e}^s$, $[\mathbf{u}]|_{\Gamma^-} := \mathbf{u} - \mathbf{u}^-$ $[\mathbf{u}]|_{\Gamma_{\mathrm{D}}} := \mathbf{u} - \mathbf{h}$, and $[\mathbf{u}]|_{\Gamma_0} := \mathbf{u} - \mathbf{u}_0$. Here ρ is the mass density, and $\mathbf{e}^s$ is the "spatial" d-covector given by $\mathbf{e}^1 \wedge \ldots \wedge \mathbf{e}^d$. The exterior derivative d_Γ on the d-manifold Γ has the component representation

$$\mathrm{d}_\Gamma \mathbf{u} := \frac{\partial u^i}{\partial \xi^\alpha} \mathbf{e}_i \otimes \mathbf{g}^\alpha \tag{11}$$

in which $\mathbf{e}_i$ are Cartesian basis vectors in E^{d+1}, $\xi^\alpha; \alpha = 1, d$ are parametric coordinates on Γ, and $\mathbf{g}^\alpha$ are 1-forms on Γ corresponding to the ξ^α directions.

Equation (7) represents balance of linear momentum on $\mathcal{Q}$, while equation (8) balances the flux of linear momentum across $\Gamma^- \cup \Gamma_{\mathrm{N}} \cup \Gamma_0$. Note that the latter condition includes the local initial velocity condition (6) when $\Gamma_0 \neq \emptyset$. Equation (9) is the Maxwell condition for displacement compatibility across the local internal, Dirichlet and initial boundaries. When combined with an appropriate auxilliary constraint, such as (10), the Maxwell condition guarantees absolute compatibility of displacements across $\Gamma^- \cup \Gamma_{\mathrm{D}} \cup \Gamma_0$.

2.4 Weak Form of the Localized Problem

Let $\mathcal{V}$ be a space of suitably smooth vector-valued functions on $\mathcal{Q}$ that satisfy $\int_{\Gamma^- \cup \Gamma_{\mathrm{D}} \cup \Gamma_0} [\mathbf{u}] = \mathbf{0}$, and let $\tilde{\mathbf{M}}$ be a $(d-1)$-form constructed from the

components of $\breve{\mathbf{M}}(\mathrm{d}\mathbf{w})$ (see example below). For $\mathbf{w} \in \mathcal{V} : \mathbf{w} = w^i \mathbf{e}_i;\ i = 1, d$, we define $\dot{\mathbf{w}}$ as the vector-valued function on $\mathcal{Q}$ with Cartesian components $\dot{w}^i = \partial w^i / \partial t$. Then the weighted residual statement of the local problem is,

Find $\mathbf{u} \in \mathcal{V} \ni$

$$-\int_{\mathcal{Q}} \dot{\mathbf{w}} \wedge (\mathrm{d}\mathbf{M} + \mathbf{b}) + \int_{\Gamma^- \cup \Gamma_\mathrm{N} \cup \Gamma_0} \dot{\mathbf{w}} \wedge [\mathbf{M}] + \int_{\Gamma^- \cup \Gamma_\mathrm{D} \cup \Gamma_0} \tilde{\mathbf{M}} \wedge \mathrm{d}_\Gamma [\mathbf{u}] = \mathbf{0}\ \forall\ \mathbf{w} \in \mathcal{V}. \tag{12}$$

An application of Stokes Theorem yields the weak form of the local problem:

Find $\mathbf{u} \in \mathcal{V} \ni$

$$\int_{\mathcal{Q}} (\mathrm{d}\dot{\mathbf{w}} \wedge \mathbf{M} - \dot{\mathbf{w}} \wedge \mathbf{b}) - \int_{\Gamma^-} \dot{\mathbf{w}} \wedge \mathbf{M}^- - \int_{\Gamma_\mathrm{N}} \dot{\mathbf{w}} \wedge \mathbf{g} - \int_{\Gamma_0} \dot{\mathbf{w}} \wedge \rho \dot{\mathbf{u}}_0 \mathbf{e}^s - \int_{\Gamma_\mathrm{D} \cup \Gamma^+} \dot{\mathbf{w}} \wedge \mathbf{M} + \int_{\Gamma^- \cup \Gamma_\mathrm{D} \cup \Gamma_0} \tilde{\mathbf{M}} \wedge \mathrm{d}_\Gamma [\mathbf{u}] = \mathbf{0}\ \forall\ \mathbf{w} \in \mathcal{V}. \tag{13}$$

The restriction of $\mathbf{u}$ and $\mathbf{w}$ to a common finite-dimensional subspace of $\mathcal{V}$ yields the local Galerkin finite element approximation.

2.5 Element-wise Conservation Properties

Finite element methods derived from (13) conserve linear momentum locally over each element $\mathcal{Q}$. To illustrate this point, consider the special case in which $d = 1$ and Young's modulus, denoted E, is uniform. Let $n^1 = \partial t / \partial \xi^1$ and $n^t = -\partial x_1 / \partial \xi^1$, where ξ^1 is an arc-length parameter along Γ. Then $\mathbf{u} = u^1 \mathbf{e}_1$, $\mathbf{w} = w^1 \mathbf{e}_1$, $\mathbf{n} = n^1 \mathbf{e}_1 + n^t \mathbf{e}_t$, $\mathbf{M} = \mathbf{e}_1 \left\{ (\rho \partial w_1 / \partial t)\, \mathbf{e}^1 + (E \partial w_1 / \partial x_1)\, \mathbf{e}^t \right\}$, and we define $\tilde{\mathbf{M}} = (\rho \partial w_1 / \partial t + E \partial w_1 / \partial x_1) \mathbf{e}_1$. We also have

$$\mathrm{d}_\Gamma [\mathbf{u}] = \left(\frac{\partial [u^1]}{\partial t} n^1 - \frac{\partial [u^1]}{\partial x_1} n^t \right) \left(-n^t \mathbf{e}^1 + n^x \mathbf{e}^t \right). \tag{14}$$

Thus,

$$\tilde{\mathbf{M}} \wedge \mathrm{d}_\Gamma [\mathbf{u}] = \left(\rho \frac{\partial w_1}{\partial t} + E \frac{\partial w_1}{\partial x_1} \right) \left(\frac{\partial [u^1]}{\partial t} n^1 - \frac{\partial [u^1]}{\partial x_1} n^t \right) \mathbf{e}_1 \left(-n^t \mathbf{e}^1 + n^x \mathbf{e}^t \right). \tag{15}$$

Now let $w^1 = c x_1 + t + \bar{c}$. Then, $\partial w^1 / \partial x_1 = c$, $\partial w^1 / \partial t = 1$, and $\mathrm{d}\dot{\mathbf{w}} = \mathbf{0}$. We can choose c and $\bar{c}$ such that $\int_{\Gamma^- \cup \Gamma_\mathrm{D} \cup \Gamma_0} [\mathbf{w}] = \mathbf{0}$ (ensures $\mathbf{w} \in \mathcal{V}$) and $\int_{\Gamma^- \cup \Gamma_\mathrm{D} \cup \Gamma_0} \tilde{\mathbf{M}} \wedge \mathrm{d}_\Gamma [\mathbf{u}] = \mathbf{0}$. Then the weak form (13) yields conservation of linear momentum on element $\mathcal{Q}$:

$$\int_{\mathcal{Q}} \mathbf{b} + \int_{\Gamma^-} \mathbf{M}^- + \int_{\Gamma_\mathrm{N}} \mathbf{g} + \int_{\Gamma_0} \rho \dot{\mathbf{u}}_0 \mathbf{e}^s + \int_{\Gamma_\mathrm{D} \cup \Gamma^+} \mathbf{M} = \mathbf{0}. \tag{16}$$

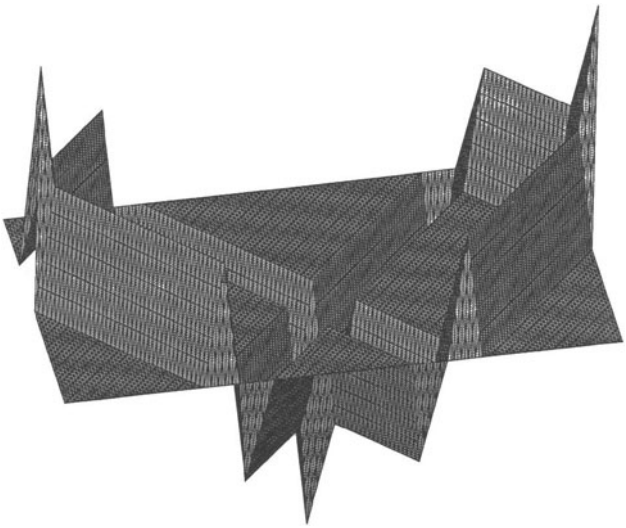


Fig. 1. The discrete and exact solutions match on a characteristic grid. Here u defines the height field; the coordinate origin is at the upper left corner.

$\frac{1}{h}$	$\|e\|_{L_2}$		$\|e_{,x}\|_{L_2}$		$\|e_{,t}\|_{L_2}$	
	Present	Ref. [5]	Present	Ref. [5]	Present	Ref. [5]
20	2.66×10^{-2}	3.30×10^{-2}	1.62	1.76	0.163	0.198
40	1.50×10^{-2}	1.48×10^{-2}	1.45	1.45	8.61×10^{-2}	0.102
80	7.17×10^{-3}	6.92×10^{-3}	1.11	1.10	5.41×10^{-2}	7.65×10^{-2}

Table 1. Convergence study of new formulation and method in Ref. [5].

3 Numerical Example

Consider the problem $-u_{,xx} + \ddot{u} = 0$ on the domain, $\mathcal{D} = \{(x,t) : 0 < x < 1; 0 < t < 2\}$. We enforce $u(0,\cdot) = u(1,\cdot) = 0$, and the initial conditions, $\dot{u}(x,\cdot) = 0$ and $u(x,\cdot) = \{0,\, 8\,(x - 0.25)\,,\, -8\,(x - 0.5)\,,\, 0\}$ for $\{0 < x < 0.25,\ 0.25 < x < 0.375,\ 0.375 < x < 0.5,\ 0.5 < x < 1.0\}$. We solve the problem on a mesh of quadratic 6-node triangular finite elements, using both the formulation presented in Ref. [5] and the new method described here. The element diameter in the x-direction is denoted h. Both methods yield the exact solution to within machine precision when applied to a mesh aligned with the characteristic directions (i.e., when the diagonal mesh angle is $\theta = \pm 45°$ — see Fig. 1). Both methods are approximate when the mesh is not characteristic. Table 1 and Fig. 2 present a convergence study of the solution error e and its derivatives at $t = 2.0$ for a mesh with diagonal angle $\theta = \pm 41°$.

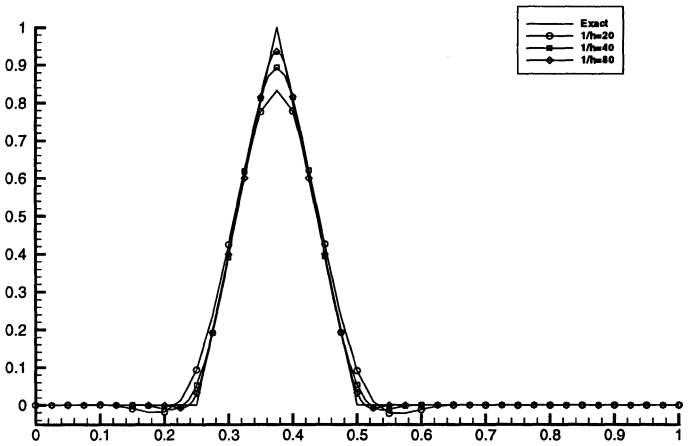


Fig. 2. Convergence study of solution on grid with $\theta = \pm 41°$ at $t = 2.0$ for $1/h = 20$, 40 and 80.

4 Acknowledgments

The work reported here was supported, in part, by the Center for Process Simulation and Design (NSF DMS 98-73945) and the Center for Simulation of Advanced Rockets (DOE LLNL B341494). The authors wish to thank E. Fried for his advice and suggestions concerning the space-time formulation.

References

1. Gurtin, M. E.: *The Linear Theory of Elasticity*, Handbuch der Physik $6a_2$, Springer-Verlag, Berlin (1972).
2. Hughes, T. J. R. and Hulbert, G. M.: Space-time finite element methods for elastodynamics: formulations and error estimates, Comput. Methods Appl. Mech. Engrg. 66 (1988) 339 – 363.
3. Johnson, C.: Discontinuous Galerkin finite element methods for second order hyperbolic problems, Comput. Methods Appl. Mech. Engrg. 107 (1993) 145-157.
4. Lowrie, R. B., Roe, P. L. and van Leer, B.: Space-time methods for hyperbolic conservation laws, In *Barriers and Challenges in Computational Fluid Dynamics*, ICASE/LaRC Interdisciplinary Series in Science and Engineering, Vol. 6, Kluwer, (1998) 79-98.
5. Richter, G. R.: An explicit finite element method for the wave equation, Appl. Num. Math. 16 (1994) 65-80.
6. Li, X. D. and Wiberg, N. -E.: Implementation and adaptivity of a space-time finite element method for structural dynamics, Comput. Methods Appl. Mech. Engrg. 156 (1998) 211-229.

Nonconforming, Enhanced Strain, and Mixed Finite Element Methods – A Unified Approach

Zhimin Zhang *

Department of Mathematics and Statistics
Texas Tech University, Lubbock, TX 79409

Abstract. Both nonconforming and enhanced strain methods are analyzed under the framework of the mixed method. The notion of selective nonconforming or selective enhanced strain methods are introduced.

1. Introduction. Let $\Omega \subset R^2$ be a polygonal domain. Consider on Ω, the plain strain problem in which we seek for the unknown displacement $\boldsymbol{u} \in V$ such that,

$$(P_\lambda) \quad B_\lambda(\boldsymbol{u}, \boldsymbol{v}) = 2\mu(\boldsymbol{\epsilon}(\boldsymbol{u}), \boldsymbol{\epsilon}(\boldsymbol{v})) + \lambda(\operatorname{div} \boldsymbol{u}, \operatorname{div} \boldsymbol{v}) = f(\boldsymbol{v}), \quad \forall \boldsymbol{v} \in V,$$

with

$$\epsilon_{ij}(\boldsymbol{v}) = \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right), \quad f(\boldsymbol{v}) = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v} dx_1 dx_2 + \int_{\partial\Omega} \boldsymbol{g} \cdot \boldsymbol{v} ds.$$

Here $\boldsymbol{\epsilon} = (\epsilon_{ij})$ is the strain tensor, μ and λ are the Lamé parameters given by

$$\mu = \frac{E}{2(1+\nu)}, \quad \lambda = \frac{E\nu}{(1+\nu)(1-2\nu)},$$

where $\nu \in [0, 1/2)$ is the Poisson ratio and E is the Young's modulus. The space V is defined according to different boundary conditions, for example, $V = H_0^1(\Omega) \times H_0^1(\Omega)$ for the pure displacement problem. See [2] for the pure traction problem.

Under certain regularity condition for $\boldsymbol{f}$, $\boldsymbol{g}$, and Ω, (P_λ) has a unique solution for any $\nu \in [0, 1/2)$. However, the traditional finite element approximation of (P_λ) fails to converge when $\nu \to 1/2$ which is called Poisson locking. We shall discuss in this article some special finite element methods to treat the Poisson locking. In order to avoid technical complexity, our discussion will be focussed on meshes that can be obtained by affine mappings from a reference element, in which case the Jacobi is a constant. The reader is referred to [8] for general quadrilateral meshes.

2. Different Approaches. We introduce several related methods.

* This work was partially supported by the National Science Foundation Grants DMS-9626193, DMS-9622690, and INT-9605050.

The mixed method. If $p = -\lambda \operatorname{div} \boldsymbol{v}$ is taken as an independent unknown, we then have the following mixed variational formulation in which we are looking for $(\boldsymbol{u}, p) \in V \times W$ such that,

$$(M_\lambda) \qquad \begin{aligned} 2\mu(\epsilon(\boldsymbol{u}), \epsilon(\boldsymbol{v})) - (\operatorname{div} \boldsymbol{v}, p) &= f(\boldsymbol{v}), \quad \forall \boldsymbol{v} \in V, \\ \lambda^{-1}(p, q) + (\operatorname{div} \boldsymbol{u}, q) &= 0, \quad \forall q \in W, \end{aligned}$$

where $W = L_0^2(\Omega)$ is the subspace of $L^2(\Omega)$ consisting of functions with zero mean value. In order for (M_λ) to have a unique solution for any $\lambda \in (0, \infty]$, a stability requirement, the inf-sup condition, must be satisfied in addition to some regularity requirement on Ω and $\boldsymbol{f}$. Special care must be taken to design finite element spaces which satisfy a discrete inf-sup condition to approximate V and W. A major drawback is that the problem changes from one of minimization to the discovery of a saddle-point, and consequently, the resulting discrete system is much larger than the original one and changes from symmetric positive definite to non-symmetric indefinite. The reader is referred to [1,3] for the literature.

The nonconforming method. This method uses the displacement formulation (P_λ). The type of methods we shall discuss in this work are constructed by amending the commonly used finite elements with some "bubbles" in each element. To be more specific, the finite element space can be decomposed into $V^h = S^h + B_h$, where S^h is the traditional conforming finite element space and B_h is the nonconforming part that is not necessarily continuous across elements. We are most interested in lower-order finite elements for S^h, for example, the bilinear element. We begin with the following variational formulation: Find $\boldsymbol{u}_h^c + \boldsymbol{u}_h^b \in S^h + B_h = V^h$ such that

$$(\mathrm{NC}) \qquad B_\lambda^h(\boldsymbol{u}_h^c + \boldsymbol{u}_h^b, \boldsymbol{v}^c + \boldsymbol{v}^b) = f(\boldsymbol{v}^c), \quad \forall \boldsymbol{v}^c + \boldsymbol{v}^b \in V^h.$$

In B_λ^h, ϵ_h and div_h are used in place of ϵ and div, respectively, to indicate that differentiation is performed element-wise, since the bubble functions may not be continuous across the element boundaries.

The variational formulation (NC) can be de-coupled to (by letting $\boldsymbol{v}^b = 0$ and $\boldsymbol{v}^c = 0$, respectively)

$$B_\lambda^h(\boldsymbol{u}_h^c + \boldsymbol{u}_h^b, \boldsymbol{v}^c) = f(\boldsymbol{v}^c), \quad B_\lambda^h(\boldsymbol{u}_h^c + \boldsymbol{u}_h^b, \boldsymbol{v}^b) = 0.$$

Note that the second equation is valid element-wise. Therefore, the bubble function $\boldsymbol{u}_h^b$ can be solved in terms of $\boldsymbol{u}_h^c$ on each element and substituted into the first equation. The process results in a discrete system with the only unknown $\boldsymbol{u}_h^c$. This procedure is called "static condensation" in the engineering community. By properly choosing B_h, the bubble functions $\boldsymbol{u}^b$ and $\boldsymbol{v}^b$ will "stabilize" the discrete system and thereby overcome the locking. The overall computational cost is compatible with the counterpart conforming method (without B_h).

The enhanced strain method. In contrast to the long history of the mixed and nonconforming methods, the enhanced strain method appeared in

1990s, see [6]. Define

$$\boldsymbol{\Gamma}^h = \boldsymbol{\epsilon}_h(V^h) = \{\boldsymbol{\gamma} = \boldsymbol{\epsilon}_h(\boldsymbol{v}) \text{ for some } \boldsymbol{v} \in V^h\},$$

the method looks for $(\boldsymbol{u}_h^c, \tilde{\boldsymbol{\epsilon}}) \in S^h \times \boldsymbol{\Gamma}^h$, such that

$$(D[\boldsymbol{\epsilon}(\boldsymbol{u}_h^c) + \tilde{\boldsymbol{\epsilon}}], [\boldsymbol{\epsilon}(\boldsymbol{v}^c) + \boldsymbol{\gamma}]) = f(\boldsymbol{v}^c), \quad \forall (\boldsymbol{v}^c, \boldsymbol{\gamma}) \in S^h \times \boldsymbol{\Gamma}^h, \tag{2.1}$$

where C is the 4th-order tensor of elastic moduli. For the isotropic material, we can write D as a matrix and $\boldsymbol{\epsilon}$ as a vector

$$D = \begin{pmatrix} \lambda+2\mu & \lambda & 0 \\ \lambda & \lambda+2\mu & 0 \\ 0 & 0 & \mu \end{pmatrix} = \lambda \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \mu \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\boldsymbol{\epsilon} = (\epsilon_{11}, \epsilon_{22}, 2\epsilon_{12})^T.$$

A straight forward calculation reveals that (2.1) is equivalent to the following variational formulation

$$\text{(ES)} \qquad 2\mu(\boldsymbol{\epsilon}(\boldsymbol{u}_h^c) + \tilde{\boldsymbol{\epsilon}}, \boldsymbol{\epsilon}(\boldsymbol{v}^c) + \boldsymbol{\gamma}) + \lambda(\operatorname{div} \boldsymbol{u}_h^c + \operatorname{tr} \tilde{\boldsymbol{\epsilon}}, \operatorname{div} \boldsymbol{v}^c + \operatorname{tr} \boldsymbol{\gamma}) = f(\boldsymbol{v}^c),$$

where "tr" is the trace operator with $\operatorname{tr} \boldsymbol{\gamma} = \gamma_{11} + \gamma_{22}$.

Equivalence of nonconforming and enhanced strain methods. The variational formulation (NC) can be written as

$$\begin{aligned} &2\mu(\boldsymbol{\epsilon}(\boldsymbol{u}_h^c) + \boldsymbol{\epsilon}_h(\boldsymbol{u}_h^b), \boldsymbol{\epsilon}(\boldsymbol{v}^c) + \boldsymbol{\epsilon}_h(\boldsymbol{v}^b)) \\ &+\lambda(\operatorname{div} \boldsymbol{u}_h^c + \operatorname{div}_h \boldsymbol{u}_h^b \operatorname{div} \boldsymbol{v}^c + \operatorname{div}_h \boldsymbol{v}^b) = f(\boldsymbol{v}^c), \end{aligned}$$

Comparing with (ES), we see that they are the same with $\tilde{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_h(\boldsymbol{u}_h^b)$ and $\boldsymbol{\gamma} = \boldsymbol{\epsilon}_h(\boldsymbol{v}^b)$ by the definition of $\boldsymbol{\Gamma}^h$. Hence, from now on, we shall concentrate on the nonconforming method.

The selective nonconforming (enhanced strain) method. The idea comes from selectively reduced integration method for the plate bending problem [4, p.327]. In the case of the plain elasticity equation, we only need to amend the bubbles (or enhanced strain) to the divergence term which will cause trouble when $\lambda \to \infty$ ($\nu \to 1/2$). We propose two strategies to modify (NC).

$$\text{(SN1)} \quad 2\mu(\boldsymbol{\epsilon}(\boldsymbol{u}_h^c), \boldsymbol{\epsilon}(\boldsymbol{v}^c)) + 2\mu(\boldsymbol{\epsilon}_h(\boldsymbol{u}_h^b), \boldsymbol{\epsilon}_h(\boldsymbol{v}^b)) + \lambda(\operatorname{div}_h \boldsymbol{u}_h, \operatorname{div}_h \boldsymbol{v}) = f(\boldsymbol{v}^c);$$

$$\text{(SN2)} \qquad 2\mu(\boldsymbol{\epsilon}(\boldsymbol{u}_h^c), \boldsymbol{\epsilon}(\boldsymbol{v}^c)) + \lambda(\operatorname{div}_h \boldsymbol{u}_h, \operatorname{div}_h \boldsymbol{v}) = f(\boldsymbol{v}^c).$$

Note that $\boldsymbol{u}_h = \boldsymbol{u}_h^c + \boldsymbol{u}_h^b$ and $\boldsymbol{v} = \boldsymbol{v}^c + \boldsymbol{v}^b$.

Equivalence of nonconforming and mixed methods. Introducing $p_h = -\lambda(\operatorname{div} \boldsymbol{u}_h^c + \operatorname{div}_h \boldsymbol{u}_h^b)$, (SN2) can be cast into the following mixed formulation: Find $(\boldsymbol{u}_h^c + \boldsymbol{u}_h^b; p_h) \in (S^h + B_h) \times W_h$ such that

$$2\mu(\boldsymbol{\epsilon}(\boldsymbol{u}_h^c), \boldsymbol{\epsilon}(\boldsymbol{v}^c)) - (\operatorname{div} \boldsymbol{v}^c + \operatorname{div}_h \boldsymbol{v}^b, p_h) = f(\boldsymbol{v}^c), \quad \forall \boldsymbol{v} \in V^h, \tag{2.2}$$

$$\text{(MN2)} \qquad \lambda^{-1}(p_h, q) + (\operatorname{div} \boldsymbol{u}_h^c + \operatorname{div}_h \boldsymbol{u}_h^b, q) = 0, \quad \forall q \in W_h, \tag{2.3}$$

with $W_h \supset \operatorname{div}_h V^h$. It is straightforward to show that under the affine mapping of the mesh (which we assume consists of triangles and/or parallelograms), (SN2) and (MN2) are equivalent. Note that $(\operatorname{div}_h \boldsymbol{v}^b, p_h) = 0$ (by letting $\boldsymbol{v}^c = 0$ in (2.2)).

Summary. Under the affine mapping of the mesh, both the nonconforming method and the enhanced strain method are equivalent to some mixed methods. Therefore, we can adopt the well established theoretical framework for mixed methods to analyze the nonconforming method and the enhanced strain method while still enjoying their simple displacement variational formulation in numerical implementation.

3. Some Theoretical Issues. In order for a mixed method to work, certain conditions must be satisfied. Most important of all are a discrete Korn's inequality and the inf-sup condition. When nonconforming or enhanced strain methods are involved, we consider, in addition, the consistent error.

To fix the idea, let S^h be the bilinear element space on $\mathcal{T}_h$, a partition of Ω, and let $\boldsymbol{v}^b \in B_h$ be defined such that on the reference plane, $\hat{\boldsymbol{v}}^b$ equals

$$\begin{pmatrix} \lambda_1^K(1-\xi^2) \\ \lambda_2^K(1-\eta^2) \end{pmatrix}, \quad \text{or} \quad \begin{pmatrix} \lambda_1^K(1-(\xi^2+\eta^2)/2) \\ \lambda_2^K(1-(\xi^2+\eta^2)/2) \end{pmatrix}.$$

We define $\|\boldsymbol{v}^b\|_h^2 = \sum_{K\in\mathcal{T}_h} |\boldsymbol{v}^b|_{1,K}^2$. It is straightforward to verify that $\|\cdot\|_h$ is a norm on V^h. We also define

$$Z_h = \{\boldsymbol{v} \in V^h \,|\, (\operatorname{div}_h \boldsymbol{v}, q) = \lambda^{-1}(p_h, q),\ \forall q \in W_h\}$$

and Π_0, the L^2 projection from $L^2(\Omega)$ onto the space of piecewise constant

$$\bar{W}_h = \{\bar{q} \in W_h,\ \bar{q}|_K \in P_0(K)\ \ \forall K \in \mathcal{T}_h\}.$$

Certain properties should be satisfied by $\boldsymbol{v}^b \in B_h$.

$$(\operatorname{div}_h \boldsymbol{v}^b, \bar{q}) = 0, \quad \forall \bar{q} \in \bar{W}_h; \tag{3.1}$$

$$c\|\operatorname{div} \boldsymbol{v}^b\|_{0,K} \geq |\boldsymbol{v}^b|_{1,K}, \quad \forall \boldsymbol{v}^b \in B_h, \quad K \in \mathcal{T}_h; \tag{3.2}$$

and for any $q \in W_h$ and $\boldsymbol{v}^c \in S^h$, there exists $\boldsymbol{v}^b \in B_h$ such that

$$(\operatorname{div}_h(\boldsymbol{v}^c + \boldsymbol{v}^b), (I-\Pi_0)q) = \|(I-\Pi_0)q\|_0^2, \tag{3.3}$$

$$\|\boldsymbol{v}^b\|_h^2 \leq C(\|(I-\Pi_0)q\|_0^2 + |\boldsymbol{v}^c|_1^2). \tag{3.4}$$

Here c and C are fixed positive constants. We can show that our two bubble functions satisfy all properties (3.1)–(3.4) with $c = 1$ and $C = 2$. Property (3.1) is obvious. The proof of (3.2) comes from a straight forward calculation. Details can be found in [8] for the proof of (3.3) and (3.4).

Coercivity: the discrete Korn's inequality. Under the condition (3.2), (SN2) and (MN2) satisfy the discrete Korn's second inequality, ie., there exists a constant $\alpha > 0$ independent of h, such that

$$|(\boldsymbol{\epsilon}(\boldsymbol{v}^c), \boldsymbol{\epsilon}(\boldsymbol{v}^c))| \geq \alpha(|\boldsymbol{v}^c|_1^2 + \|\boldsymbol{v}^b\|_h^2), \quad \forall \boldsymbol{v}^c + \boldsymbol{v}^b \in Z_h. \tag{3.5}$$

Proof: Given $\boldsymbol{v} = \boldsymbol{v}^c + \boldsymbol{v}^b \in Z_h$, we have, $(\mathrm{div}_h(\boldsymbol{v}^c + \boldsymbol{v}^b), q) = -\lambda^{-1}(p_h, q)$ for all $q \in W_h$. Set $q = \mathrm{div}_h \boldsymbol{v}^b$, recall $(p_h, \mathrm{div}_h \boldsymbol{v}^b) = 0$, and we have

$$\|\mathrm{div}_h \boldsymbol{v}^b\|^2 = -(\mathrm{div}\, \boldsymbol{v}^c, \mathrm{div}_h \boldsymbol{v}^b) \leq \|\mathrm{div}\, \boldsymbol{v}^c\| \|\mathrm{div}_h \boldsymbol{v}^b\|.$$

Therefore, $\|\mathrm{div}_h \boldsymbol{v}^b\| \leq \|\mathrm{div}\, \boldsymbol{v}^c\| \leq \sqrt{2}|\boldsymbol{v}^c|_1$. Realizing that (3.2) implies

$$c\|\mathrm{div}_h \boldsymbol{v}^b\| \geq \|\boldsymbol{v}^b\|_h,$$

we have, $c\sqrt{2}|\boldsymbol{v}^c|_1 \geq \|\boldsymbol{v}^b\|_h$. On the other hand, the conforming part $\boldsymbol{v}^c$ satisfy the second Korn's inequality $|(\boldsymbol{\epsilon}(\boldsymbol{v}^c), \boldsymbol{\epsilon}(\boldsymbol{v}^c))| \geq \gamma |\boldsymbol{v}^c|_1^2$. After some simple algebraic manipulation, we obtain

$$|(\boldsymbol{\epsilon}(\boldsymbol{v}^c), \boldsymbol{\epsilon}(\boldsymbol{v}^c))| \geq \frac{\gamma}{1 + 2c^2}(|\boldsymbol{v}^c|_1^2 + \|\boldsymbol{v}^b\|_h^2).$$

Hence, the assertion is proved with $\alpha = \gamma/(1 + 2c^2)$. □

Remark. The nonconforming method (SN1) (or its counterpart (MN1)) automatically satisfies (3.5) for any $\boldsymbol{v} \in V^h$ (not only for $\boldsymbol{v} \in Z_h$) without condition (3.2), since the nonconforming part satisfies the Korn's second inequality element-wise.

Consistent error. The consistent error is introduced by the nonconforming term and can be expressed (for (MN2)) as

$$E_h(\boldsymbol{u}, \boldsymbol{v}) = 2\mu(\boldsymbol{\epsilon}(\boldsymbol{u}), \boldsymbol{\epsilon}(\boldsymbol{v}^c)) - (\mathrm{div}_h \boldsymbol{v}, p) - f(\boldsymbol{v}^c) = -(\mathrm{div}_h \boldsymbol{v}^b, p).$$

Note that

$$2\mu(\boldsymbol{\epsilon}(\boldsymbol{u}), \boldsymbol{\epsilon}(\boldsymbol{v}^c)) - (\mathrm{div}\, \boldsymbol{v}^c, p) - f(\boldsymbol{v}^c) = 0.$$

Use (3.1) and (3.2), the consistent error can be estimated as following.

$$\sup_{\boldsymbol{v} \in V^h, \boldsymbol{v} \neq 0} \frac{-(\mathrm{div}_h\, \boldsymbol{v}^b, p)}{\|\boldsymbol{v}^b\|_h} = \sup_{\boldsymbol{v} \in V^h, \boldsymbol{v} \neq 0} \frac{-(\mathrm{div}_h\, \boldsymbol{v}^b, (I - \Pi_0)p)}{\|\boldsymbol{v}^b\|_h} \leq Ch\|p\|_1.$$

Stability: the inf-sup condition. We need to characterize a subspace of W_h,

$$\ker b_h' = \{q \in W_h | \quad (\mathrm{div}_h \boldsymbol{v}, q) = 0,\ \forall \boldsymbol{v} \in V^h\}.$$

(Recall that $\ker b' = \{q \in W | \quad (\mathrm{div}\, \boldsymbol{v}, q) = 0,\ \forall \boldsymbol{v} \in V\} = \{0\}$.)

Proposition. Under the conditions (3.1) and (3.3),

$$\ker b_h' = \{\bar{q} \in \bar{W}_h | \quad (\mathrm{div}\, \boldsymbol{v}^c, \bar{q}) = 0,\ \forall \boldsymbol{v}^c \in S^h\}. \tag{3.6}$$

Proof: For $\boldsymbol{v} = \boldsymbol{v}^c + \boldsymbol{v}^b \in V_h$ and $q = \bar{q} + (I - \Pi_0)q \in W_h$, we have

$$(\mathrm{div}_h \boldsymbol{v}, q) = (\mathrm{div}\, \boldsymbol{v}^c, \bar{q}) + (\mathrm{div}_h \boldsymbol{v}, (I - \Pi_0)q)$$

by (3.1). Given $q \in W_h$, $\boldsymbol{v}^c \in S^h$, we can assume that $(\mathrm{div}\, \boldsymbol{v}^c, \bar{q}) \geq 0$ (otherwise, use $-\boldsymbol{v}^c$). Choose $\boldsymbol{v}^b \in B_h$ according to (3.3) which yields,

$$(\mathrm{div}_h \boldsymbol{v}, q) = (\mathrm{div}\, \boldsymbol{v}^c, \bar{q}) + \|(I - \Pi_0)q\|_0^2.$$

Therefore, $(\mathrm{div}_h \boldsymbol{v}, q) = 0$ iff $(\mathrm{div}\, \boldsymbol{v}^c, \bar{q}) = 0$ and $(I - \Pi_0)q = 0$ which establishes (3.6). □

The kernel is exactly the same as for the Q_1-P_0 element and it is well known that this kernel is the checkerboard function [3]. Therefore, by the same filtering procedure (see [3]) and with the help of (3.3)–(3.4), we are able to establish the inf-sup condition for a modified pair $V_*^h \times W_h^*$. Indeed, if we are only interested in the displacement $\boldsymbol{u}$, the filtering is not necessary. By the general theory of the mixed method, we are able to establish

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_h+\|p-p_h^*\|_0 \le C(\inf_{\boldsymbol{v}\in V_*^h} \|\boldsymbol{u}-\boldsymbol{v}\|_h+\inf_{q\in W_h^*} \|p-q\|_0+\sup_{\boldsymbol{v}\in V_*^h\setminus\{0\}} \frac{E_h(\boldsymbol{u},\boldsymbol{v})}{\|\boldsymbol{v}\|_h}).$$

By the approximation property of $V_*^h \times W_h^*$ and the consistent error, an error bound

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_h + \|p - p_h^*\|_0 \le Ch(\|\boldsymbol{u}\|_2 + \|p\|_1)$$

can be obtained. See [7] for details.

4. A Final Remark. Numerical experiments (see [5]) indicate that for nearly incompressible material ($\lambda >> 1$), the nonconforming method and the enhanced strain method with their selective counterparts are compatible with the Q_1-P_0 element, while for regular problems their performance is much improved over the bilinear element. Therefore the methods we have discussed here can be used for general plane stress and plane strain elasticity problems for all the range of the Poisson ratio $\nu \in [0, 1/2)$.

References

1. F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
2. R.S. Falk, Nonconforming finite element methods for the equations of linear elasticity, Math. Comp., 57 (1991), 529–550.
3. V. Girault and P.A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Theory and Algorithms, Springer-Verlag, Berlin, 1986.
4. T.J.R. Hughes, *The Finite Element Method*, Linear Static and Dynamic Finite Element Analysis, Prentice Hall, New Jersey, 1987.
5. H.M. Karunadasa, *Selective nonconforming finite element methods for elasticity problems*, Thesis, Department of Mathematics, Texas Tech, 1995.
6. B.D. Reddy and J.C. Simo, Stability and convergence of a class of enhanced strain finite element methods, SIAM J. Numer. Anal. 32 (1995), 1705–1728.
7. Z. Zhang, Stability analysis of Wilson's element for incompressible elasticity, Tech. Report No. 94-ZZ2, June, 1994, Department of Mathematics, Texas Tech University.
8. Z. Zhang, Analysis of some quadrilateral nonconforming elements for incompressible elasticity, SIAM J. Numer. Anal., 34 (1997), 640–663.

Editorial Policy

§1. Volumes in the following four categories will be published in LNCSE:

i) Research monographs
ii) Lecture and seminar notes
iii) Conference proceedings
iv) Textbooks

Those considering a book which might be suitable for the series are strongly advised to contact the publisher or the series editors at an early stage.

§2. Categories i) and ii). These categories will be emphasized by Lecture Notes in Computational Science and Engineering. **Submissions by interdisciplinary teams of authors are encouraged.** The goal is to report new developments – quickly, informally, and in a way that will make them accessible to non-specialists. In the evaluation of submissions timeliness of the work is an important criterion. Texts should be well-rounded, well-written and reasonably self-contained. In most cases the work will contain results of others as well as those of the author(s). In each case the author(s) should provide sufficient motivation, examples, and applications. In this respect, Ph.D. theses will usually be deemed unsuitable for the Lecture Notes series. Proposals for volumes in these categories should be submitted either to one of the series editors or to Springer-Verlag, Heidelberg, and will be refereed. A provisional judgment on the acceptability of a project can be based on partial information about the work: a detailed outline describing the contents of each chapter, the estimated length, a bibliography, and one or two sample chapters – or a first draft. A final decision whether to accept will rest on an evaluation of the completed work which should include

- at least 100 pages of text;
- a table of contents;
- an informative introduction perhaps with some historical remarks which should be accessible to readers unfamiliar with the topic treated;
- a subject index.

§3. Category iii). Conference proceedings will be considered for publication provided that they are both of exceptional interest and devoted to a single topic. One (or more) expert participants will act as the scientific editor(s) of the volume. They select the papers which are suitable for inclusion and have them individually refereed as for a journal. Papers not closely related to the central topic are to be excluded. Organizers should contact Lecture Notes in Computational Science and Engineering at the planning stage.

In exceptional cases some other multi-author-volumes may be considered in this category.

§4. Category iv) Textbooks on topics in the field of computational science and engineering will be considered. They should be written for courses in CSE education. Both graduate and undergraduate level are appropriate. Multidisciplinary topics are especially welcome.

§5. Format. Only works in English are considered. They should be submitted in camera-ready form according to Springer-Verlag's specifications. Electronic material can be included if appropriate. Please contact the publisher. Technical instructions and/or TEX macros are available via http://www.springer.de/author/tex/help-tex.html; the name of the macro package is "LNCSE – LaTEX2e class for Lecture Notes in Computational Science and Engineering". The macros can also be sent on request.

General Remarks

Lecture Notes are printed by photo-offset from the master-copy delivered in camera-ready form by the authors. For this purpose Springer-Verlag provides technical instructions for the preparation of manuscripts. See also *Editorial Policy*.

Careful preparation of manuscripts will help keep production time short and ensure a satisfactory appearance of the finished book. The actual production of a Lecture Notes volume normally takes approximately 12 weeks.

The following terms and conditions hold:

Categories i), ii), and iii):
Authors receive 50 free copies of their book. No royalty is paid. Commitment to publish is made by letter of intent rather than by signing a formal contract. Springer-Verlag secures the copyright for each volume.

For conference proceedings, editors receive a total of 50 free copies of their volume for distribution to the contributing authors.

Category iv):
Regarding free copies and royalties, the standard terms for Springer mathematics monographs and textbooks hold. Please write to Peters@springer.de for details. The standard contracts are used for publishing agreements.

All categories:
Authors are entitled to purchase further copies of their book and other Springer mathematics books for their personal use, at a discount of 33,3 % directly from Springer-Verlag.

Addresses:

Professor M. Griebel
Institut für Angewandte Mathematik
der Universität Bonn
Wegelerstr. 6
D-53115 Bonn, Germany
e-mail: griebel@iam.uni-bonn.de

Professor D. E. Keyes
Computer Science Department
Old Dominion University
Norfolk, VA 23529–0162, USA
e-mail: keyes@cs.odu.edu

Professor R. M. Nieminen
Laboratory of Physics
Helsinki University of Technology
02150 Espoo, Finland
e-mail: rni@fyslab.hut.fi

Professor D. Roose
Department of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200A
3001 Leuven-Heverlee, Belgium
e-mail: dirk.roose@cs.kuleuven.ac.be

Professor T. Schlick
Department of Chemistry and
Courant Institute of Mathematical
Sciences
New York University
and Howard Hughes Medical Institute
251 Mercer Street, Rm 509
New York, NY 10012-1548, USA
e-mail: schlick@nyu.edu

Springer-Verlag, Mathematics Editorial
Tiergartenstrasse 17
D-69121 Heidelberg, Germany
Tel.: *49 (6221) 487-185
e-mail: peters@springer.de
http://www.springer.de/math/
peters.html

Lecture Notes in Computational Science and Engineering

Vol. 1 D. Funaro, *Spectral Elements for Transport-Dominated Equations.* 1997. X, 211 pp. Softcover. ISBN 3-540-62649-2

Vol. 2 H. P. Langtangen, *Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming. 1999. XXIII, 682 pp. Hardcover. ISBN 3-540-65274-4

Vol. 3 W. Hackbusch, G. Wittum (eds.), *Multigrid Methods V.* Proceedings of the Fifth European Multigrid Conference held in Stuttgart, Germany, October 1-4, 1996. 1998. VIII, 334 pp. Softcover. ISBN 3-540-63133-X

Vol. 4 P. Deuflhard, J. Hermans, B. Leimkuhler, A. E. Mark, S. Reich, R. D. Skeel (eds.), *Computational Molecular Dynamics: Challenges, Methods, Ideas.* Proceedings of the 2nd International Symposium on Algorithms for Macromolecular Modelling, Berlin, May 21-24, 1997. 1998. XI, 489 pp. Softcover. ISBN 3-540-63242-5

Vol. 5 D. Kröner, M. Ohlberger, C. Rohde (eds.), *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws.* Proceedings of the International School on Theory and Numerics for Conservation Laws, Freiburg / Littenweiler, October 20-24, 1997. 1998. VII, 285 pp. Softcover. ISBN 3-540-65081-4

Vol. 6 S. Turek, *Efficient Solvers for Incompressible Flow Problems.* An Algorithmic and Computational Approach. 1999. XVII, 352 pp, with CD-ROM. Hardcover. ISBN 3-540-65433-X

Vol. 7 R. von Schwerin, *Multi Body System SIMulation.* Numerical Methods, Algorithms, and Software. 1999. XX, 338 pp. Softcover. ISBN 3-540-65662-6

Vol. 8 H.-J. Bungartz, F. Durst, C. Zenger (eds.), *High Performance Scientific and Engineering Computing.* Proceedings of the International FORTWIHR Conference on HPSEC, Munich, March 16-18, 1998. 1999. X, 471 pp. Softcover. 3-540-65730-4

Vol. 9 T. J. Barth, H. Deconinck (eds.), *High-Order Methods for Computational Physics.* 1999. VII, 582 pp. Hardcover. 3-540-65893-9

Vol. 10 H. P. Langtangen, A. M. Bruaset, E. Quak (eds.), *Advances in Software Tools for Scientific Computing.* 2000. X, 357 pp. Softcover. 3-540-66557-9

Vol. 11 B. Cockburn, G. E. Karniadakis, C.-W. Shu (eds.), *Discontinuous Galerkin Methods.* Theory, Computation and Applications. 2000. XI, 470 pp. Hardcover. 3-540-66787-3

For further information on these books please have a look at our mathematics catalogue at the following URL: `http://www.springer.de/math/index.html`