The IMA Volumes in Mathematics and its Applications

Xiaobing Feng
Ohannes Karakashian
Yulong Xing *Editors*

Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations

2012 John H Barrett Memorial Lectures

Springer

The IMA Volumes in Mathematics and its Applications
Volume 157

For further volumes:
http://www.springer.com/series/811

Institute for Mathematics and its Applications (IMA)

The Institute for Mathematics and its Applications was established by a grant from the National Science Foundation to the University of Minnesota in 1982. The primary mission of the IMA is to foster research of a truly interdisciplinary nature, establishing links between mathematics of the highest caliber and important scientific and technological problems from other disciplines and industries. To this end, the IMA organizes a wide variety of programs, ranging from short intense workshops in areas of exceptional interest and opportunity to extensive thematic programs lasting a year. IMA Volumes are used to communicate results of these programs that we believe are of particular value to the broader scientific community.

The full list of IMA books can be found at the Web site of the Institute for Mathematics and its Applications:

http://www.ima.umn.edu/springer/volumes.html.

Presentation materials from the IMA talks are available at

http://www.ima.umn.edu/talks/.

Video library is at

http://www.ima.umn.edu/videos/.

Fadil Santosa, Director of the IMA

Xiaobing Feng • Ohannes Karakashian
Yulong Xing
Editors

Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations

2012 John H Barrett Memorial Lectures

Springer

Editors
Xiaobing Feng
Department of Mathematics
The University of Tennessee
Knoxville, TN, USA

Ohannes Karakashian
Department of Mathematics
The University of Tennessee
Knoxville, TN, USA

Yulong Xing
Department of Mathematics
The University of Tennessee
Knoxville, TN, USA

ISSN 0940-6573
ISBN 978-3-319-01817-1 ISBN 978-3-319-01818-8 (eBook)
DOI 10.1007/978-3-319-01818-8
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013949279

Mathematics Subject Classification (2010): 65M12, 65M15, 65M20, 65N12, 65N15, 65N30

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

This volume was based on the 2012 Barrett Lectures on "Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations (PDEs)" which was partially funded by the Institute for Mathematics and its Applications (IMA) at the University of Minnesota. The workshop took place at the University of Tennessee at Knoxville from May 9–11, 2012. We would like to thank all the participants for making this a stimulating and productive workshop. In particular we would like to thank the organizers, Xiaobing Feng Ohannes Karakashian and Yulong Xing for organizing this volume which came out of the invited speakers at the workshop.

We also take this opportunity to thank the National Science Foundation for its support of the IMA.

Minneapolis, MN

Fadil Santosa
Jiaping Wang

Preface

The John H. Barrett Memorial Lectures at the University of Tennessee, Knoxville (UTK) were established in honor of John H. Barrett, a distinguished researcher in ordinary differential equations and department head, at the time of his death in 1969. The Lectures have been given annually since 1970 in a variety of mathematical fields by a succession of distinguished lecturers. The topic of the Barrett Lectures changes from year to year and is chosen to reflect research interests within the Department of Mathematics at the University of Tennessee. Since 1993 the Lectures have consisted of two or three one-hour survey talks by each of two or three leading researchers, representing different themes/directions in a single field. The Lectures are partly funded by a grant from Mathematics Department of the University of Tennessee and have often received additional support from the National Science Foundation. They attract wide interest, with an audience of between 40 and 60 participants from the whole country, in addition to faculty and students from UTK and the Oak Ridge National Laboratory. They represent one of the few long-standing lecture series in mathematics in the southeastern USA.

The 2012 Barrett Lectures, which is the 42nd Lecture in the series, were held on the campus of the University of Tennessee at Knoxville from May 9–11, 2012. The topic of the 2012 Barrett Lectures was "Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations (PDEs)," which is a hot topic and has a broad appeal to researchers from applied sciences and engineering. One of the primary goals of the Barrett Lectures is to bring prominent researchers working in such active areas to UTK, as a service to the university and to the southeastern region of the USA. As one of the few long running lecture series in mathematics in the southeastern USA, plus the popularity and broad appeal of its topic, it was expected that there was a large attendance of the Lectures in 2012. About 70 people from the USA and Europe attended

the Lectures and half of the attendees are junior researchers (assistant professors, postdocs, and graduate students).

The main speakers of the 2012 Barrett Lectures were *Franco Brezzi* of University of Pavia (Italy) and *Chi-Wang Shu* of Brown University. Each of them delivered three one-hour survey lectures with Franco Brezzi focusing on theoretical aspects of discontinuous Galerkin methods for elliptic problems and Chi-Wang Shu on discontinuous Galerkin methods for evolution problems and their applications. The titles of their talks were:

1. Franco Brezzi (University of Pavia, Italy)
   - Part I: Theoretical aspects of discontinuous Galerkin (DG) methods for stationary problems: mathematical background of DG methods.
   - Part II: Classical DG methods for second and fourth order elliptic problems.
   - Part III: Connections between DG and other methods.
2. Chi-Wang Shu (Brown University)
   - Discontinuous Galerkin methods for time-dependent problems: survey and recent developments: Parts I–III

In addition to the two main speakers, there were also ten one-hour invited talks for the 2012 Lectures. These ten speakers and titles of their talks were

1. Slimane Adjerid (Virginia Tech): Accurate error estimates and superconveregnce for DG methods.
2. Susanne Brenner (Louisiana State University): C^0 interior penalty methods.
3. Bernardo Cockburn (University of Minnesota): Devising superconvergent DG methods.
4. Clint Dawson (University of Texas at Austin): Local time stepping in DG methods and applications to the shallow water equations.
5. Leszek Demkowicz (University of Texas at Austin): Discontinuous Petrov–Galerkin methods with optimal test functions.
6. Jean-Luc Guermond (Texas A & M University): Discontinuous Galerkin methods for the radiative transport equation.
7. Donatella Marini (University of Pavia, Italy): Virtual elements and DG.
8. Charalambos Makridakis (University of Crete, Greece): Transport, dispersion, and local reconstructions in discontinuous Galerkin methods.
9. Ricardo Nochetto (University of Maryland): Time-discrete higher order ALE formulations: a DG approach.
10. Beatrice Riviere (Rice University): Coupled free flows and porous media flows.

This book contains articles from 11 speakers, each of whom is a leading researcher in the field of discontinuous Galerkin finite element methods and its applications. Following the tradition of the Barrett Lectures, several of these articles are in-depth survey papers with an expository discussion that should make this book a useful reference for researchers both in and out-

side the field, including other applied science and engineering communities, young researchers, and graduate students.

The 2012 Barrett Lectures were partially funded by a grant from the National Science Foundation (DMS-1203237), by the Institute for Mathematics and its Applications (IMA) at University of Minnesota, and by Research Office and College of Arts and Sciences as well as Department of Mathematics at the University of Tennessee, Knoxville. The organizers, together with all attendees, are grateful to these funding agencies. Their generous support made the Lectures possible and, among other things, allowed the organizers to fund the participation of young researchers including graduate students and recent postdocs.

Finally, we would like to express our thanks to Ms. Connie Mroz and Jane Parker, the secretaries of the 2012 Barrett Lectures, who made all organizational details run smoothly, and Juvy Melton, who helped to do the budget and all the paper work for grant applications. We would also like to thank Ben Walker, Angela Woofter, Thomas Lewis, and Cody Lorton for their various help during the Lectures.

This is the first time when the proceedings of the Barrett Lectures is published as a volume in the IMA book series. We are grateful to Professor Fadil Santosa, the director of IMA, for his enthusiasm and encouragement to publish the proceedings. We would also like to thank Katherine Cramer of IMA and Achi Dosanjh of Springer for their help during the course of the preparation.

Knoxville, TN

Xiaobing Feng
Ohannes Karakashian
Yulong Xing

Contents

A QUICK TUTORIAL ON DG METHODS FOR ELLIPTIC PROBLEMS

F. BREZZI* AND L.D. MARINI†

Abstract. In this paper we recall a few basic definitions and results concerning the use of DG methods for elliptic problems. As examples we consider the Poisson problem and the linear elasticity problem. A hint on the nearly incompressible case is given, just to show one of the possible advantages of DG methods over continuous ones. At the end of the paper we recall some physical principles for linear elasticity problems, just to open the door towards possible new developments.

AMS Classification 65N12, 65N15, 65N30

Key words. Discontinuous Galerkin, Elliptic problems, Linear elasticity

1. Introduction. The main purpose of this paper is to present the basic features of Discontinuous Galerkin Methods for elliptic problems. We will give some hints on the basic mathematical tools typically used to study and analyze them and on their capability to avoid some common troubles (as the discretization of nearly incompressible materials). We will state approximation properties and show how to derive a-priori estimates. We will also present some possible variants and relationships with other approaches (as mixed or hybrid methods for linear elasticity) that possibly deserve a deeper analysis.

The paper is addressed to readers with a more engineering oriented background, and an interest in continuum mechanics, with the idea to help them in getting more familiar with the basic concepts and features of DG methods, that indeed, according to the latest developments, show an interesting potential also in structural problems.

Actually, applications of DG methods to other problems, and in particular to hyperbolic problems, conservation laws and the like, started already forty years ago, and are fully developed nowadays (see, e.g., [19, 35]). In this book these applications are discussed at a much higher level (starting from the "parallel" contribution of Chi Wang Shu [36]); this is quite natural, since the interested people are (in general) already acquainted with all the basic instruments and with the applications to the more common problems.

Instead, most practitioners in structural engineering and continuum mechanics, so far, are not yet familiar with the use of DG methods, that have been pushed forward mainly by applied mathematicians and more

*Istituto Universitario di Studi Superiori (IUSS), Piazza della Vittoria 15, 27100 Pavia, Italy and Department of Mathematics, King Abulaziz University, PO Box 80203, Jeddah 21589, Saudi Arabia, brezzi@imati.cnr.it

†Dipartimento di Matematica, Università di Pavia, Via Ferrata 1, 27100 Pavia, Italy, marini@imati.cnr.it

X. Feng et al. (eds.), *Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations*, The IMA Volumes in Mathematics and its Applications 157, DOI 10.1007/978-3-319-01818-8_1,

"mathematically oriented" engineers. Hence the idea of addressing people who are less familiar with the DG machinery but are interested in trying them on their problems.

We will not discuss issues related to a-posteriori estimates and mesh-adaptivity, a very interesting subject that, however, goes beyond the scope of this paper. For this we refer, for instance, to [30–32]. For the same reason, we will not discuss matters related to the solution of the final linear systems, such as the construction of efficient solvers and preconditioners (see, e.g., [9, 25]).

The paper is organized as follows. In Sect. 2 we recall some basic instruments, such as Poincaré, trace, and inverse inequalities, that will be useful in the sequel. In Sect. 3 typical tools for dealing with discontinuous functions are introduced: jumps and averages, norms and bounds for the edge contributions. Sect. 4 is devoted to the treatment of the Poisson problem; the most common DG schemes are derived and proved to be stable and consistent. Error estimates are also recalled. In Sect. 5 linear elasticity problems are treated, including the nearly incompressible case where the use of DG approximations proves to be particularly well suited for dealing with the so-called locking phenomenon. Finally, in Sect. 6 we recall some basic physical principles that are at the basis of alternative variational formulations (always for linear elasticity). The use of DG discretizations for many of these formulations is still at the beginning, and their potential is, in our opinion, still to be fully assessed.

Throughout the paper we shall follow the usual notation for Sobolev norms and seminorms, as, for instance, in [18]. Hence, for a geometric object $\mathcal{O}$ (as an edge, or an element, or a general domain) and a smooth-enough function v defined on $\mathcal{O}$, we will denote by

$$\|v\|^2_{0,\mathcal{O}} \equiv |v|^2_{0,\mathcal{O}} \equiv \int_{\mathcal{O}} v^2 \, d\mathcal{O}$$

the (square) norm of v in $\mathbb{L}^2(\mathcal{O})$. On the other hand, the notation $|v|^2_{k,\mathcal{O}}$ will indicate, for k integer ≥ 1, the square of the seminorm of v obtained summing all the squares of the $\mathbb{L}^2$ norms of all the derivatives of order k. Hence, in two dimensions,

$$|v|^2_{1,\mathcal{O}} \equiv \left|\frac{\partial v}{\partial x_1}\right|^2_{0,\mathcal{O}} + \left|\frac{\partial v}{\partial x_2}\right|^2_{0,\mathcal{O}}$$

$$|v|^2_{2,\mathcal{O}} \equiv \left|\frac{\partial^2 v}{\partial x_1^2}\right|^2_{0,\mathcal{O}} + \left|\frac{\partial^2 v}{\partial x_1 \partial x_2}\right|^2_{0,\mathcal{O}} + \left|\frac{\partial^2 v}{\partial x_2^2}\right|^2_{0,\mathcal{O}},$$

and so on.

2. Some Basic Mathematical Instruments. We start with a few very basic inequalities. We will give a rather detailed proof in one dimension, and often only a general idea on the case of several dimensions.

2.1. Poincaré Inequality. Let $v \in C^1([0,T])$ with $v = 0$ at $t_0 \in [0,T]$. Then, using the fundamental theorem of calculus we get

$$v(t) = \int_{t_0}^{t} v'(\tau)\, d\tau, \qquad \text{then, taking the absolute values,}$$

$$|v(t)| \le \int_0^T |v'(\tau)|\, d\tau \qquad \text{then squaring both sides and using C-S}$$

$$|v(t)|^2 \le \left(\int_0^T |v'(\tau)|\, d\tau^2 \right) \le T \int_0^T |v'(\tau)|^2\, d\tau; \text{ integrating from 0 to } T$$

$$\int_0^T |v(t)|^2\, dt \le T^2 \int_0^T |v'(\tau)|^2\, d\tau. \tag{2.1}$$

2.2. Trace Inequalities. Let $v \in C^1([0,T])$ Then, using the fundamental theorem of calculus on the function v^2:

$$v^2(0) = v^2(t) - \int_0^t (v^2(\tau))'\, d\tau \qquad \text{and then taking the absolute value}$$

$$v^2(0) \le v^2(t) + \int_0^T |2v(\tau)v'(\tau)|\, d\tau \text{ then multiplying and dividing by } \sqrt{T}$$

$$v^2(0) \le v^2(t) + \int_0^T 2\frac{|v(\tau)|}{\sqrt{T}} \sqrt{T}|v'(\tau)|\, d\tau \qquad \text{and using } 2ab \le a^2 + b^2$$

$$v^2(0) \le v^2(t) + \int_0^T \left(\frac{|v(\tau)|^2}{T} + T|v'(\tau)|^2 \right) d\tau; \quad \text{integrating from 0 to } T$$

$$T\, v^2(0) \le \int_0^T v^2(t)\, dt + T \int_0^T \left(\frac{v^2(\tau)}{T} + T(v'(\tau))^2 \right) d\tau; \text{ and dividing by } T$$

$$v^2(0) \le \int_0^T \left(\frac{2}{T} v^2(\tau) + T(v'(\tau))^2 \right) d\tau \le \frac{2}{T} \|v\|_0^2 + T|v|_1^2. \tag{2.2}$$

2.3. Comments on the Above Inequalities. Note that both in (2.1) and in (2.2) the *physical dimensions* of the two terms coincide. In particular in (2.1) we have

$$\left[\int v^2\, dt \right] \equiv [v]^2[t] \equiv [v]^2[t] \left[\frac{[t]^2}{[t]^2} \right] \equiv [t]^2 \left[\frac{[v]^2[t]}{[t]^2} \right]$$
$$\equiv [t]^2 \left[\frac{[v]}{[t]} \right]^2 [t] \equiv [t]^2 \left[\int |v'|^2\, dt \right].$$

Similarly, considering (2.2) we easily check that

$$[v]^2 \equiv \left[\frac{1}{t} \right] [v]^2[t] \equiv \frac{1}{[t]} \left[\int v^2\, dt \right]$$

and

$$[v]^2 \equiv [t] \left[\frac{v}{t}\right]^2 [t] \equiv [t] \left[\int |v'|^2 \, dt\right]$$

A rough interpretation of the trace inequality could be: if the value at one point is big, then either the function has a big integral [and you can use the first piece in the right-hand side of (2.2)] or it has a big derivative [and you can use the second piece in the right-hand side of (2.2)]. See Fig. 1.

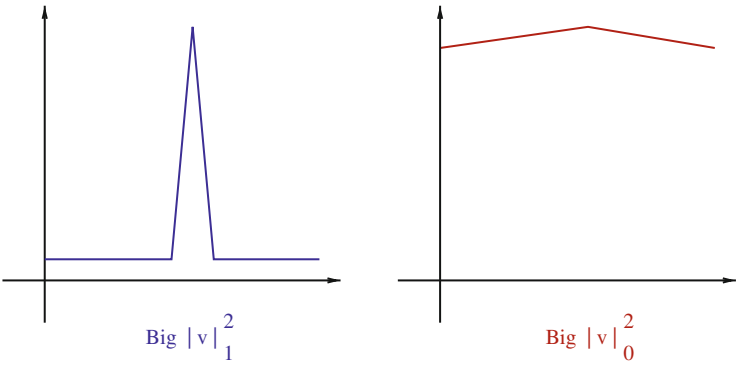


FIG. 1. Trace inequality: If you are big at one point, either you go down quickly (and have a big $|\cdot|_1$ norm), or you stay up, and have a big $|\cdot|_0$ norm.

If you are big at one point (say, at zero), either you go down quickly (and have a big $|\cdot|_1$ norm), or you stay up, and have a big $|\cdot|_0$ norm. See Fig. 1.

2.4. 2-D Versions. We summarize here the two-dimensional versions of the above inequalities, with some picture that indicates a possible proof (using the one-dimensional results) in some particular geometries.

In Fig. 2 we illustrate Poincaré inequality for functions v vanishing on the edge of $\partial\Omega$ contained in the x axis.
From the 1-d case

$$\|v(x,\cdot)\|^2_{0,]0,T[} \le T^2 |v(x,\cdot)|^2_{1,]0,T[}$$

we deduce

$$\|v\|^2_{0,\Omega} \le T^2 \left\| \frac{\partial v}{\partial t} \right\|^2_{0,\Omega}.$$

At a more general level, we already saw that in the estimate (2.1) the physical dimensions match. It is easy to see that, in a more general bounded domain $K \subset I\!R^d$ with characteristic length ℓ we have

$$\left[\|v\|^2_{0,K}\right] = [v]^2 [\ell]^d \qquad \text{and} \qquad \left[|v|^2_{1,K}\right] = [v]^2 [\ell]^{d-2}$$

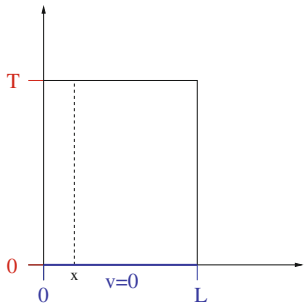


FIG. 2. Poincaré inequality in two dimensions

so that a natural guess is

$$\|v\|^2_{0,K} \leq (d(K))^2 |v|^2_{1,K} \tag{2.3}$$

where $d(K)$ is the diameter of K, and where we have to assume, for instance, that v has zero mean value on K (or some other condition that allows to take care of constant functions).

Possibly this is a good moment to point out that the widely used definition

$$\|v\|^2_{1,K} := \|v\|^2_{0,K} + |v|^2_{1,K}$$

doesn't make any sense, unless everything has been adimensionalized: a practice rather unhealthy from the engineering point of view.

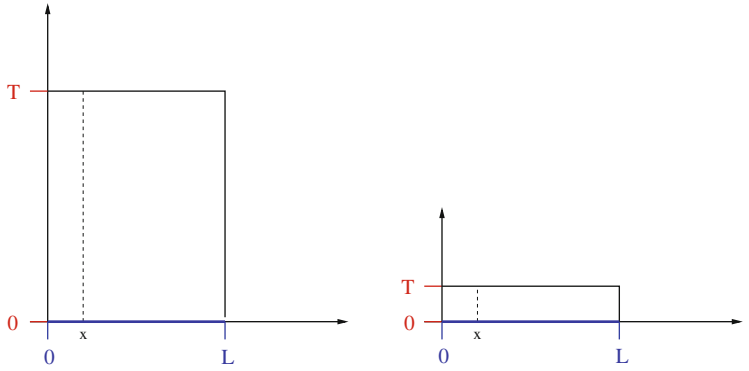


FIG. 3. Trace inequality in two dimensions

Concerning instead the trace inequality, from the one-dimensional case we have, again for the rectangle $K \equiv]0, L[\times]0, T[$

$$|v^2(x,0)| \leq \frac{2}{T}\|v(x,t)\|^2_{0,\,]0,T[} + T\left\|\frac{\partial v}{\partial t}(x,t)\right\|^2_{0,\,]0,T[} \qquad \forall x \in]0, L[.$$

Here, and in what follows, $]a,b[$ denotes the open interval (a,b). By integrating in x from 0 to L we have:

$$\|v(\cdot,0)\|^2_{0,\,]0,L[} \le \frac{2}{T}\|v\|^2_{0,K} + T\left|\frac{\partial v}{\partial t}\right|^2_{0,\,K}$$

from which we reasonably guess the more general version

$$\|v\|^2_{0,\partial K} \le C\left(\ell^{-1}\|v\|^2_{0,K} + \ell|v|^2_{1,K}\right) \tag{2.4}$$

where both the constant C and the characteristic length ℓ can depend on several geometric features (see Fig. 3 for a simple example where the bound on the L^2 norm of the trace on the lower edge of the rectangle depends on the height T of the rectangle), but the constant C does not depend on *the size* of K.

2.5. Inverse Inequalities. In a finite dimensional space, all norms are equivalent, in the sense that for any two norms $\|\cdot\|_@$ and $\|\cdot\|_\#$ there exist two positive constants c and C such that

$$c\,\|v\|_@ \le \|v\|_\# \le C\,\|v\|_@ \qquad \text{for every } v \text{ in the space.}$$

However if the norms are, say, $\|v\|_{0,K}$ and $\|v\|_{1,K}$ the constants c and C might **depend on the size of** K. Indeed we already saw in (2.3) that in the inequality

$$\|v\|^2_{0,K} \le C|v|^2_{1,K} \tag{2.5}$$

the constant C should have physical dimension

$$[\,C\,] = [\ell]^2$$

and actually behave as the square of the diameter of K.

On the other hand, it is natural to ask the question whether one could have (in one dimension, to start with) an inequality of the type

$$h^2|v|^2_{1,\,]0,h[} \le C\|v\|^2_{0,\,]0,h[}$$

for some constant C. But taking

$$v = \sin\left(\frac{2\pi kx}{h}\right) \qquad k \in \mathbb{N}, \tag{2.6}$$

we have

$$\|v\|^2_{0,\,]0,h[} = \frac{h}{2} \qquad h^2|v|^2_{1,\,]0,h[} = 4\pi^2k^2\frac{h}{2}$$

and our dreams dissolve.

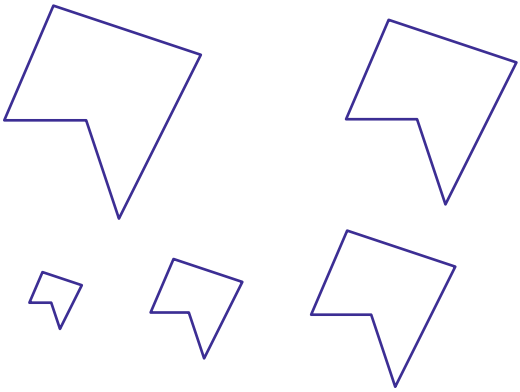


FIG. 4. Homothetic elements

However....we cannot fit **all** the functions (2.6), for **all** the possible integers k, in a single *finite dimensional space*! Hence the inequality

$$h^2|v|^2_{1,]0,h[} \leq C\|v\|^2_{0,]0,h[}$$

has still some possibilities, **if** we are ready to accept a constant C that depends on the finite dimensional subspace I am using. For instance, for $v = (x/h)^r$ (with r integer ≥ 1) we have

$$h^2|v|^2_{1,]0,h[} = \frac{h\,r^2}{2r-1} \quad \text{and} \quad \|v\|^2_{0,]0,h[} = \frac{h}{2r+1} \tag{2.7}$$

and we might get away with a constant C that depends on the degree of the polynomials (and actually this **is** the case). For instance, in the case of (2.7) we have

$$h^2|v|^2_{1,]0,h[} \leq \frac{r^2(2r+1)}{2r-1}\|v\|^2_{0,]0,h[} \leq 3r^2\,\|v\|^2_{0,]0,h[}.$$

More generally, for a family of homothetic elements (see Fig. 4) and an integer r there exists a $C = C(r)$ such that

$$|v|^2_{1,K} \leq Ch^{-2}\|v\|^2_{0,K}$$

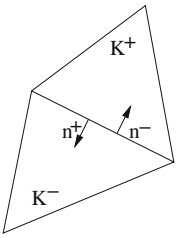
where $h =$ diameter of K, and the inequality holds for every v polynomial of degree $\leq r$, and even more generally (see e.g. [18])

$$|v|^2_{s,K} \leq Ch^{-2(s-k)}\|v\|^2_{k,K} \qquad \text{for } s,\ k \text{ integers with } s \geq k \tag{2.8}$$

For a much wider and deeper review of the basic mathematical instruments for dealing with Finite Elements we refer, for instance, to [13].

3. Some Inequalities for DG Elements. As we want to deal with spaces of piecewise polynomials that can be discontinuous from one element to the neighboring one, it is natural to begin by considering the simplest case of two triangles (as in the next figure) and functions that are polynomials separately in each triangle (and possibly discontinuous from one triangle to the other).

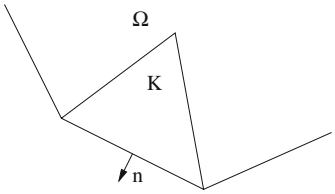
3.1. Definition of Averages and Jumps. If K^+ and K^- are two elements with an edge e in common, we denote by $\mathbf{n}^+$ and $\mathbf{n}^-$ the outward unit normal at e of K^+ and K^-, respectively.



Then for every pair (v^+, v^-) of smooth functions on K^+ and K^-, respectively, and for every pair $(\boldsymbol{\tau}^+, \boldsymbol{\tau}^-)$ of smooth vector valued functions on K^+ and K^-, respectively, we set

$$\{\!\{v\}\!\} := \frac{1}{2}(v^+ + v^-), \qquad [\![v]\!] := v^+\mathbf{n}^+ + v^-\mathbf{n}^-$$
$$\{\!\{\boldsymbol{\tau}\}\!\} := \frac{1}{2}(\boldsymbol{\tau}^+ + \boldsymbol{\tau}^-), \qquad [\![\boldsymbol{\tau}]\!] := \boldsymbol{\tau}^+ \otimes \mathbf{n}^+ + \boldsymbol{\tau}^- \otimes \mathbf{n}^-$$

where $\mathbf{a} \otimes \mathbf{b} := \frac{1}{2}(\mathbf{a}\mathbf{b}^T + \mathbf{b}\mathbf{a}^T)$. We will also use the so-called *scalar jump*: $[\![\boldsymbol{\tau}]\!]_s \equiv [\![\boldsymbol{\tau}]\!]_{nn} = \boldsymbol{\tau}^+ \cdot \mathbf{n}^+ + \boldsymbol{\tau}^- \cdot \mathbf{n}^-$



On a boundary edge, instead, for every smooth function v and for every smooth vector valued function $\boldsymbol{\tau}$ we set

$$\{\!\{v\}\!\} := v \qquad [\![v]\!] := v\mathbf{n}$$
$$\{\!\{\boldsymbol{\tau}\}\!\} := \boldsymbol{\tau}, \qquad [\![\boldsymbol{\tau}]\!] := \boldsymbol{\tau} \otimes \mathbf{n}, \qquad [\![\boldsymbol{\tau}]\!]_s := \boldsymbol{\tau} \cdot \mathbf{n}$$

3.2. Piecewise Integrals. Given a decomposition (that for simplicity we assume compatible) of our computational domain Ω we denote:

- The set of all elements by $\mathcal{T}_h$,
- The set of all edges by $\mathcal{E}_h$,
- The set of all *internal* edges by $\mathcal{E}_h^0$,
- The set of all *boundary* edges by $\mathcal{E}_h^\partial$.

For the sake of simplicity we also assume $\mathcal{T}_h$ to be *quasi-uniform*, meaning that there exists a positive constant γ such that

$$h_{\min} \geq \gamma\, h_{\max}, \tag{3.1}$$

where $h_{\min}$ and $h_{\max}$ are the minimum and maximum diameter of the elements of $\mathcal{T}_h$, respectively. This will allow us to simplify notation, and use h to denote the characteristic length of all the elements of $\mathcal{T}_h$. Moreover we set:

$$(f,g)_{\mathcal{T}_h} := \sum_{K\in\mathcal{T}_h} \int_K f\, g \,\mathrm{d}x \qquad < f,g >_{\mathcal{E}_h} := \sum_{e\in\mathcal{E}_h} \int_e f\, g \,\mathrm{d}s$$

$$< f,g >_{\mathcal{E}_h^0} := \sum_{e\in\mathcal{E}_h^0} \int_e f\, g \,\mathrm{d}s \qquad < f,g >_{\mathcal{E}_h^\partial} := \sum_{e\in\mathcal{E}_h^\partial} \int_e f\, g \,\mathrm{d}s$$

3.3. The Magic Formula. For any piecewise smooth scalar function v, and for any piecewise smooth vector valued function $\boldsymbol{\tau}$ we have now

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} v\, \boldsymbol{\tau}\cdot\mathbf{n}_K \,\mathrm{d}s = < [\![v]\!], \{\!\{\boldsymbol{\tau}\}\!\} >_{\mathcal{E}_h} + < \{\!\{v\}\!\}, [\![\boldsymbol{\tau}]\!]_s >_{\mathcal{E}_h^0}. \tag{3.2}$$

The (elementary) proof is based on the algebraic equality:

$$a_1b_1 - a_2b_2 = \frac{1}{2}(a_1+a_2)(b_1-b_2) + \frac{1}{2}(b_1+b_2)(a_1-a_2)$$

3.4. Continuity of Edge Contributions. For piecewise smooth scalar functions u and v, using on each edge

$$|\{\!\{\nabla u\}\!\}| \leq (|\nabla u^+| + |\nabla u^-|)/2, \tag{3.3}$$

and the trace inequality (2.4) we have

$$\begin{aligned} < [\![v]\!], \{\!\{\nabla u\}\!\} >_{\mathcal{E}_h} &\leq \sum_{e\in\mathcal{E}_h} \left| \int_e [\![v]\!] \cdot \{\!\{\nabla u\}\!\}\, ds \right| \\ &\leq C \left(\sum_{e\in\mathcal{E}_h} \frac{1}{h} \|[\![v]\!]\|_{0,e}^2 \right)^{1/2} \left(\sum_{K\in\mathcal{T}_h} (\|\nabla u\|_{0,K}^2 + h^2 |\nabla u|_{1,K}^2) \right)^{1/2}. \end{aligned} \tag{3.4}$$

If u is a piecewise polynomial of degree $\leq r$, (3.4) becomes, thanks to the inverse inequality (2.8),

$$< [\![v]\!], \{\!\{\nabla u\}\!\} >_{\mathcal{E}_h} \leq C_r \left(\sum_{e\in\mathcal{E}} \frac{1}{h} \|[\![v]\!]\|_{0,e}^2 \right)^{1/2} \left(\sum_{K\in\mathcal{T}_h} \|\nabla u\|_{0,K}^2 \right)^{1/2}. \tag{3.5}$$

We define now, for v piecewise smooth and $k \in \mathbb{N}$:

$$\|v\|^2_{\text{jump}} := \sum_{e\in\mathcal{E}_h} \frac{1}{h}\|[\![v]\!]\|^2_{0,e}, \quad |\nabla v|^2_{k,h} := \sum_{K\in\mathcal{T}_h} |\nabla v|^2_{k,K}, \quad |v|^2_{k+1,h} := |\nabla v|^2_{k,h}.$$

Often we will write $\|\cdot\|_j$ instead of $\|\cdot\|_{\text{jump}}$. We also set:

$$\|v\|^2_{\text{DG}} := \|v\|^2_{\text{jump}} + |\nabla v|^2_{0,h} + h^2|\nabla v|^2_{1,h} \tag{3.6}$$

that, using (2.8), for piecewise polynomials of degree less than or equal to a given degree r is equivalent to

$$\|v\|^2_{\text{DG}} \simeq \|v\|^2_{\text{jump}} + |\nabla v|^2_{0,h}. \tag{3.7}$$

Then our continuity Eqs. (3.4) and (3.5) become, respectively,

$$< [\![v]\!], \{\!\{\nabla u\}\!\} >_{\mathcal{E}_h} \le C\|v\|_j(|u|_{1,h} + h^2|u|_{2,h}) \le C\|u\|_{\text{DG}}\|v\|_{\text{DG}} \tag{3.8}$$

and

$$< [\![v]\!], \{\!\{\nabla u\}\!\} >_{\mathcal{E}_h} \le C_r\|v\|_j|u|_{1,h} \le C_r\|u\|_{\text{DG}}\|v\|_{\text{DG}}. \tag{3.9}$$

In a quite similar way, using on each edge

$$\{\!\{v\}\!\} \le (|v^+| + |v^-|)/2 \qquad |[\![\nabla u]\!]| \le (|\nabla u^+| + |\nabla u^-|)/2 \tag{3.10}$$

one proves the inequalities

$$< \{\!\{v\}\!\}, [\![\nabla u]\!]_s >_{\mathcal{E}_h} \le C\|u\|_{\text{DG}}\|v\|_{\text{DG}} \tag{3.11}$$

and

$$< \{\!\{v\}\!\}, [\![\nabla u]\!]_s >_{\mathcal{E}_h} \le C_r\|u\|_{\text{DG}}\|v\|_{\text{DG}}. \tag{3.12}$$

Finally, for v piecewise smooth on a domain $\mathcal{O}$, the trace inequality gives

$$\|v\|^2_{\text{jump}} \le C\,(h^{-2}\|v\|^2_{0,\mathcal{O}} + |v|^2_{1,h}). \tag{3.13}$$

4. DG for the Poisson Problem. We consider now one of the simplest possible elliptic problems, in order to understand the behavior of DG methods. We will deal only with the more popular variants (SIPG, NIPG, IIPG). For a more detailed analysis of the numerous other variants of DG methods for the Poisson problem, we refer, for instance, to [5].

Given a two-dimensional domain Ω and $f \in L^2(\Omega)$ we look for u such that

$$-\Delta u = f \quad \text{in } \Omega \qquad \text{and} \qquad u = 0 \quad \text{on } \partial\Omega \tag{4.1}$$

Given a decomposition of Ω into triangles (for simplicity) we want to use a DG method. We fix, once and for all, the degree r of the local polynomials, and *we define V_h as the space of functions v_h that are piecewise polynomials of degree $\leq r$ on Ω and can be discontinuous from one triangle to another.* For a $v_h \in V_h$ we have

$$\int_\Omega -\Delta u\, v_h \mathrm{d}x = \sum_{T\in\mathcal{T}_h} \left(\int_T \nabla u \cdot \nabla v_h \mathrm{d}x - \int_{\partial T} v_h \nabla u \cdot \mathbf{n}_T \mathrm{d}s \right).$$

For u = exact solution and $v_h \in V_h$ we have

$$\int_\Omega -\Delta u\, v_h \mathrm{d}x = \sum_{T\in\mathcal{T}_h} \left(\int_T \nabla u \cdot \nabla v_h \mathrm{d}x - \int_{\partial T} v_h \nabla u \cdot \mathbf{n}_T \mathrm{d}s \right)$$

that using (3.2) becomes

$$\begin{aligned} &= (\nabla u, \nabla v_h)_{\mathcal{T}_h} - < \{\!\{\nabla u\}\!\}, [\![v_h]\!] >_{\mathcal{E}_h} - < [\![\nabla u]\!]_s, \{\!\{v_h\}\!\} >_{\mathcal{E}_h^0} \\ &= (\nabla u, \nabla v_h)_{\mathcal{T}_h} - < \{\!\{\nabla u\}\!\}, [\![v_h]\!] >_{\mathcal{E}_h}. \quad \text{since } [\![\nabla u]\!] = 0 \end{aligned}$$

4.1. The Three Main Variants. We recall that if u is the solution of problem (4.1), then for every piecewise polynomial v_h we have

$$(-\Delta u\,,\, v_h)_{\mathcal{T}_h} = (\nabla u, \nabla v_h)_{\mathcal{T}_h} - < \{\!\{\nabla u\}\!\}, [\![v_h]\!] >_{\mathcal{E}_h}.$$

For $\delta = -1, 1, 0$ (three variants) and $\alpha_{\text{stab}} > 0$ we define the discrete problem as: *Find $u_h \in V_h$ such that*

$$\begin{aligned} (f, v_h)_{\mathcal{T}_h} = (\nabla u_h, \nabla v_h)_{\mathcal{T}_h} &- < \{\!\{\nabla u\}\!\}, [\![v_h]\!] >_{\mathcal{E}_h} \\ &+ \delta < \{\!\{\nabla v_h\}\!\}, [\![u_h]\!] >_{\mathcal{E}_h} + \frac{\alpha_{\text{stab}}}{h} < [\![u_h]\!], [\![v_h]\!] >_{\mathcal{E}_h}. \end{aligned}$$

We point out that the terms in the last line are zero for $u_h = u$.

We now set

$$\begin{aligned} a_\delta(u_h, v_h) := (\nabla u_h, \nabla v_h)_{\mathcal{T}_h} &- < \{\!\{\nabla u_h\}\!\}, [\![v_h]\!] >_{\mathcal{E}_h} \\ &+ \delta < \{\!\{\nabla v_h\}\!\}, [\![u_h]\!] >_{\mathcal{E}_h} + \frac{\alpha_{\text{stab}}}{h} < [\![u_h]\!], [\![v_h]\!] >_{\mathcal{E}_h} \end{aligned} \tag{4.2}$$

so that the discrete problem becomes

$$a_\delta(u_h, v_h) = (f, v_h)_{\mathcal{T}_h} \qquad \forall v_h \in V_h.$$

We have now to check consistency and stability of all the variants, in order to prove optimal error bounds

4.2. Consistency. We note first that, for every δ and for every $\alpha_{\rm stab}$, when u is the exact solution we have, for all $v_h \in V_h$:

$$a_\delta(u, v_h) = (\nabla u, \nabla v_h)_{\mathcal{T}_h} - < \{\!\{\nabla u\}\!\}, [\![v_h]\!] >_{\mathcal{E}_h} = (f, v_h).$$

Hence, if u_h solves $a_\delta(u_h, v_h) = (f, v_h)_{\mathcal{T}_h}$ for all $v_h \in V_h$ we have the *Galerkin Orthogonality*

$$a_\delta(u - u_h, v_h) = 0 \qquad \forall v_h \in V_h. \tag{4.3}$$

Recalling that, for piecewise smooth u and v,

$$\begin{aligned} a_\delta(u, v) := &(\nabla u, \nabla v)_{\mathcal{T}_h} - < \{\!\{\nabla u\}\!\}, [\![v]\!] >_{\mathcal{E}_h} \\ &+ \delta < \{\!\{\nabla v\}\!\}, [\![u]\!] >_{\mathcal{E}_h} + \frac{\alpha_{\rm stab}}{h} < [\![u]\!], [\![v]\!] >_{\mathcal{E}_h}, \end{aligned}$$

and using the definition of the *jump-norm* together with (3.8) and (3.11) we gather easily that for all piecewise smooth u and v we have

$$a_\delta(u, v) \le C \, \|u\|_{\rm DG} \, \|v\|_{\rm DG}$$

with a constant C independent of the mesh-size.

4.3. Stability. We first recall that, in the subspace V_h, from the inverse inequality (2.8)

$$\|v_h\|^2_{\rm DG} = \|v_h\|^2_j + \|\nabla v_h\|^2_{0,h} + h^2 \|\nabla v_h\|_{1,h} \simeq \|v_h\|^2_j + \|\nabla v_h\|^2_{0,h}.$$

Therefore, from the definition (4.2) we have:

$$a_\delta(v_h, v_h) := |v_h|^2_{1,h} + \alpha_{\rm stab} \|v_h\|^2_j + (\delta - 1) < \{\!\{\nabla v_h\}\!\}, [\![v_h]\!] >_{\mathcal{E}_h}$$

so that

$$a_\delta(v_h, v_h) \ge |v_h|^2_{1,h} + \alpha_{\rm stab} \|v_h\|^2_j - |\delta - 1| C |v_h|_{1,h} \, \|v_h\|_j \tag{4.4}$$

with a constant C independent of the mesh size. At this point it is convenient to recall that, given a quadratic form $x^2 + \alpha_s y^2 - 2\beta xy$, the associated matrix

$$\begin{pmatrix} 1 & -\beta \\ -\beta & \alpha_s \end{pmatrix}$$

is positive definite if and only if $\alpha_s > \beta^2$. In other words, for β fixed, we will always have

$$x^2 + \alpha_s y^2 - \beta xy \ge \alpha^* (x^2 + y^2)$$

for some constant $\alpha^* > 0$, whenever α_s is *big enough*. Going back to (4.4) we deduce that, for every δ,

$$\begin{aligned} a_\delta(v_h, v_h) &\ge |v_h|^2_{1,h} + \alpha_{\rm stab} \|v_h\|^2_j - |\delta - 1| C |v_h|_{1,h} \, \|v_h\|_j \\ &\ge \alpha^* \|v_h\|^2_{\rm DG} \quad \forall v_h \in V_h \end{aligned} \tag{4.5}$$

for some constant $\alpha^* > 0$, whenever $\alpha_{\rm stab}$ is big enough.

4.4. The Corresponding Methods and Some Variants. At this point we recall that we had three choices for δ, namely $\delta = -1, 1, 0$, in the discrete bilinear form

$$a_\delta(u_h, v_h) := (\nabla u_h, \nabla v_h)_{\mathcal{T}_h} - < \{\!\{\nabla u_h\}\!\}, [\![v_h]\!] >_{\mathcal{E}_h} + \delta < \{\!\{\nabla v_h\}\!\}, [\![u_h]\!] >_{\mathcal{E}_h} + \frac{\alpha_{\text{stab}}}{h} < [\![u_h]\!], [\![v_h]\!] >_{\mathcal{E}_h} .$$

We can now comment that for all the three methods we have consistency (actually: Galerkin orthogonality) and stability (in the subspace) with a constant independent of the mesh size. We can further comment that, in particular

- For $\delta = -1$ (SIPG, [3, 38]) we have a *symmetric method*
- For $\delta = 1$ (NIPG, [10, 34]) we have stability for all $\alpha_{\text{stab}} > 0$
- For $\delta = 0$ (IIPG, [24, 37]) we have a simpler expression.

We can also consider other variants. Always for $\delta = -1, 1, 0$ we denote by Π^e_{r-1} the $L^2(e)$ projection onto the polynomials of degree $\leq r-1$ on e. We consider the following variants

$$a_\delta(u_h, v_h) := (\nabla u_h, \nabla v_h)_{\mathcal{T}_h} - < \{\!\{\nabla u_h\}\!\}, [\![v_h]\!] >_{\mathcal{E}_h} + \delta < \{\!\{\nabla v_h\}\!\}, [\![u_h]\!] >_{\mathcal{E}_h} + \frac{\alpha_{\text{stab}}}{h} < \Pi^e_{r-1}[\![u_h]\!], \Pi^e_{r-1}[\![v_h]\!] >_{\mathcal{E}_h} .$$

For $r = 1$, these variants are denoted, respectively, SIPG-0, NIPG-0, and IIPG-0. In particular, IIPG-0 has several nice features that allow an easier construction of solvers and/or pre-conditioners [8, 9].

Other variants include the possibility of adding, on top of the stabilizing term

$$\alpha_{\text{stab}} h^{-1} < [\![u_h]\!], [\![v_h]\!] >_{\mathcal{E}_h}, \tag{4.6}$$

an additional stabilizing term of the type

$$\beta_{\text{stab}} h < [\![\nabla u_h]\!]_s, [\![\nabla v_h]\!]_s >_{\mathcal{E}_h}$$

(see, e.g., [16]).

Finally we point out that, for $\delta = 1$, and piecewise linear elements ($r = 1$), we can eliminate the jump-penalty term (4.6) and obtain stability by inserting a bubble function into the local space [1, 2, 15, 17].

4.5. Convergence. Let u_I be an approximation of u in V_h. Setting $\delta_h := u_h - u_I$ we have

$$\begin{aligned} \alpha^* \|\delta_h\|^2_{\text{DG}} &\leq a_\delta(\delta_h, \delta_h) && \text{(use the definition of } \delta_h) \\ &= a_\delta(u_h - u_I, \delta_h) && \text{(use (4.3))} \\ &= a_\delta(u - u_I, \delta_h) && \text{(use (3.8))} \\ &\leq M \, \|u - u_I\|_{\text{DG}} \, \|\delta_h\|_{\text{DG}} \end{aligned} \tag{4.7}$$

so that

$$\|u-u_h\|_{\mathrm{DG}} \le \|u-u_I\|_{\mathrm{DG}} + \|\delta_h\|_{\mathrm{DG}} \le \left(1+\frac{M}{\alpha_*}\right)\|u-u_I\|_{\mathrm{DG}}. \tag{4.8}$$

4.6. Approximation. We assume that u_I is an approximation of u in V_h with the following property: There exists an integer r (the degree of the local polynomials) and a constant C such that

$$|u_I-u|_{s,K} \le Ch^{r+1-s}|u|_{r+1,K} \tag{4.9}$$

for all integers s with $0 \le s \le r$, for all h and for all element $K \in \mathcal{T}_h$.

Using (4.9) we bound first the jump norm:

$$\begin{aligned}\|u_I-u\|_j^2 &= \sum_{e\in\mathcal{E}_h}\frac{1}{h}\|[\![u_I-u]\!]\|_{0,e}^2 \le 2\sum_{K\in\mathcal{T}_h}\sum_{e\in\partial K}\frac{1}{h}\|u_I-u\|_{0,e}^2\\ &\le C\sum_{K\in\mathcal{T}_h}\left(h^{-2}\|u_I-u\|_{0,K}^2+|u_I-u|_{1,K}^2\right) \le C\,h^{2r}|u|_{r+1,K}^2.\end{aligned}$$

Now, always using (4.9) we bound the second part of the DG norm:

$$\begin{aligned}&\sum_{K\in\mathcal{T}_h}\left(|\nabla(u_I-u)|_{0,K}^2+h^2\,|\nabla(u_I-u)|_{1,K}^2\right)\\ &\quad\le \sum_{K\in\mathcal{T}_h}\left(|u_I-u|_{1,K}^2+h^2\,|u_I-u|_{2,K}^2\right) \le C\,h^{2r}|u|_{r+1,K}^2.\end{aligned}$$

We conclude that under the assumption (4.9) we have

$$\begin{aligned}\|u_I-u\|_{\mathrm{DG}}^2 &= \|[\![u_I-u]\!]\|_j^2+\sum_{K\in\mathcal{T}_h}\left(|\nabla(u_I-u)|_{0,K}^2+h^2|\nabla(u_I-u)|_{1,K}^2\right)\\ &\le C\,h^{2r}|u|_{r+1,K}^2.\end{aligned}$$

5. Linear Elasticity.

5.1. The Problem. Given a domain Ω and a distributed load $\mathbf{f}$, we define

$$A_\mu\boldsymbol{u} := -\mathbf{div}\boldsymbol{\varepsilon}(\boldsymbol{u}) \qquad A_\lambda := -\nabla\mathrm{div}\boldsymbol{u} \qquad \mathcal{A} := 2\mu A_\mu + \lambda A_\lambda$$

where μ and λ are the *Lamé* coefficients, depending on the material, and $\boldsymbol{\varepsilon}(\boldsymbol{v}) := (1/2)(\nabla\boldsymbol{v}+(\nabla\boldsymbol{v})^T)$ is the usual symmetric gradient. Then we look for $\boldsymbol{u}$ such that

$$\mathcal{A}\boldsymbol{u} = \mathbf{f} \text{ in } \Omega \quad \text{ and } \quad \boldsymbol{u} = 0 \text{ on } \partial\Omega \tag{5.1}$$

The bilinear forms associated with the operators A_μ, A_λ, and $\mathcal{A}$ are given by

$$a_\mu(\boldsymbol{u},\boldsymbol{v}) := \int_\Omega \boldsymbol{\varepsilon}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{v})\mathrm{d}x, \tag{5.2}$$

$$a_\lambda(\boldsymbol{u},\boldsymbol{v}) := \int_\Omega \mathrm{div}\boldsymbol{u}\,\mathrm{div}\boldsymbol{v}\mathrm{d}x, \tag{5.3}$$

and

$$a(\boldsymbol{u},\boldsymbol{v}) := 2\mu a_\mu(\boldsymbol{u},\boldsymbol{v}) + \lambda a_\lambda(\boldsymbol{u},\boldsymbol{v}), \tag{5.4}$$

respectively. Hence, setting $\mathbf{V} := (H_0^1(\Omega))^2$, the variational formulation of (5.1) reads: *find* $\boldsymbol{u} \in$ *such that*

$$a(\boldsymbol{u},\boldsymbol{v}) := 2\mu a_\mu(\boldsymbol{u},\boldsymbol{v}) + \lambda a_\lambda(\boldsymbol{u},\boldsymbol{v}) = (\mathbf{f},\boldsymbol{v}) \quad \forall \boldsymbol{v} \in \mathbf{V}. \tag{5.5}$$

REMARK 5.1. *We point out that from the variational formulation* (5.5)*, taking as usual* $\boldsymbol{v} = \boldsymbol{u}$ *and using the Korn inequality*

$$C_{\text{Korn}}\|\boldsymbol{v}\|_{\mathbf{V}}^2 \le a_\mu(\boldsymbol{v},\boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \mathbf{V} \tag{5.6}$$

we easily have

$$2\mu\, C_{\text{Korn}}\|\boldsymbol{u}\|_{\mathbf{V}}^2 + \lambda\|\text{div}\boldsymbol{u}\|_{0,\Omega}^2 \le 2\mu a_\mu(\boldsymbol{u},\boldsymbol{u}) + \lambda\|\text{div}\boldsymbol{u}\|_{0,\Omega}^2 = (\mathbf{f},\boldsymbol{u}) \tag{5.7}$$

and therefore

$$\sqrt{\mu}\|\boldsymbol{u}\|_{\mathbf{V}} + \sqrt{\lambda}\|\text{div}\boldsymbol{u}\|_{0,\Omega} \le C\,\|\mathbf{f}\|_{|\mathbf{V}'} \tag{5.8}$$

with a constant C *independent of* μ *and* λ*. On the other hand, we also have easily*

$$2\mu a_\mu(\boldsymbol{u},\boldsymbol{v}) + \lambda a_\lambda(\boldsymbol{u},\boldsymbol{v}) \le C\,(2\mu+\lambda)\|\boldsymbol{u}\|_{\mathbf{V}}\,\|\boldsymbol{v}\|_{\mathbf{V}} \tag{5.9}$$

with a constant C *independent of* μ *and* λ*.*

5.2. Discretization. Assume now that we have again a decomposition $\mathcal{T}_h$ of Ω into elements K. On every element K we set

$$a_K(\boldsymbol{u},\boldsymbol{v}) = 2\mu \int_K \boldsymbol{\varepsilon}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{v})\mathrm{d}x + \lambda \int_K \text{div}\boldsymbol{u}\,\text{div}\boldsymbol{v}\mathrm{d}x.$$

and we recall the Green formula:

$$\begin{aligned} a_K(\boldsymbol{u},\boldsymbol{v}) = &-2\mu \int_K (A_\mu \boldsymbol{u})\cdot\boldsymbol{v}\mathrm{d}x - \lambda \int_K (\nabla\text{div}\boldsymbol{u})\cdot\boldsymbol{v}\mathrm{d}x \\ &+ \int_{\partial K} \left(2\mu \mathbf{M}^\mu_{\mathbf{n}_K}(\boldsymbol{u}) + \lambda \mathbf{M}^\lambda_{\mathbf{n}_K}(\boldsymbol{u})\right)\cdot\boldsymbol{v}\mathrm{d}s \end{aligned}$$

where $M^\mu_{\mathbf{n}_K}(\boldsymbol{u}) := \boldsymbol{\varepsilon}(\boldsymbol{u})\cdot\mathbf{n}_K$ and $M^\lambda_{\mathbf{n}_K}(\boldsymbol{u}) := (\text{div}\boldsymbol{u})\mathbf{n}_K$. We also recall that the stress field $\boldsymbol{\sigma}$ is given by

$$\boldsymbol{\sigma} := 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda\text{div}\boldsymbol{u}\mathbb{I}, \qquad (\text{in short: } \boldsymbol{\sigma} = \mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}))$$

where $\mathbb{I}$ is the *identity matrix*. We rewrite

$$a_K(\boldsymbol{u},\boldsymbol{v}) = (\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}),\boldsymbol{\varepsilon}(\boldsymbol{v}))_K, \quad \mathbf{M}_{\mathbf{n}_K}(\boldsymbol{u}) = (2\mu\mathbf{M}^\mu_{\mathbf{n}_K} + \lambda\mathbf{M}^\lambda_{\mathbf{n}_K})(\boldsymbol{u}).$$

The Green formula can then be written as

$$(\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}),\boldsymbol{\varepsilon}(\boldsymbol{v}))_K = (\mathcal{A}\boldsymbol{u},\boldsymbol{v})_K + < \mathbf{M}_{\mathbf{n}_K}(\boldsymbol{u}),\boldsymbol{v} >_{\partial K} .$$

We now introduce, in the spirit of the previous sections, the space $\mathbf{V}_h$ of piecewise polynomial (possibly discontinuous) vectors, concentrating our attention, for simplicity, on the piecewise linear case. For $\boldsymbol{u}$ and $\boldsymbol{v}$ piecewise smooth, summing over K and then applying the correspondent (for this case) of the "magic trick," we have

$$\sum_K (\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}),\boldsymbol{\varepsilon}(\boldsymbol{v}))_K = (\mathcal{A}\boldsymbol{u},\boldsymbol{v})_{\mathcal{T}_h} + < \{\!\{\mathbf{M}_{\mathbf{n}_K}(\boldsymbol{u})\}\!\}, [\![\boldsymbol{v}]\!]\cdot\mathbf{n} >_{\mathcal{E}_h} + < [\![\mathbf{M}_{\mathbf{n}_K}(\boldsymbol{u})]\!]\cdot\mathbf{n}, \{\!\{\boldsymbol{v}\}\!\} >_{\mathcal{E}_h}.$$

When $\boldsymbol{u}$ is the exact solution and $\boldsymbol{v} = \boldsymbol{v}_h$ is an element of $\mathbf{V}_h$ we obviously have $[\![\mathbf{M}_{\mathbf{n}_K}(\boldsymbol{u})]\!]\cdot\mathbf{n} = 0$. Hence

$$(\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}),\boldsymbol{\varepsilon}(\boldsymbol{v}_h))_{\mathcal{T}_h} - < \{\!\{\mathbf{M}_{\mathbf{n}_K}(\boldsymbol{u})\}\!\}, [\![\boldsymbol{v}_h]\!]\cdot\mathbf{n} >_{\mathcal{E}_h} = (\mathbf{f},\boldsymbol{v}_h)_{\mathcal{T}_h}. \tag{5.10}$$

5.3. The Discretized Problem. As before, from (5.10) we take inspiration in order to write the discretized problem. Taking again into account that the regularity of the exact solution implies $[\![\mathbf{M}_{\mathbf{n}_K}(\boldsymbol{u})]\!]\cdot\mathbf{n} = 0$ as well as $[\![\boldsymbol{u}]\!] = 0$, we introduce the bilinear form

$$B_h(\boldsymbol{u},\boldsymbol{v}) := (\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}),\boldsymbol{\varepsilon}(\boldsymbol{v}))_{\mathcal{T}_h} - < \{\!\{\mathbf{M}_{\mathbf{n}_K}(\boldsymbol{u})\}\!\}, [\![\boldsymbol{v}]\!]\cdot\mathbf{n} >_{\mathcal{E}_h} + \delta < \{\!\{\mathbf{M}_{\mathbf{n}_K}(\boldsymbol{v})\}\!\}, [\![\boldsymbol{u}]\!]\cdot\mathbf{n} >_{\mathcal{E}_h} + \frac{\alpha_{\rm stab}}{h} < [\![\boldsymbol{u}]\!], [\![\boldsymbol{v}]\!] >_{\mathcal{E}_h}, \tag{5.11}$$

where again we can take $\delta = -1, 1, 0$ (three methods) and $\alpha_{\rm stab} > 0$ is a stabilization parameter. We consider then the discretized problem

$$\text{Find } \boldsymbol{u}_h \in \mathbf{V}_h \text{ such that:} \quad B_h(\boldsymbol{u}_h,\boldsymbol{v}_h) = (\mathbf{f},\boldsymbol{v}_h)_{\mathcal{T}_h} \quad \forall \boldsymbol{v}_h \in \mathbf{V}_h. \tag{5.12}$$

It is immediate to see, from (5.10) and (5.12), that *Galerkin orthogonality* holds:

$$B_h(\boldsymbol{u}-\boldsymbol{u}_h,\boldsymbol{v}_h) = 0 \qquad \forall \boldsymbol{v}_h \in \mathbf{V}_h. \tag{5.13}$$

Moreover, defining, as in (3.6),

$$\|\mathbf{v}\|^2_{\rm DG} := |\boldsymbol{v}|^2_{1,h} + \sum_K h_K^2 |\boldsymbol{v}|^2_{2,K} + \sum_e \frac{1}{h_e}\|[\![\boldsymbol{v}]\!]\|^2_{0,e},$$

we have, with arguments quite similar to the ones of the previous section and using the DG version of (5.6) (see [12]), that for $\alpha_{\rm stab}$ big enough we have *stability*:

$$\exists \kappa_s > 0 \text{ such that } \kappa_s \mu \|\boldsymbol{v}_h\|^2_{\rm DG} \le B_h(\boldsymbol{v}_h,\boldsymbol{v}_h) \qquad \forall \boldsymbol{v}_h \in \mathbf{V}_h, \tag{5.14}$$

with a κ_s independent of μ, λ, and h. Similarly, using (5.9) in every element and following again the same arguments used for Poisson problem in the previous section, we can also prove *continuity*

$$\exists M > 0 \text{ s. t. } B_h(\boldsymbol{u}, \boldsymbol{v}) \leq M\,(\mu + \lambda)\,\|\boldsymbol{u}\|_{\mathrm{DG}}\|\boldsymbol{v}\|_{\mathrm{DG}} \quad \forall \boldsymbol{u},\, \boldsymbol{v} \in H^2(\mathcal{T}_h), \tag{5.15}$$

with an M independent of μ, λ, and h.

5.4. The Nearly Incompressible Case. As we saw, for every λ and μ positive we have stability [see (5.14)] and continuity [see (5.15)]. However, for $\lambda >> \mu$ we have a mismatch between the stability and the continuity constant.

Let us see the effects of this on the classical error estimate. Let $\boldsymbol{u}_I$ be an approximation of the solution $\boldsymbol{u}$ in $\mathbf{V}_h$. Setting $\boldsymbol{\eta}_h := \boldsymbol{u}_h - \boldsymbol{u}_I$ we have, as in (4.7) and (4.8),

$$\begin{aligned} \kappa_s \mu \|\boldsymbol{\eta}_h\|^2_{\mathrm{DG}} &\leq B_h(\boldsymbol{\eta}_h, \boldsymbol{\eta}_h) = B_h(\boldsymbol{u}_h - \boldsymbol{u}_I, \boldsymbol{\eta}_h) \\ &= B_h(\boldsymbol{u} - \boldsymbol{u}_I, \boldsymbol{\eta}_h) \leq M(\mu + \lambda)\,\|\boldsymbol{u} - \boldsymbol{u}_I\|_{\mathrm{DG}}\,\|\boldsymbol{\eta}_h\|_{\mathrm{DG}} \end{aligned} \tag{5.16}$$

so that

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\mathrm{DG}} \leq \|\boldsymbol{u} - \boldsymbol{u}_I\|_{\mathrm{DG}} + \|\boldsymbol{\eta}_h\|_{\mathrm{DG}} \leq \left(1 + \frac{M\,(\mu + \lambda)}{\mu \kappa_s}\right) \|\boldsymbol{u} - \boldsymbol{u}_I\|_{\mathrm{DG}}$$

and for $\lambda >> \mu$ we are in deep trouble. Actually, if instead of DG methods we were using traditional H^1-conforming methods we would face the so-called *locking phenomenon*, and the solution $\boldsymbol{u}_h$ of our discretized problem would be bounded, but would not converge to the exact solution $\boldsymbol{u}$.

With DG methods, instead, we have good results: let us see why. As a first step, we recall the so-called *inf-sup* condition for the continuous problem (5.5)

$$\exists\, \beta > 0 \text{ such that } \quad \inf_{q \in Q} \sup_{\boldsymbol{v} \in \mathbf{V}} \frac{(\mathrm{div}\boldsymbol{v}, q)}{\|q\|_Q\, \|\boldsymbol{v}\|_{\mathbf{V}}} \geq \beta > 0, \tag{5.17}$$

where $Q := L^2(\Omega)/\mathbb{R}$ is the subspace of $L^2(\Omega)$ made of functions with zero mean value.

5.5. Solving Troubles with DG . We can now start proving error bounds for the discretized problem (5.12). We restart as in (5.16), setting now $\boldsymbol{\delta}_h := \boldsymbol{u}_h - \boldsymbol{u}_I$, and stop at

$$\kappa_s \mu \|\boldsymbol{\delta}_h\|^2_{\mathrm{DG}} \leq B_h(\boldsymbol{\delta}_h, \boldsymbol{\delta}_h) = B_h(\boldsymbol{u}_h - \boldsymbol{u}_I, \boldsymbol{\delta}_h) = B_h(\boldsymbol{u} - \boldsymbol{u}_I, \boldsymbol{\delta}_h). \tag{5.18}$$

Instead of bounding brutally the last term, we now observe that

$$\begin{aligned} B_h(\boldsymbol{u} - \boldsymbol{u}_I, \boldsymbol{\delta}_h) \leq\ & 2\mu\, C \|\boldsymbol{u} - \boldsymbol{u}_I\|_{\mathrm{DG}}\, \|\boldsymbol{\delta}_h\|_{\mathrm{DG}} + \tfrac{\alpha_{\mathrm{stab}}}{h}\, |<[\![\boldsymbol{u} - \boldsymbol{u}_I]\!]_s, [\![\boldsymbol{\delta}_h]\!]_s>_{\mathcal{E}_h}| \\ & + \lambda\, |(\mathrm{div}(\boldsymbol{u} - \boldsymbol{u})_I, \mathrm{div}\boldsymbol{\delta}_h)_{\mathcal{T}_h} + <\{\mathrm{div}(\boldsymbol{u} - \boldsymbol{u}_I)\}, [\![\boldsymbol{\delta}_h]\!] \cdot \mathbf{n}>_{\mathcal{E}_h} \\ & + \delta\, <\{\mathrm{div}\boldsymbol{\delta}_h\}, [\![\boldsymbol{u} - \boldsymbol{u}_I]\!] \cdot \mathbf{n}>_{\mathcal{E}_h}|\,. \end{aligned} \tag{5.19}$$

The only way to bound this with a constant that does not depend on λ would be to find a $\boldsymbol{u}_I$, in the subspace $\mathbf{V}_h$, such that:

$$\textbf{(a)} \qquad \int_K \mathrm{div}(\boldsymbol{u}-\boldsymbol{u}_I)\mathrm{d}x = 0 \qquad \forall K;$$

$$\textbf{(b)} \qquad \int_e (\boldsymbol{u}-\boldsymbol{u}_I)\cdot\mathbf{n}\mathrm{d}s = 0 \qquad \forall e;$$

$$\textbf{(c)} \qquad \|\boldsymbol{u}-\boldsymbol{u}_I\|_{r,K} \le C\,h^{2-r}\|\boldsymbol{u}\|_2 \qquad \forall K, \quad r=0,1.$$

Property **(a)** would cancel the term $(\mathrm{div}(\boldsymbol{u}-\boldsymbol{u}_I), \mathrm{div}\boldsymbol{\delta}_h)_{\mathcal{T}_h}$, since $\mathrm{div}\boldsymbol{\delta}_h$ (for our piecewise linear elements) is constant in each element. Moreover, (since $\mathrm{div}\boldsymbol{u}_I$ is piecewise constant) it will also imply

$$\|\lambda\mathrm{div}(\boldsymbol{u}-\boldsymbol{u}_I)\|_{0,K} \le C\,h_K\|\lambda\mathrm{div}\boldsymbol{u}\|_{1,K}$$

on every element K.

Property **(b)** would cancel the term $<\{\mathrm{div}\boldsymbol{\delta}_h\}, [\![\boldsymbol{u}-\boldsymbol{u}_I]\!]\cdot\mathbf{n}>_{\mathcal{E}_h}$ since, again, $\mathrm{div}\boldsymbol{\delta}_h$ is constant in each element (and therefore its trace is constant on each edge).

Property **(c)** takes care of the the jump terms. Indeed, combined with (3.13) it will provide

$$\|\boldsymbol{u}-\boldsymbol{u}_I\|_j^2 \le C\,(h^{-2}\|\boldsymbol{u}-\boldsymbol{u}_I\|_{0,\mathcal{O}}^2 + |\boldsymbol{u}-\boldsymbol{u}_I|_{1,h}^2) \le C\,h^2\|\boldsymbol{u}\|_{2,\mathcal{O}}^2.$$

Recalling the $H(\mathrm{div})$-conforming Finite Elements (as, for instance, the BDM_1 spaces [14]), we see that such a $\boldsymbol{u}_I$ can be easily constructed, and our work is concluded.

We note that the BDM_1 is not a subspace of $\mathbf{V}$, so that the above construction could not be used to prove convergence for traditional continuous Galerkin approximations.

REMARK 5.2. *The use of $\boldsymbol{u}_I$ in the above construction was instrumental to derive error bounds for fully discontinuous approximations. The idea, however, can be used to construct semi-discontinuous approximations, that is, with $\mathbf{V}_h \subset H(\mathrm{div})$ only, thus guaranteeing continuity of the normal component but not of the tangential component. This approach was used, for instance, in [23] for the Stokes problem.*

6. Alternative Formulations. In this section we recall some basic physical principles that are the basis for several numerical methods. For convenience and simplicity we restrict our attention to linear elasticity problems, although the range of applications (of the physical principles and of the related numerical methods) is much wider.

6.1. Minimum Potential Energy. The primal formulation of the linear elasticity problem (say, with homogeneous Dirichlet boundary conditions all over $\partial\Omega$) that we saw already in the previous section is based on the *minimum potential energy* principle:

$$\frac{1}{2}(\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}),\boldsymbol{\varepsilon}(\boldsymbol{v})) - (\mathbf{f},\mathbf{v}) = \text{ minimum}, \tag{6.1}$$

that is equivalent to our variational Eq. (5.5), that we repeat here for convenience of the reader

$$a(\boldsymbol{u},\boldsymbol{v}) \equiv (\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}),\boldsymbol{\varepsilon}(\boldsymbol{v})) = (\mathbf{f},\boldsymbol{v}) \quad \forall \boldsymbol{v} \in \mathbf{V} = (H_0^1(\Omega))^d.$$

6.2. Complementary Energy. We introduce the following notation

$$\boldsymbol{\Sigma} := (L^2(\Omega))^{d\times d}_{\rm sym}$$
$$\forall \mathbf{g} \quad \text{we set} \quad \boldsymbol{\Sigma}_{\mathbf{g}} := \{\boldsymbol{\tau} \in \boldsymbol{\Sigma} \ \text{ with } \ \mathbf{div}\,\boldsymbol{\tau} + \mathbf{g} = 0\}$$

that we are going to use mainly for $\mathbf{g} = \mathbf{f}$ or $\mathbf{g} = \mathbf{0}$. The dual formulation of elasticity problems is based on the *complementary energy* principle:

$$\frac{1}{2}(\mathbb{C}^{-1}\boldsymbol{\sigma},\boldsymbol{\sigma}) = \ \text{minimum over } \boldsymbol{\Sigma}_{\mathbf{f}}, \tag{6.2}$$

giving rise to the variational equation

$$\boldsymbol{\sigma} \in \boldsymbol{\Sigma}_{\mathbf{f}} \quad \text{and} \quad (\mathbb{C}^{-1}\boldsymbol{\sigma},\boldsymbol{\tau}) = 0 \quad \forall \boldsymbol{\tau} \in \boldsymbol{\Sigma}_{\mathbf{0}}.$$

6.3. The Hellinger–Reissner Principle. The Hellinger–Reissner principle is at the basis of the two more common mixed formulations. The principle reads:

$$\frac{1}{2}(\mathbb{C}^{-1}\boldsymbol{\sigma},\boldsymbol{\sigma}) - (\boldsymbol{\varepsilon}(\boldsymbol{u}),\boldsymbol{\sigma}) + (\mathbf{f},\boldsymbol{u}) = \ \text{stationary}. \tag{6.3}$$

The (Euler–Lagrange) equations of (6.3) are:

$$\begin{cases} (\mathbb{C}^{-1}\boldsymbol{\sigma},\boldsymbol{\tau}) - (\boldsymbol{\varepsilon}(\boldsymbol{u}),\boldsymbol{\tau}) = 0 \quad \forall \boldsymbol{\tau} \in \boldsymbol{\Sigma} \\ (\boldsymbol{\varepsilon}(\boldsymbol{v}),\boldsymbol{\sigma}) = (\mathbf{f},\boldsymbol{v}) \quad \forall \boldsymbol{v} \in \mathbf{V} \end{cases} \tag{6.4}$$

This is the *primal mixed* formulation for elasticity.

On the other hand, the Euler–Lagrange equations (6.4) become, upon integration by parts,

$$\begin{cases} (\mathbb{C}^{-1}\boldsymbol{\sigma},\boldsymbol{\tau}) + (\boldsymbol{u},\mathbf{div}\,\boldsymbol{\tau}) \quad = 0 \quad \forall \boldsymbol{\tau} \in \boldsymbol{\Sigma} \text{ with } \mathbf{div}\boldsymbol{\tau} \in (L^2(\Omega))^d \\ (\boldsymbol{v},\mathbf{div}\,\boldsymbol{\sigma}) \quad = -(\mathbf{f},\boldsymbol{v}) \quad \forall \boldsymbol{v} \in (L^2(\Omega))^d \end{cases} \tag{6.5}$$

This is the *dual mixed* formulation for elasticity.

6.4. Discontinuous Approximations. In the discretization of (6.3) one clearly chooses either $-(\boldsymbol{\varepsilon}(\boldsymbol{u}),\boldsymbol{\tau})$ or $(\boldsymbol{u},\mathbf{div}\,\boldsymbol{\tau})$ depending on whether one takes *continuous displacements* or *continuous (normal) stresses*, and this, as we have seen, corresponds to using *primal mixed* or *dual mixed* methods, respectively.

Clearly, if **both** displacements ($\boldsymbol{v}$) **and** stresses ($\boldsymbol{\tau}$) are approximated by **discontinuous** piecewise polynomials, the two above terms are no longer equal. Indeed one has

$$(\boldsymbol{\varepsilon}(\boldsymbol{v}), \boldsymbol{\tau})_h + (\boldsymbol{v}, \mathbf{div}\, \boldsymbol{\tau})_h = < \{\!\{\boldsymbol{\tau}\}\!\}, [\![\boldsymbol{v}]\!] > + < [\![\boldsymbol{\tau}]\!], \{\!\{\boldsymbol{v}\}\!\} > \tag{6.6}$$

in the best tradition of DG methods. Here, and in what follows, for functions v, w piecewise smooth, $(v, w)_h$ will indicate the scalar product:

$$(v, w)_h = \sum_{K \in \mathcal{T}_h} (v, w)_{0,K}.$$

6.5. Towards Hybrid Methods. Formula (6.6) opens the door towards Hybrid methods. Assume that your discretization allows you to know the displacements only at the interelement boundaries (to fix the ideas, because the displacement trial and test functions are defined, inside each element, to be the solution of some PDE). On the other hand, in this case you can reasonably take them (that is, the displacements) to be *single valued* on the skeleton, so that $[\![\boldsymbol{v}]\!] = 0$ in (6.6). Then, if $\mathbf{div}_h\, \boldsymbol{\tau} = 0$ in each element K, (6.6) becomes

$$(\boldsymbol{\varepsilon}(\boldsymbol{v}), \boldsymbol{\tau})_h = < [\![\boldsymbol{\tau}]\!], \{\!\{\boldsymbol{v}\}\!\} >, \tag{6.7}$$

so that in (6.4) you can write $< [\![\boldsymbol{\tau}]\!], \{\!\{\boldsymbol{v}\}\!\} >$ instead of $(\boldsymbol{\varepsilon}(\boldsymbol{v}), \boldsymbol{\tau})_h$. Similarly, when $\mathbf{div}_h\, \boldsymbol{\sigma} + \mathbf{f} = 0$ Eq. (6.6) gives

$$(\boldsymbol{\varepsilon}(\boldsymbol{v}), \boldsymbol{\sigma})_h - (\mathbf{f}, \boldsymbol{v}) = < [\![\boldsymbol{\sigma}]\!], \{\!\{\boldsymbol{v}\}\!\} >, \tag{6.8}$$

that can be used in the second equation of (6.4).

Using both (6.7) and (6.8) (always for $\mathbf{div}_h \boldsymbol{\tau} = 0$ and $\mathbf{div}_h \boldsymbol{\sigma} + \mathbf{f} = 0$, respectively) in (6.4), we have then

$$\begin{cases} (\mathbb{C}^{-1}\boldsymbol{\sigma}, \boldsymbol{\tau}) - < [\![\boldsymbol{\tau}]\!], \{\!\{\boldsymbol{u}\}\!\} >= 0 \quad \forall\, \boldsymbol{\tau} \in \boldsymbol{\Sigma_0} \\ < [\![\boldsymbol{\sigma}]\!], \{\!\{\boldsymbol{v}\}\!\} >= 0 \quad \forall\, \boldsymbol{v} \in \mathbf{V} \end{cases} \tag{6.9}$$

6.6. Dual Hybrid Methods. The general strategy for constructing a dual hybrid method is as follows.

Pick up a *particular solution* $\boldsymbol{\sigma}_f$ (such that $\mathbf{div}_h \boldsymbol{\sigma} + \mathbf{f} = 0$), and write $\boldsymbol{\sigma} = \boldsymbol{\sigma}_f + \boldsymbol{\sigma}_0$ with $\boldsymbol{\sigma}_0$ to be found. Then look for $\boldsymbol{\sigma}_0$ and $\boldsymbol{u}$ such that

$$\begin{cases} (\mathbb{C}^{-1}(\boldsymbol{\sigma}_0 + \boldsymbol{\sigma}_f), \boldsymbol{\tau}_0) - < [\![\boldsymbol{\tau}_0]\!], \{\!\{\boldsymbol{u}\}\!\} > = 0 \quad \forall\, \boldsymbol{\tau}_0 \in \boldsymbol{\Sigma}_0 \\ < [\![\boldsymbol{\sigma}_0 + \boldsymbol{\sigma}_f]\!], \{\!\{\boldsymbol{v}\}\!\} > = 0 \quad \forall\, \boldsymbol{v} \in \mathbf{V}. \end{cases} \tag{6.10}$$

Note that the values of $\boldsymbol{u}$ and $\boldsymbol{v}$ are used *only at the interelement boundaries*. Separating $\boldsymbol{\sigma}_f$, and considering $\boldsymbol{\sigma}_0$ as the *true stress unknown*, we have then the final formulation: *Find* $\boldsymbol{\sigma}_0 \in \boldsymbol{\Sigma_0}$ *and* $\boldsymbol{u}$ *on the skeleton such that*

$$\begin{cases} (\mathbb{C}^{-1}\boldsymbol{\sigma}_0, \boldsymbol{\tau}_0) - < [\![\boldsymbol{\tau}_0]\!], \{\!\{\boldsymbol{u}\}\!\} > = -(\mathbb{C}^{-1}\boldsymbol{\sigma}_f, \boldsymbol{\tau}_0) \quad \forall\, \boldsymbol{\tau}_0 \in \boldsymbol{\Sigma_0} \\ < [\![\boldsymbol{\sigma}_0]\!], \{\!\{\boldsymbol{v}\}\!\} > = - < [\![\boldsymbol{\sigma}_f]\!], \{\!\{\boldsymbol{v}\}\!\} > \quad \forall\, \boldsymbol{v} \in \mathbf{V}. \end{cases} \tag{6.11}$$

Note: When you discretize (6.11) you will need *sufficiently many* $\boldsymbol{\tau}_0$ to control $\{\!\{\boldsymbol{u}\}\!\}$....

6.7. Primal Hybrid. Assume now that, in the primal formulation (6.1), we start with *discontinuous* $\boldsymbol{u}$ and $\boldsymbol{v}$. One possibility to do this would be to proceed as in the previous section. Another possibility, however, is to consider that we are actually dealing with a minimization problem, and to consider the interelement continuity (here $[\![\boldsymbol{u}]\!] = 0$) as a *constraint*. Then we could introduce a Lagrange multiplier (that will turn out to be the normal component of the stress field $\boldsymbol{\sigma}$ at the interelement boundaries), obtaining the two equations

$$\begin{cases} (\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{v}))_h + < \{\!\{\boldsymbol{\sigma}\}\!\}, [\![\boldsymbol{v}]\!] > = (\mathbf{f}, \boldsymbol{v}) \quad \forall\, \boldsymbol{v} \\ < \{\!\{\boldsymbol{\tau}\}\!\}, [\![\boldsymbol{u}]\!] >= 0 \quad \forall\, \boldsymbol{\tau}. \end{cases} \tag{6.12}$$

where $\boldsymbol{u}$ and $\boldsymbol{v}$ are (a priori) discontinuous from one element to the other while $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ are defined *only at the interelement boundaries*.

Equation (6.12) are the basis for the **Primal Hybrid Methods**. Note that, in this case, you will need *sufficiently many* $\boldsymbol{v}$'s to control $\{\!\{\boldsymbol{\sigma}\}\!\}$.

6.8. Nonconforming Methods. In discretizing (6.12) you can restrict yourself to consider displacement fields $\boldsymbol{u}$ and $\boldsymbol{v}$ in some subspace (of discontinuous p.w. polynomials) $\mathbf{V}_h$. To fix the ideas, assume that the elements of $\mathbf{V}_h$ are, piecewise, polynomials of degree k for some $k \geq 1$. In a similar way, you will assume that you have, at the interelement boundaries, a space $\boldsymbol{\Sigma}_h$ to discretize the normal components of the stress field, made of piecewise (actually: *egdewise*) polynomials of degree m. We can assume, for the sake of simplicity, that $m < k$ (otherwise, in general, the *inf-sup* condition would fail, since you will not have *sufficiently many* $\boldsymbol{v}$'s to control $\{\!\{\boldsymbol{\sigma}\}\!\}$). At this point you might restrict your attention to displacements that belong to the space $V_{\rm nc}$ defined by

$$V_{\rm nc} := \{\boldsymbol{v} \in \mathbf{V}_h \ \text{ such that } \ < \{\!\{\boldsymbol{\tau}\}\!\}, [\![\boldsymbol{v}]\!] >= 0 \quad \forall\, \boldsymbol{\tau} \in \boldsymbol{\Sigma}_h\}.$$

Then you will just look for $\boldsymbol{u} \in V_{\rm nc}$ such that

$$(\mathbb{C}\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{v}))_h = (\mathbf{f}, \boldsymbol{v}) \quad \forall\, \boldsymbol{v} \in V_{\rm nc}.$$

This could obviously be seen as using a *Nonconforming Finite Element Method*.

In a quite similar way you could instead start from (6.11), and introduce a discretized space $\boldsymbol{\Sigma}_h$ (made of piecewise polynomial symmetric tensors) and a discretized space $\mathbf{V}$ made of edgewise polynomial vectors on the skeleton. Then you could think of using an $H(\mathbf{div})$-nonconforming space of the form

$$\boldsymbol{\Sigma}_{\rm nc} := \{\boldsymbol{\tau} \in \boldsymbol{\Sigma}_h \ \text{ such that } \ < \{\!\{\boldsymbol{v}\}\!\}, [\![\boldsymbol{\tau}]\!] >= 0 \quad \forall\, \boldsymbol{v} \in \mathbf{V}_h\}.$$

6.9. Hybridizing Dual Mixed Methods. Let us go back to the Hellinger–Reissner principle for *dual mixed* elements (6.5), that uses "continuous stresses" (i.e., "$H(\mathbf{div})$-conforming") and discontinuous displacements. Assume now that you want to use, a priori, *discontinuous stresses* $\boldsymbol{\sigma}$, and enforce back their continuity by means of a Lagrange multiplier.

Then you will consider spaces $\mathbf{V}_h$ and $\boldsymbol{\Sigma}_h$ made of discontinuous piecewise polynomials, and a space of edgewise polynomials $\mathbb{M}_h$, and look for $\boldsymbol{\sigma} \in \boldsymbol{\Sigma}$, $\boldsymbol{u} \in \mathbf{V}_h$, and $\mathcal{U} \in \mathbb{M}_h$ such that

$$(\mathbb{C}^{-1}\boldsymbol{\sigma}, \boldsymbol{\tau}) + (\boldsymbol{u}, \mathbf{div}\, \boldsymbol{\tau})_h - < \{\!\{\mathcal{U}\}\!\}, [\![\boldsymbol{\tau}]\!] >= 0 \quad \forall\, \boldsymbol{\tau} \in \boldsymbol{\Sigma}_h \tag{6.13}$$

$$-(\boldsymbol{v}, \mathbf{div}\, \boldsymbol{\sigma})_h = (\mathbf{f}, \boldsymbol{v}) \quad \forall\, \boldsymbol{v} \in \mathbf{V}_h \tag{6.14}$$

$$< \{\!\{\mathcal{V}\}\!\}, [\![\boldsymbol{\sigma}]\!] >= 0 \quad \forall\, \mathcal{V} \in \mathbb{M} \tag{6.15}$$

Noting that in (6.13)–(6.15) $\mathbf{V}_h$ and $\boldsymbol{\Sigma}_h$ are "bubbles-spaces" (meaning that you can easily have a basis made of vectors and tensors (respectively) having support in a single element), we can eliminate $\boldsymbol{\sigma}$ and $\boldsymbol{u}$ by *static condensation*, and end up with a system of the type

$$\Lambda(\{\!\{\mathcal{U}\}\!\}, \{\!\{\mathcal{V}\}\!\}) =< F, \{\!\{\mathcal{V}\}\!\} > \quad \forall\, \mathcal{V}$$

whose matrix is, in general, symmetric and positive definite. Remember, however, that you will still need some sort of *inf-sup* condition. Indeed, recalling the hybridized formulation (6.13)–(6.15), if you are interested only in the $\mathcal{U}$ variable (eliminating the others by static condensation), you cannot avoid an *inf-sup* condition: you need *sufficiently many* $\boldsymbol{\tau}$*'s to control* $\{\!\{\mathcal{U}\}\!\}$ (that appears only in the first Eq. (6.13)).

This procedure, originally introduced by Fraeijs de Veubeke [26], has been first analyzed for Poisson problem in [4] and is used in a rather systematic way when dealing with mixed finite element methods for scalar elliptic problems. Apart from the paramount advantage of going back to a single elliptic problem, the procedure has many additional advantages:

- The Lagrange multiplier $\mathcal{U}$ is a good approximation of $\boldsymbol{u}$ at the interfaces. You can postprocess $\mathcal{U}$ and get an approximation of $\boldsymbol{u}$ *one order better* than the original one coming from the mixed formulation (see, e.g., [4]).
- In many cases, $\mathcal{U}$ can be computed directly using suitable *nonconforming* discretizations of the *primal formulation* (see e.g. [33])
- In many problems, $\mathcal{U}$ can be identified with the *flux variable* of Finite Volumes and DG Methods, with many interesting features to be exploited (see, e.g., [20, 21]).

However, the application to linear elasticity problems is less spectacular, since the combined need to work with symmetric stress fields, to have an *inf-sup* condition and to have $H(\mathrm{div})$ compatibility is a considerable source of troubles. See, for instance, [6, 11, 22, 27, 29] for some recent attempts using reduced symmetry (and the references therein for earlier attempts), and see as well [7] and [28] for an attempt to use nonconforming elements.

REFERENCES

[1] P.F. Antonietti, F. Brezzi and L.D. Marini *Bubble stabilization of discontinuous Galerkin methods,* Comput. Methods Appl. Mech. Engrg. **198** (2009), pp. 1651–1659.

[2] P.F. Antonietti, F. Brezzi and L.D. Marini *Stabilizations of the Baumann-Oden DG formulation: the 3D case,* Boll. Unione Mat. Ital. **9** (2008), pp. 629–643.

[3] D.N. Arnold *An interior penalty finite element method with discontinuous elements,* SIAM J. Numer. Anal. **19** (1982), pp. 742–760.

[4] D.N. Arnold and F. Brezzi *Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates,* RAIRO Modél. Math. Anal. Numér. **19** (1985), pp. 7–32.

[5] D.N. Arnold, F. Brezzi, B. Cockburn, and L.D. Marini*Unified analysis of discontinuous Galerkin methods for elliptic problems,* SIAM J. Numer. Anal. **39** (2001/02), pp. 1749–1779.

[6] D.N. Arnold, R. Falk, and R. Winther *Mixed finite element methods for linear elasticity with weakly imposed symmetry,* Math. Comput **76** (2007), pp. 1699–1723.

[7] D.N. Arnold and R. Winther *Nonconforming mixed elements for elasticity,* Math. Models Methods Appl. Sci. **13** (2003), pp. 295–307.

[8] B. Ayuso de Dios, F. Brezzi, O. Havle, and L.D. Marini*L2-estimates for the DG IIPG-0 scheme,* Numer. Methods Partial Differential Equations **28** (2012), pp. 1440–1465.

[9] B. Ayuso de Dios and L. Zikatanov *Uniformly convergent iterative methods for discontinuous Galerkin discretizations,* J. Sci. Comput. **40** (2009), pp. 4–36.

[10] C.E. Baumann and J.T. Oden *A discontinuous hp finite element method for convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg. **175** (1999), pp. 311–341.

[11] D. Boffi, F. Brezzi, M. Fortin *Reduced symmetry elements in linear elasticity* , Commun. Pure Appl. Anal. **8** (2009), pp. 95–121.

[12] S.C. Brenner *Korn's inequalities for piecewise H^1 vector fields,* Math. Comp. **73** (2004), pp. 1067–1087.

[13] S.C. Brenner, L.R. Scott, *The mathematical theory of finite element methods* , Texts in Applied Mathematics, 15. Springer-Verlag, New York, 1994.

[14] F. Brezzi, J. Douglas, Jr., and L.D. Marini *Two families of mixed finite elements for second order elliptic problems,* Numer. Math. **47** (1985), pp. 217–235.

[15] F. Brezzi and L.D. Marini *Bubble stabilization of discontinuous Galerkin methods,* in Advances in numerical mathematics, (W. Fitzgibbon, R. Hoppe, J. Periaux, O. Pironneau, Y. Vassilevski, Eds) Institute of Numerical Mathematics of the Russian Academy of Sciences, Moscow (2006), pp. 25–36.

[16] E. Burman, P. Hansbo*A stabilized non-conforming finite element method for incompressible flow,* Comput. Methods Appl. Mech. Engrg. **195** (2006), pp. 2881–2899.

[17] E. Burman and B. Stamm*Symmetric and non-symmetric discontinuous Galerkin methods stabilized using bubble enrichment,* C.R. Math. Acad. Sci. Paris **346** (2008), pp. 103–106.

[18] P.G. Ciarlet, *The finite element method for elliptic problems*, North-Holland, 1978.

[19] B. Cockburn *Discontinuous Galerkin methods,*ZAMM Z. Angew. Math. Mech. **83** (2003), 731754.

[20] B. Cockburn and J. Gopalakrishnan *A characterization of hybridized mixed methods for second order elliptic problems,* SIAM J. Numer. Anal. **42** (2004), pp. 283–301.

[21] B. Cockburn, J. Gopalakrishnan and R. Lazarov *Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems,* SIAM J. Numer. Anal. **47** (2009), pp. 1319–1365.

[22] B. Cockburn, J. Gopalakrishnan and J. Guzmán *A new elasticity element made for enforcing weak stress symmetry,* Math. Comput. **79** (2010), pp. 1331–1349.

[23] B. Cockburn, G. Kanschat and D. Shötzau *A note on discontinuous Galerkin divergence-free solutions of the Navier-Stokes equations,* J. Sci. Comput. **31** (2007), pp. 61–73.

[24] C. Dawson, S. Sun, and M.F. Wheeler *Compatible algorithms for coupled flow and transport*, Comput. Methods Appl. Mech. Engrg. **193** (2004), pp. 2565–2580.

[25] X. Feng and O.A. Karakashian A. Two-level additive Schwarz methods for a discontinuous Galerkin approximation of second order elliptic problems, SIAM J. Numer. Anal. **39** (2001), pp. 1343–1365.

[26] B.X. Fraeijis de Veubeke *Displacement and equilibrium models in the finite element method,* in Stress Analysis, O.C. Zienkiewicz and G. Holister Eds., Wiley (1965).

[27] R. Falk *Finite elements for linear elasticity,* in Mixed Finite Elements: Compatibility Conditions Stress Analysis, Lecture Notes in Math. v. 1939, D. Boffi and L. Gastaldi Eds., Springer, Heidelberg (2008).

[28] J. Gopalakrishnan and J. Guzmán *Symmetric nonconforming mixed finite elements for linear elasticity,* SIAM J. Numer. Anal. **49** (2011), pp. 1504–1520.

[29] J. Guzmán *A unified analysis of several mixed methods for elasticity with weak stress symmetry,* J. Sci. Comput. **44** (2010) 156–169.

[30] O.A. Karakashian and F. Pascal *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems,* SIAM J. Numer. Anal. **41** (2003), pp. 2374–2399.

[31] O.A. Karakashian and F. Pascal *Convergence of adaptive discontinuous Galerkin approximations of second-order elliptic problems*, SIAM J. Numer. Anal. **45** (2007), pp. 641–665.

[32] C. Lovadina and L.D. Marini *A-Posteriori Error Estimates for Discontinuous Galerkin Approximations of Second Order Elliptic Problems,* J. Sci. Comput. **40** (2009), pp. 340–359.

[33] L.D. Marini *An inexpensive method for the evaluation of the solution of the lowest order Raviart-Thomas mixed method,* SIAM J. Numer. Anal. **22** (1985), pp. 493–496.

[34] B. Rivière, M.F. Wheeler, and V. Girault *A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems,*SIAM J. Numer. Anal. **39** (2001), pp. 902–931.

[35] C.W. Shu *Discontinuous Galerkin methods: general approach and stability,* Numerical solutions of partial differential equations, 149201, Adv. Courses Math. CRM Barcelona, Birkhuser, Basel, 2009

[36] C.W. Shu *Paper in this book*

[37] S. Sun and M.F. Wheeler *Symmetric and nonsymmetric discontinuous Galerkin methods for reactive transport in porous media,* SIAM J. Numer. Anal. **43** (2005), pp. 195–219.

[38] M.F. Wheeler *An elliptic collocation-finite element method with interior penalties,* SIAM J. Numer. Anal. **15** (1978), pp. 152–161.

DISCONTINUOUS GALERKIN METHOD FOR TIME-DEPENDENT PROBLEMS: SURVEY AND RECENT DEVELOPMENTS

CHI-WANG SHU*

Abstract. In these lectures we give a general survey on discontinuous Galerkin methods for solving time-dependent partial differential equations. We also present a few recent developments on the design, analysis, and application of these discontinuous Galerkin methods.

Key words. Discontinuous Galerkin method, Time-dependent partial differential equations, Superconvergence, Positivity-preserving, δ-functions.

AMS(MOS) subject classifications. Primary 65M60, 65M20, 65M12, 65M15.

1. Introduction. Discontinuous Galerkin (DG) methods belong to the class of finite element methods. The finite element function space corresponding to DG methods consists of piecewise polynomials (or other simple functions) which are allowed to be completely discontinuous across element interfaces. Therefore, using finite element terminologies, DG methods are the most extreme case of nonconforming finite element methods.

The first DG method was introduced in 1973 by Reed and Hill in a Los Alamos technical report [72]. It solves the equations for neutron transport, which are time independent linear hyperbolic equations. A major development of the DG method is carried out by Cockburn et al. in a series of papers [23, 25, 27–29], in which the authors have established a framework to easily solve nonlinear time-dependent hyperbolic equations, such as the Euler equations of compressible gas dynamics. The DG method of Cockburn et al. belongs to the class of method-of-lines, namely the DG discretization is used only for the spatial variables, and explicit, nonlinearly stable high order Runge–Kutta methods [81] are used to discretize the time variable. Other important features of the DG method of Cockburn et al. include the usage of exact or approximate Riemann solvers as interface fluxes and total variation bounded (TVB) nonlinear limiters [79] to achieve non-oscillatory properties for strong shocks, both of which are borrowed from the methodology of high resolution finite volume schemes.

The DG method has found rapid applications in such diverse areas as aeroacoustics, electro-magnetism, gas dynamics, granular flows, magnetohydrodynamics, meteorology, modeling of shallow water, oceanography, oil recovery simulation, semiconductor device simulation, transport of contaminant in porous media, turbomachinery, turbulent flows, viscoelastic

*Division of Applied Mathematics, Brown University, Providence, RI 02912, USA, shu@dam.brown.edu

X. Feng et al. (eds.), *Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations*, The IMA Volumes in Mathematics and its Applications 157, DOI 10.1007/978-3-319-01818-8_2,

flows and weather forecasting, among many others. For earlier work on DG methods, we refer to the survey paper [24], and other papers in that Springer volume, which contains the conference proceedings of the First International Symposium on Discontinuous Galerkin Methods held at Newport, Rhode Island in 1999. The lecture notes [21] is a good reference for many details, as well as the extensive review paper [31]. The review paper [99] covers the local DG method for partial differential equations (PDEs) containing higher order spatial derivatives. More recently, there are three special journal issues devoted to the DG method [32, 33, 35], which contain many interesting papers on DG method in all aspects including algorithm design, analysis, implementation, and applications. There are also a few recent books and lecture notes [38, 54, 59, 75, 80] on DG methods.

2. DG Methods for Hyperbolic Conservation Laws. As we mentioned in the previous section, the first DG method [72] was designed to solve linear hyperbolic equations in neutron transport. Let us use the following simple example to demonstrate the idea of this method. We consider a one-dimensional linear steady state hyperbolic equation

$$u_x = f, \quad x \in [0,1]; \qquad u(0) = g. \tag{2.1}$$

First, we divide [0,1] into N cells

$$0 = x_{\frac{1}{2}} < x_{\frac{3}{2}} < \cdots < x_{N+\frac{1}{2}} = 1,$$

and denote

$$I_j = \left(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}\right), \quad x_j = \frac{1}{2}\left(x_{j-\frac{1}{2}} + x_{j+\frac{1}{2}}\right), \quad h_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$$

as the cells, cell centers and cell lengths, respectively. We also define $h = h_{\max} = \max_j h_j$ and $h_{\min} = \min_j h_j$, and we consider only regular meshes, that is $h_{\max} \le \lambda h_{\min}$ where $\lambda \ge 1$ is a constant during mesh refinement. If $\lambda = 1$, then the mesh is uniformly distributed. Define the discontinuous Galerkin finite element space as

$$V_h^k = \{v : v|_{I_j} \in \mathcal{P}^k(I_j), j = 1, \cdots, N\}, \tag{2.2}$$

where $\mathcal{P}^k(I_j)$ denotes the space of polynomials in I_j of degree at most k. This polynomial degree k can actually change from cell to cell, but we assume it is a constant in these lectures for simplicity. The DG scheme for solving (2.1) is: find $u_h \in V_h^k$, such that for any $v_h \in V_h^k$ and all $1 \le j \le N$,

$$\begin{aligned} &-\int_{I_j} u_h (v_h)_x dx + (u_h)^-_{j+\frac{1}{2}} (v_h)^-_{j+\frac{1}{2}} \\ &-(u_h)^-_{j-\frac{1}{2}} (v_h)^+_{j-\frac{1}{2}} = \int_{I_j} f v_h dx. \end{aligned} \tag{2.3}$$

Here we define $(u_h)^-_{\frac{1}{2}} = g$ using the given boundary condition in (2.1). If a local basis of $P^k(I_j)$ is chosen and denoted as $\varphi_j^\ell(x)$ for $\ell = 0, 1, \cdots, k$,

we can express the numerical solution as

$$u_h(x) = \sum_{\ell=0}^{k} u_j^\ell \varphi_j^\ell(x), \qquad x \in I_j,$$

and we should solve for the coefficients

$$u_j = \begin{pmatrix} u_j^0 \\ \vdots \\ u_j^k \end{pmatrix}, \tag{2.4}$$

which, according to the scheme (2.3), satisfy the linear equation

$$A_j u_j = b_j \tag{2.5}$$

where A_j is a $(k+1) \times (k+1)$ matrix whose (ℓ, m)th entry is given by

$$a_j^{\ell,m} = -\int_{I_j} \varphi_j^m(x)(\varphi_j^\ell(x))_x dx + \varphi_j^m(x_{j+\frac{1}{2}})\varphi_j^\ell(x_{j+\frac{1}{2}}) \tag{2.6}$$

and the ℓth entry of the right-hand-side vector b_j is given by

$$b_j^\ell = u_h(x^-_{j-\frac{1}{2}})\varphi_j^\ell(x_{j-\frac{1}{2}}) + \int_{I_j} f(x)\varphi_j^\ell(x)dx,$$

which depends on the information of u_h in the left cell I_{j-1}, if it is in the computational domain, or on the boundary condition, if it is outside the computational domain (i.e., when $j = 1$). It is easy to verify that the matrix A_j in (2.5) with entries given by (2.6) is invertible, hence the numerical solution u_h in the cell I_j can be easily obtained by solving the small linear system (2.5), once the solution at the left cell I_{j-1} is already known, or if the left cell is outside the computational domain. Therefore, we can obtain the numerical solution u_h in the following ordering: first we obtain it in the cell I_1, since its left boundary is equipped with the prescribed boundary condition in (2.1). We then obtain the solution in the cell I_2, as the numerical solution u_h in its left cell I_1 is already available. This process can be repeated sequentially to obtain solutions in I_j with $j = 3, 4, \ldots$ until we obtain the solution u_h for all cells in the computational domain.

Notice that this method does not involve any large linear system solvers and is very easy to implement. The first order version ($k = 0$) is a well-known upwind finite difference scheme; however, it is more difficult to generalize the same scheme for higher order finite difference schemes which involve a wide stencil. On the other hand, this DG scheme can be designed for any polynomial degree k, and it is easy to be generalized to two and higher spatial dimensions. In [56], Lesaint and Raviart proved that this DG method is convergent with the optimal order of accuracy, namely

$O(h^{k+1})$, in the L^2 norm, when piecewise tensor product polynomials of degree k are used as basis functions in multi-dimensions. Numerical experiments indicate that the convergence rate is also optimal when the usual piecewise polynomials of degree k are used instead in multi-dimensions.

Notice that, even though the method (2.3) is designed for the steady state problem (2.1), it can be easily used on initial-boundary value problems of linear time-dependent hyperbolic equations: we just need to identify the time variable t as one of the spatial variables. Also, this DG method can be easily designed and efficiently implemented on arbitrary triangulations. L^2 error estimates of $O(h^{k+1/2})$ where k is again the polynomial degree and h is the mesh size can be obtained when the solution is sufficiently smooth, for arbitrary meshes, see, e.g., [53]. This estimate is actually sharp for the most general situation [67]; however, in many cases the optimal $O(h^{k+1})$ error bound can be proved [22, 74]. In actual numerical computations, one almost always observe the optimal $O(h^{k+1})$ accuracy.

Unfortunately, even though the method (2.3) is easy to implement, accurate, and efficient, it cannot be easily generalized to linear systems, where the characteristic information comes from different directions, or to nonlinear problems, where the characteristic wind direction depends on the solution itself. This difficulty can be overcome when the DG discretization is only used for the spatial variables, and the time discretization is achieved by the explicit Runge–Kutta methods. This is the approach of the so-called Runge–Kutta discontinuous Galerkin (RKDG) method [23, 25, 27–29]. We demonstrate the RKDG method with the one-dimensional conservation law

$$u_t + f(u)_x = 0. \tag{2.7}$$

The semi-discrete DG method for solving (2.7) is defined as follows: find the unique function $u_h = u_h(t) \in V_h^k$ such that, for all test functions $v_h \in V_h^k$ and all $1 \le j \le N$, we have

$$\begin{aligned} \int_{I_j} (u_h)_t \, v_h dx - \int_{I_j} f(u_h)(v_h)_x dx \\ + \hat{f}_{j+\frac{1}{2}}(v_h)^-_{j+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}}(v_h)^+_{j-\frac{1}{2}} = 0. \end{aligned} \tag{2.8}$$

Here, $\hat{f}_{i+\frac{1}{2}}$ is the numerical flux, which is a single-valued function defined at the cell interfaces and in general depends on the values of the numerical solution u_h from both sides of the interface

$$\hat{f}_{i+\frac{1}{2}} = \hat{f}(u_h(x^-_{i+\frac{1}{2}}, t), u_h(x^+_{i+\frac{1}{2}}, t)).$$

We use the so-called monotone fluxes from finite difference and finite volume schemes for solving conservation laws, which satisfy the following conditions:

- Consistency: $\hat{f}(u, u) = f(u)$;
- Continuity: $\hat{f}(u^-, u^+)$ is at least Lipschitz continuous with respect to both arguments u^- and u^+.

- Monotonicity: $\hat{f}(u^-, u^+)$ is a non-decreasing function of its first argument u^- and a non-increasing function of its second argument u^+. Symbolically $\hat{f}(\uparrow, \downarrow)$.

We refer to, e.g., [57] for more details about monotone fluxes.

The semi-discrete version of this DG scheme can again be written in the compact form

$$\frac{d}{dt}u_j = A(u_{j-1}) + B(u_j) + C(u_{j+1})$$

where the vectors u_j are defined in (2.4), and A, B, C are vector functions. If the conservation law is linear, then A, B, C are linear operators, namely

$$A(u_{j-1}) = Au_{j-1}; \qquad B(u_j) = Bu_j; \qquad C(u_{j+1}) = Cu_{j+1}$$

with constant matrices A, B, and C (scaled by the local mesh sizes). This makes the implementation of the RKDG method local and highly efficient. It also makes the method easy for parallel implementation. The method can achieve almost 100 % parallel efficiency for static meshes and over 80 % parallel efficiency for dynamic load balancing with adaptive meshes [9, 73].

As a finite element method, the DG method can be designed in multi-dimensions on arbitrary triangulations (even those with hanging nodes) in the same fashion as in the one-dimensional case. It is easy to handle complicated geometry and boundary conditions.

2.1. Stability. It is well known that weak solutions of (2.7) may not be unique and the unique, physically relevant weak solution (the so-called entropy solution) satisfies the following entropy inequality

$$U(u)_t + F(u)_x \leq 0 \tag{2.9}$$

in distribution sense, for any convex entropy $U(u)$ satisfying $U''(u) \geq 0$ and the corresponding entropy flux $F(u) = \int^u U'(u)f'(u)du$. It will be nice if a numerical approximation to (2.7) also shares a similar entropy inequality as (2.9). It is usually quite difficult to prove a discrete entropy inequality for finite difference or finite volume schemes, especially for high order schemes and when the flux function $f(u)$ in (2.7) is not convex or concave. However, it turns out that it is easy to prove that the solution u_h to the semi-discrete DG scheme (2.8) satisfies a cell entropy inequality [51]:

$$\frac{d}{dt}\int_{I_j} U(u_h)\,dx + \hat{F}_{j+\frac{1}{2}} - \hat{F}_{j-\frac{1}{2}} \leq 0 \tag{2.10}$$

for the square entropy $U(u) = \frac{u^2}{2}$, with a consistent entropy flux

$$\hat{F}_{i+\frac{1}{2}} = \hat{F}\left(u_h\left(x^-_{i+\frac{1}{2}}, t\right), u_h\left(x^+_{i+\frac{1}{2}}, t\right)\right)$$

satisfying $\hat{F}(u, u) = F(u)$.

An easy corollary of the cell entropy inequality is the following L^2 stability. For periodic or compactly supported boundary conditions for the computational domain $[a, b]$, the solution u_h to the semi-discrete DG scheme (2.8) satisfies the following L^2 stability

$$\frac{d}{dt}\int_a^b (u_h)^2 dx \leq 0, \tag{2.11}$$

or

$$\|u_h(\cdot, t)\| \leq \|u_h(\cdot, 0)\|. \tag{2.12}$$

Here and below, an unmarked norm is the usual L^2 norm.

Notice that both the cell entropy inequality (2.10) and the L^2 stability (2.11) are valid even when the exact solution of the conservation law (2.7) is discontinuous. Both conclusions are valid in multi-dimensions on arbitrary triangulations [51], and for both scalar equations and symmetric hyperbolic systems [39]. They also hold true for fully discrete RKDG methods for linear conservation laws [110].

2.2. Error Estimates and Superconvergence. If we assume the exact solution of (2.7) is smooth, we can obtain optimal L^2 error estimates. Namely, the solution u_h of the DG scheme (2.8) for the PDE (2.7) with a smooth solution u, using the space of kth degree piecewise polynomials (2.2), satisfies the following error estimate

$$\|u - u_h\| \leq Ch^{k+1} \tag{2.13}$$

where C depends on u and its derivatives but is independent of h. Such error estimates can be obtained for the general nonlinear scalar conservation law (2.7) and symmetrizable hyperbolic systems, and for both semi-discrete DG methods and fully discretized RKDG methods, see [108–110]. The results also hold in multi-dimensions in tensor-product meshes and basis functions.

In recent years, there are a lot of efforts in the literature to obtain superconvergence results for DG methods solving hyperbolic conservation laws. These results consist of two categories.

The first category is to explore the superconvergence of the DG solution to the exact smooth solution in negative norms for linear hyperbolic equations:

$$\|u - u_h\|_{-k} \leq Ch^{2k+1} \tag{2.14}$$

where C depends on u and its derivatives but is independent of h [26]. Here $\|\cdot\|_{-k}$ is the negative Sobolev norm defined by

$$\|v\|_{-k} = \max_{\varphi \in H^k, \varphi \neq 0} \frac{(v, \varphi)}{\|\varphi\|_{H^k}}$$

where $(\cdot,\cdot)$ is the standard L^2 inner product and H^k is standard Sobolev space of order k. This result (and similar results for divided differences on uniform meshes), together with a local post-processing technique [12], allows us to obtain a post-processed solution $w_h = P(u_h)$ (where P is a local post-processing operator) on uniform meshes which is superconvergent in the strong L^2 norm:

$$\|u - w_h\| \leq Ch^{2k+1} \tag{2.15}$$

where C depends on u and its derivatives but is independent of h [26]. These results have been generalized to one-sided post-processing near the boundaries [76], structured triangular meshes [65], nonuniform meshes [34], and nonlinear problems [50]. It has also been applied to aeroacoustics [77] and computer graphics [82].

The second category is to explore the superconvergence of the DG solution to a special projection of the exact smooth solution, or superconvergence of the DG solution to the exact smooth solution at certain Gauss–Radau quadrature points.

The superconvergence of the DG solution to a special projection of the exact smooth solution takes the form

$$\|Pu - u_h\| \leq Ch^{k+1+\alpha} \tag{2.16}$$

where Pu is a projection (of the Gauss–Radau type) of the exact solution u into the finite element space V_h^k, and $\alpha > 0$ is the rate of superconvergence. In [18], Cheng and Shu started this line of study by obtaining (2.16) with $\alpha = 1/2$ for linear, time-dependent hyperbolic equations in one-dimension, with uniform meshes and periodic boundary conditions. The proof is based on Fourier analysis and is carried out only for the piecewise linear $k = 1$ case; however, numerical results confirm the validity for higher k's. Another important consequence of this superconvergence result is that the constant C in (2.16) only grows linearly with time t, therefore the standard error $\|u - u_h\|$ does not grow for a very long time $t \sim 1/\sqrt{h}$. This analysis verifies an observation by practitioners, that the error of the DG solution for wave propagation does not seem to grow much with time. The result in [18] is improved in [20] to general polynomial degree k, on nonuniform regular meshes, and without periodic boundary conditions. The technique used in [20] is a finite element type, not a Fourier analysis. In [105], the result in [20] is further improved to $\alpha = 1$. This half-order increase in the analysis is highly nontrivial and involves subtle handling of cancellation of errors during time evolution. The result in [105] is optimal. In [64], (2.16) with $\alpha = 1/2$ is proved for scalar nonlinear conservation laws with a fixed wind direction in one space dimension.

The superconvergence of the DG solution to the exact smooth solution at certain Gauss–Radau quadrature points has been explored in the literature. In [2, 3], Adjerid et al. proved the $(k+2)$th order superconvergence

of the DG solutions at the downwind-biased Radau points for ordinary differential equations. Later, Adjerid and Weihart [4, 5] investigated the local DG error for multi-dimensional first-order linear symmetric and symmetrizable hyperbolic systems of partial differential equations. The authors showed the projection of the local DG error is also $(k+2)$th order superconvergent at the downwind-biased Radau points by performing a local error analysis on Cartesian meshes. The global superconvergence is given by numerical experiments. In [4, 5], only initial-boundary value problems are considered, and the local DG error estimate is valid for t sufficiently large. Subsequently, Adjerid and Baccouch [1] investigated the global convergence of the implicit residual-based a posteriori error estimates and proved that these estimates at a fixed time t converge to the true spatial error in the L^2 norm under mesh refinement. In [117], using Fourier analysis, Zhong and Shu showed that the error between the DG numerical solution and the exact solution is $(k+2)$th order superconvergent at the downwind-biased Radau points and $(2k+1)$th order superconvergent at the downwind point in each cell on uniform meshes with periodic boundary conditions for $k = 1$, 2 and 3, for linear time-dependent hyperbolic equations in one dimension, with uniform meshes and periodic boundary conditions.

One of the applications of these superconvergence results is the design of asymptotically exact *a posteriori* error indicators, which are useful in adaptive computations.

2.3. Nonlinear Limiters. For computing solutions with strong discontinuities, the cell entropy inequality (2.10) and the L^2 stability (2.11), although helpful, are often not enough to control spurious numerical oscillations. In practice, especially for nonlinear problems containing strong discontinuities, we often need to apply nonlinear limiters to control these oscillations. Most of the limiters studied in the literature come from the methodologies of finite volume and finite difference high resolution schemes.

A limiter can be considered as a post-processor of the computed DG solution. In any cell which is deemed to contain a possible discontinuity (the so-called *troubled cells*), the DG polynomial is replaced by a new polynomial of the same degree, while maintaining the original cell average for conservation. Different limiters compute this new polynomial in different fashions. The main idea is to require that the new polynomial is less oscillatory than the old one, and, if the solution in this cell happens to be smooth, then the new polynomial should have the same high order accuracy as the old one. Some of the limiters are applied to all cells, while they should take effect (change the polynomial in the cell) only in the cells near the discontinuities. The total variation diminishing (TVD) limiters [37] belong to this class. Unfortunately, such limiters tend to take effect also in some cells in which the solution is smooth, for example in cells near smooth extrema of the exact solution. Accuracy is therefore lost in such cells. The TVB limiters [79], applied to RKDG schemes in [23, 25, 27, 29],

attempt to remove this difficulty and to ensure that the limiter takes effect only in cells near the discontinuities. The TVB limiters are widely used in applications, because of their simplicity in implementation. However, the TVB limiters involve a parameter M, related to the value of the second derivative of the exact solution near smooth extrema, which must be chosen by the user for different test cases. The moment-based limiter [9] and the improved moment limiter [13] also belong to this class, and they are specifically designed for DG methods and limit the moments of the polynomial sequentially, from the highest order moment downwards. Unfortunately, the moment-based limiters may also take effect in certain smooth cells, thereby destroying accuracy in these cells.

The limiters based on the weighted essentially non-oscillatory (WENO) methodology are designed with the objective of maintaining the high order accuracy even if they take effect in smooth cells. These limiters are based on the WENO methodology for finite volume and finite difference schemes [52, 62], and involve nonlinear reconstructions of the polynomials in troubled cells using the information of neighboring cells. The WENO reconstructed polynomials have the same high order of accuracy as the original polynomials when the solution is smooth, and they are (essentially) non-oscillatory near discontinuities. Qiu and Shu [70] and Zhu et al. [119] designed WENO limiters using the usual WENO reconstruction based on cell averages of neighboring cells as in [40, 52, 78], to reconstruct the values of the solutions at certain Gaussian quadrature points in the target cells, and then rebuild the solution polynomials from the original cell average and the reconstructed values at the Gaussian quadrature points through a numerical integration for the moments. This limiter needs to use the information from not only the immediate neighboring cells but also neighbors' neighbors, making it complicated to implement in multi-dimensions, especially for unstructured meshes [40, 116, 119]. The effort in [68, 71] attempts to construct Hermite type WENO approximations, which use the information of not only the cell averages but also the lower order moments such as slopes, to reduce the spread of reconstruction stencils. However for higher order methods the information of neighbors' neighbors is still needed. More recently, Zhong and Shu [118] developed a new WENO limiting procedure for RKDG methods on structured meshes. The idea is to reconstruct the entire polynomial, instead of reconstructing point values or moments in the classical WENO reconstructions. That is, the entire reconstruction polynomial on the target cell is a convex combination of polynomials on this cell and its immediate neighboring cells, with suitable adjustments for conservation and with the nonlinear weights of the convex combination following the classical WENO procedure. The main advantage of this limiter is its simplicity in implementation, as it uses only the information from immediate neighbors and the linear weights are always positive. This simplicity is more prominent for multi-dimensional unstructured meshes, which is studied in [120] for two-dimensional unstructured triangular meshes.

The WENO limiters are typically applied only in designated "troubled cells," in order to save computational cost and to minimize the influence of accuracy in smooth regions. Therefore, a troubled cell indicator is needed, to correctly identify cells near discontinuities in which the limiters should be applied. Qiu and Shu in [69] have compared several troubled cell indicators. In practice, the TVB indicator [79] and the KXRCF indicator [55] are often the best choices.

Finally, let us mention the recently developed positivity-preserving limiters for DG schemes [114]. These limiters involve only simple scaling of the polynomials and is very inexpensive to implement. They can guarantee maximum principle in the scalar case [111, 115] and positivity-preserving for certain systems, for example positivity-preserving for density and pressure for Euler equations of compressible gas dynamics [112, 113, 115] and positivity-preserving for water height for shallow water equations [88]. They are also proved to maintain the original high order accuracy of the DG scheme. It is worth mentioning that, in [83], the RKDG method with only the positivity-preserving limiter is used to compute the very demanding gaseous detonations in two-dimensional structured and unstructured meshes, with stable and high resolution results.

2.4. Hyperbolic Equations Involving δ-Functions. In a hyperbolic conservation law

$$\begin{aligned} &u_t + f(u)_x = g(x,t), && (x,t) \in R \times (0,T], \\ &u(x,0) = u_0(x), && x \in R, \end{aligned} \tag{2.17}$$

the initial condition u_0, or the source term $g(x,t)$, or the solution $u(x,t)$ may contain δ-singularities. Such problems appear often in applications and are difficult to approximate numerically. Many numerical techniques rely on modifications with smooth kernels and hence may severely smear such singularities, leading to large errors in the approximation. On the other hand, the DG methods are based on weak formulations and can be designed directly to solve such problems without modifications, leading to very accurate results.

In [106], DG methods to solve hyperbolic Eq. (2.17) involving δ-singularities are explored. Negative-order norm error estimates for the accuracy of DG approximations to δ-singularities are investigated. First, linear hyperbolic conservation laws in one space dimension with singular initial data are investigated. It is proved that, by using piecewise kth degree polynomials, at time t, the error in the $H^{-(k+2)}$ norm over the whole domain is $(k+1/2)$th order, and the error in the $H^{-(k+1)}(\mathbb{R}\backslash\mathcal{R}_t)$ norm is $(2k+1)$th order, where $\mathcal{R}_t$ is the pollution region due to the initial singularity with the width of order $\mathcal{O}(h^{1/2}\log(1/h))$ and h is the maximum cell length. As an application of the negative-order norm error estimates, the numerical solution can be convolved with a suitable kernel which is a linear combination of B-splines, to obtain L^2 error estimate of $(2k+1)$th

order for the post-processed solution. Second, high order superconvergence error estimates for linear hyperbolic conservation laws with singular source terms are obtained in [106] by applying Duhamel's principle. Numerical examples including an acoustic equation and the nonlinear rendezvous algorithms are given to demonstrate the good performance of DG methods for solving hyperbolic equations involving δ-singularities. The results in [106] give us evidence that the DG method is a good algorithm for problems involving δ-singularities in their solutions. In future work we will apply the DG method to more nonlinear hyperbolic equations involving δ-singularities.

2.5. Generalization to Hamilton–Jacobi Equations. Time-dependent Hamilton–Jacobi equations take the form

$$\varphi_t + H(\varphi_{x_1}, \ldots, \varphi_{x_d}) = 0, \qquad \varphi(x,0) = \varphi^0(x), \tag{2.18}$$

where H is a Lipschitz continuous function. H could also depend on φ, x, and t in some applications. Hamilton–Jacobi equations appear often in many applications. Examples include front propagation, level set methods, image processing and computer vision, control and differential games.

At least in the one-dimensional case, there is a strong relationship between the Hamilton–Jacobi equation

$$\varphi_t + H(\varphi_x) = 0, \qquad \varphi(x,0) = \varphi^0(x) \tag{2.19}$$

and the hyperbolic conservation law

$$u_t + H(u)_x = 0, \qquad u(x,0) = u^0(x). \tag{2.20}$$

In fact, if we identify $u = \varphi_x$, the two Eqs. (2.19) and (2.20) are equivalent. This equivalency provides motivations for designing algorithms for one equation based on the success for another equation. Therefore, it is very natural to attempt an adaptation of the DG methods designed for the conservation laws (2.20) to solve the Hamilton–Jacobi equation (2.19).

The first attempt to design a DG method was based exactly on this observation: at least in one dimension, the viscosity solution of the Hamilton–Jacobi equation (2.19) is equivalent to the entropy solution of the conservation law (2.20), when we identify $\varphi_x = u$. Therefore, a DG scheme for solving the conservation law (2.20), as given by (2.8) (with f there replaced by H), can be directly used to approximate the derivative of the viscosity solution of the Hamilton–Jacobi equation (2.19). This leads to the following DG algorithm of Hu and Shu [41]: Find $\varphi_h \in V_h^{k+1}$, such that $u_h = (\varphi_h)_x \in V_h^k$ is determined by the DG scheme (2.8) (with f there replaced by H), and the missing degree of freedom is determined by

$$\int_{I_j} ((\varphi_h)_t + H(u_h))\, dx = 0.$$

This algorithm is well defined for one dimension. Additional complications exist for multi-dimensional cases. We take two space dimensions as an example. The Hamilton–Jacobi equation

$$\varphi_t + H(\varphi_x, \varphi_y) = 0 \tag{2.21}$$

is (in some sense) equivalent to the following system of conservation laws

$$u_t + H(u, v)_x = 0, \qquad v_t + H(u, v)_y = 0 \tag{2.22}$$

when we identify $u = \varphi_x$ and $v = \varphi_y$. We would like to still use a piecewise polynomial φ_h as our solution variable and take its derivatives to approximate u and v. The DG algorithm of Hu and Shu [41], as re-interpreted by Li and Shu [43], can be formulated as follows: Find $\varphi_h \in V_h^{k+1}$, such that $(u_h, v_h) = ((\varphi_h)_x, (\varphi_h)_y) \in W_h^k$ is determined by the standard DG scheme solving the conservation laws (2.22), and the missing degree of freedom is determined by

$$\int_{I_j} ((\varphi_h)_t + H(u_h, v_h))\, dxdy = 0.$$

Here, I_j denotes two-dimensional elements (triangles or rectangles), and W_h^k is the locally curl-free subspace of $V_h^k \times V_h^k$:

$$W_h^k = \{(u, v) \in V_h^k \times V_h^k : u_y - v_x = 0 \quad \forall (x, y) \in I_j\}.$$

Some analysis for this DG method (including L^2 stability for a specific class of the Hamiltonian H) is given in [42]. A priori L^2 error estimates for smooth solutions are given in [89].

Even though the DG schemes in [41, 43] are successful in approximating the Hamilton–Jacobi equation (2.18), it involves rewriting it as a conservation law satisfied by the derivatives of the solution φ. It is desirable to design a DG method which solves directly the solution φ to the Hamilton–Jacobi equation (2.18). The scheme of Cheng and Shu [16] serves this purpose. The scheme is defined as: find $\varphi_h \in V_h^k$, such that

$$\begin{aligned} &\int_{I_j} ((\varphi_h)_t + H((\varphi_h)_x))\, v_h(x) dx \\ &\quad + \left(\min_{x \in I_{j+1/2}} H'((\tilde{\varphi}_h)_x) \right)_- [\varphi_h]_{j+\frac{1}{2}} (v_h)^-_{j+\frac{1}{2}} \\ &\quad + \left(\max_{x \in I_{j-1/2}} H'((\tilde{\varphi}_h)_x) \right)_+ [\varphi_h]_{j-\frac{1}{2}} (v_h)^+_{j-\frac{1}{2}} = 0 \end{aligned} \tag{2.23}$$

holds for any $v_h \in V_h^k$ and all $1 \le j \le N$. Here $a_- = \min(a, 0)$, $a_+ = \max(a, 0)$, $[w] = w^+ - w^-$ denotes the jump of w, and $H'(u)$ denotes the derivative of $H(u)$ with respect to u. The interval $I_{j+1/2} = [x_j, x_{j+1}]$,

and the function $\tilde{\varphi}_h$ is the L^2 projection of φ_h (which is discontinuous at the interface point $x_{j+\frac{1}{2}}$) into $\mathcal{P}^{2k+1}(I_j \cup I_{j+1})$. It can be easily verified that when the Hamiltonian $H(u) = au$ is linear and the Hamilton–Jacobi equation is also a conservation law with a possible source term (when a depends on x), the scheme (2.23) becomes the standard DG scheme for this conservation law. Extension of this method to multi-dimensions is straightforward. Further development and application of this method to problems in optimal control are given in [10, 11]. A priori L^2 error estimates for smooth solutions are given in [89].

Another DG method which solves directly the Hamilton–Jacobi equations (2.18) is that of Yan and Osher [102]. This method is motivated by the local discontinuous Galerkin (LDG) method for solving second order partial differential equations [30], to be described in next section. We refer the readers to [102] for more details of this method. A priori L^2 error estimates for smooth solutions are given in [89].

3. DG Methods for Convection-Diffusion Equations. DG methods are most suitable for convection equations; however, they are also good methods for solving convection dominated convection diffusion equations, such as Navier–Stokes equations with high Reynolds numbers. In this section we discuss the DG methods for time-dependent convection-diffusion equations

$$u_t + \sum_{i=1}^{d} f_i(u)_{x_i} - \sum_{i=1}^{d}\sum_{j=1}^{d} (a_{ij}(u)u_{x_j})_{x_i} = 0, \tag{3.1}$$

where $(a_{ij}(u))$ is a symmetric, semi-positive definite matrix.

For equations containing higher order spatial derivatives, such as the convection-diffusion equation (3.1), discontinuous Galerkin methods designed for hyperbolic conservation laws cannot be directly applied. Let us look at the heat equation as an example

$$u_t = u_{xx}. \tag{3.2}$$

Comparing with the hyperbolic conservation law (2.7), we can treat the heat equation (3.2) also as a "conservation law" by identifying $f(u)$ in (2.7) with $-u_x$ in (3.2). Therefore, it would appear that we could change the DG scheme (2.8), designed for solving the conservation law (2.7), to the following scheme for solving the heat equation (3.2). Find $u_h \in V_h^k$ such that, for all test functions $v_h \in V_h^k$ and all $1 \le j \le N$, we have

$$\begin{aligned} &\int_{I_j} (u_h)_t\, v_h dx + \int_{I_j} (u_h)_x (v_h)_x dx \\ &\qquad - \widehat{u_x}_{j+\frac{1}{2}} (v_h)^-_{j+\frac{1}{2}} + \widehat{u_x}_{j-\frac{1}{2}} (v_h)^+_{j-\frac{1}{2}} = 0, \end{aligned} \tag{3.3}$$

where $f(u)$ in (2.8) is replaced by $-u_x$ in (3.3). Of course, we still need to define the numerical fluxes $\widehat{u_x}_{j+\frac{1}{2}}$. Upwinding and monotone fluxes are no

longer relevant, as we are dealing with the heat equation (3.2) for which there is no preferred wind direction. It would appear that the average flux

$$\widehat{u_x}_{j+\frac{1}{2}} = \frac{1}{2}\left((u_h)_x)^-_{j+\frac{1}{2}} + ((u_h)_x)^+_{j+\frac{1}{2}}\right) \tag{3.4}$$

is a reasonable choice. If we take this flux in the scheme (3.3), we would observe numerically a very strange phenomenon. The numerical solution appears to be convergent when the mesh is refined; however, it does not seem to converge to the correct solution of the PDE. One would then suspect that the scheme is stable but inconsistent. However, the scheme can be written as a standard finite difference scheme and standard linear analysis for finite difference schemes can be performed. It turns out that the scheme (3.3) with the flux (3.4) is consistent but (very weakly) unstable [31, 107]. Therefore, we should be very careful when generalizing DG schemes from first order hyperbolic equations to PDEs with higher order spatial derivatives.

3.1. Local Discontinuous Galerkin Methods. One possible way to design a stable and convergent DG method for solving convection-diffu-sion equations is to rewrite the equation into a first order system, then apply the discontinuous Galerkin method on the system. A key ingredient for the success of such methods is the correct design of interface numerical fluxes. These fluxes must be designed to guarantee stability and local solvability of all the auxiliary variables introduced to approximate the derivatives of the solution. The local solvability of all the auxiliary variables is why the method is called a LDG method in [30].

The first local discontinuous Galerkin method was developed by Cockburn and Shu [30], for the convection-diffusion equation (3.1) containing second derivatives. Their work was motivated by the successful numerical experiments of Bassi and Rebay [7] for the compressible Navier–Stokes equations.

We will use the heat equation (3.2) to demonstrate the idea of LDG schemes. We rewrite Eq. (3.2) as the following system

$$u_t - q_x = 0, \qquad q - u_x = 0, \tag{3.5}$$

which "looks like" a system of conservation laws, except that the second equation does not have a time derivative in q. We can then formally write down the DG scheme (2.8) for each equation in (3.5), resulting in the following scheme. Find $u_h, q_h \in V_h^k$ such that, for all test functions $v_h, p_h \in V_h^k$ and all $1 \le j \le N$, we have

$$\begin{aligned}
&\int_{I_j} (u_h)_t\, v_h dx + \int_{I_j} q_h\, (v_h)_x dx - \hat{q}_{j+\frac{1}{2}} (v_h)^-_{j+\frac{1}{2}} + \hat{q}_{j-\frac{1}{2}} (v_h)^+_{j-\frac{1}{2}} = 0;\\
&\int_{I_j} q_h p_h dx + \int_{I_j} u_h\, (p_h)_x dx - \hat{u}_{j+\frac{1}{2}} (p_h)^-_{j+\frac{1}{2}} + \hat{u}_{j-\frac{1}{2}} (p_h)^+_{j-\frac{1}{2}} = 0.
\end{aligned} \tag{3.6}$$

Of course, we would still need to define the numerical fluxes $\hat{u}_{j+\frac{1}{2}}$ and $\hat{q}_{j+\frac{1}{2}}$. Again, based on the fact that upwinding and monotone fluxes are no longer relevant, as we are dealing with the heat equation (3.2) for which there is no preferred wind direction, we could still try the average flux

$$\hat{u}_{j+\frac{1}{2}} = \frac{1}{2}\left((u_h)^-_{j+\frac{1}{2}} + (u_h)^+_{j+\frac{1}{2}}\right), \; \hat{q}_{j+\frac{1}{2}} = \frac{1}{2}\left((q_h)^-_{j+\frac{1}{2}} + (q_h)^+_{j+\frac{1}{2}}\right). \quad (3.7)$$

Notice that, from the second equation in the scheme (3.6), we can solve q_h explicitly and locally (in the cell I_j) in terms of u_h, by inverting the small mass matrix inside the cell I_j. This is why the method is referred to as the "local" discontinuous Galerkin method. It turns out that the LDG scheme (3.6) with the central fluxes (3.7) is stable and convergent, but it loses one order of accuracy, to $O(h^k)$ only in the L^2 norm, for odd k. A better choice of the numerical fluxes is the so-called alternating fluxes, defined as

$$\hat{u}_{j+\frac{1}{2}} = (u_h)^-_{j+\frac{1}{2}}, \qquad \hat{q}_{j+\frac{1}{2}} = (q_h)^+_{j+\frac{1}{2}}. \quad (3.8)$$

The important point is that $\hat{q}$ and $\hat{u}$ should be chosen from different directions. Thus, the choice

$$\hat{u}_{j+\frac{1}{2}} = (u_h)^+_{j+\frac{1}{2}}, \qquad \hat{q}_{j+\frac{1}{2}} = (q_h)^-_{j+\frac{1}{2}}$$

is also fine. It can be proved that the LDG scheme (3.6) with the alternating fluxes (3.8) is stable and convergent, with optimal $O(h^{k+1})$ error order in the L^2 norm, see, e.g., [80].

The beauty of the DG method is that, once it is designed for the linear equation (3.2) and proved to be stable and accurate, it can be easily generalized to fully nonlinear convection-diffusion equations

$$u_t + f(u)_x = (a(u)u_x)_x \quad (3.9)$$

with $a(u) \geq 0$. We again rewrite this equation as the following system

$$u_t + f(u)_x - (b(u)q)_x = 0, \qquad q - B(u)_x = 0, \quad (3.10)$$

where

$$b(u) = \sqrt{a(u)}, \qquad B(u) = \int^u b(u)du. \quad (3.11)$$

The semi-discrete LDG scheme is defined as follows. Find $u_h, q_h \in V_h^k$ such that, for all test functions $v_h, p_h \in V_h^k$ and all $1 \leq i \leq N$, we have

$$\int_{I_j} (u_h)_t \, v_h dx - \int_{I_j} (f(u_h) - b(u_h)q_h)(v_h)_x dx$$
$$+(\hat{f} - \hat{b}\hat{q})_{j+\frac{1}{2}}(v_h)^-_{j+\frac{1}{2}} - (\hat{f} - \hat{b}\hat{q})_{j-\frac{1}{2}}(v_h)^+_{j-\frac{1}{2}} = 0, \quad (3.12)$$

$$\int_{I_j} q_h p_h dx + \int_{I_j} B(u_h)(p_h)_x dx - \hat{B}_{j+\frac{1}{2}}(p_h)^-_{j+\frac{1}{2}} + \hat{B}_{j-\frac{1}{2}}(p_h)^+_{j-\frac{1}{2}} = 0.$$

In [30], sufficient conditions for the choices of the numerical fluxes to guarantee the stability of the scheme (3.12) are given. Here, we only discuss a particularly attractive choice, i.e. the so-called alternating fluxes discussed before for the linear heat equation, now defined as

$$\hat{b} = \frac{B(u_h^+) - B(u_h^-)}{u_h^+ - u_h^-}, \qquad \hat{q} = q_h^+, \qquad \hat{B} = B(u_h^-). \tag{3.13}$$

The important point is that $\hat{q}$ and $\hat{B}$ should be chosen from different directions. Thus, the choice

$$\hat{b} = \frac{B(u_h^+) - B(u_h^-)}{u_h^+ - u_h^-}, \qquad \hat{q} = q_h^-, \qquad \hat{B} = B(u_h^+)$$

is also fine.

Notice that, even for this fully nonlinear case, from the second equation in the scheme (3.12), we can still solve q_h explicitly and locally (in cell I_j) in terms of u_h, by inverting the small mass matrix inside the cell I_j, thus justifying the terminology "local" discontinuous Galerkin methods.

In [30], it is proved that, for the solution u_h, q_h to the semi-discrete LDG scheme (3.12), we still have the following "cell entropy inequality"

$$\frac{1}{2}\frac{d}{dt}\int_{I_j} (u_h)^2\, dx + \int_{I_j} (q_h)^2 dx + \hat{F}_{j+\frac{1}{2}} - \hat{F}_{j-\frac{1}{2}} \le 0 \tag{3.14}$$

for a consistent entropy flux

$$\hat{F}_{j+\frac{1}{2}} = \hat{F}\left(u_h\left(x^-_{j+\frac{1}{2}}, t\right), q_h\left(x^-_{j+\frac{1}{2}}, t\right); u_h\left(x^+_{j+\frac{1}{2}}, t\right), q_h\left(x^+_{j+\frac{1}{2}}\right)\right)$$

satisfying $\hat{F}(u, q; u, q) = F(u) - ub(u)q$ where, as before, $F(u) = \int^u uf'(u)du$. This, together with periodic or compactly supported boundary conditions, implies the following L^2 stability

$$\frac{d}{dt}\int_a^b (u_h)^2 dx + 2\int_a^b (q_h)^2 dx \le 0, \tag{3.15}$$

or

$$\|u_h(\cdot, t)\|^2 + 2\int_0^t \|q_h(\cdot, \tau)\|^2 d\tau \le \|u_h(\cdot, 0)\|^2. \tag{3.16}$$

A priori L^2 error estimates for smooth solutions are provided in [94].

3.2. Internal Penalty Discontinuous Galerkin Methods. Another important class of DG methods for solving diffusion equations is the class of internal penalty discontinuous Galerkin methods. We will use the simple heat equation (3.2) to demonstrate the idea. If we multiply both sides of (3.2) by a test function v and integrate over the cell I_j, and integrate by parts for the right-hand side, we obtain the equality

$$\int_{I_j} u_t v dx = -\int_{I_j} u_x v_x dx + (u_x)_{j+\frac{1}{2}} v^-_{j+\frac{1}{2}} - (u_x)_{j-\frac{1}{2}} v^+_{j-\frac{1}{2}} \tag{3.17}$$

where we have used superscripts $\pm$ on v at cell boundaries to prepare for numerical schemes involving functions which are discontinuous at those cell boundaries. Summing over j, we obtain, with periodic boundary conditions for simplicity, the following equality

$$\int_a^b u_t v dx = -\sum_{j=1}^N \int_{I_j} u_x v_x dx - \sum_{j=1}^N (u_x)_{j+\frac{1}{2}} [v]_{j+\frac{1}{2}} \tag{3.18}$$

where $[w] \equiv w^+ - w^-$ denotes the jump of w at the cell interface. If we attempt to convert the equality (3.18) into a numerical scheme, we could try the following. Find $u_h \in V_h^k$ such that, for all test functions $v_h \in V_h^k$, we have

$$\int_a^b (u_h)_t (v_h) dx = -\sum_{j=1}^N \int_{I_j} (u_h)_x (v_h)_x dx - \sum_{j=1}^N \{(u_h)_x\}_{j+\frac{1}{2}} [v_h]_{j+\frac{1}{2}} \tag{3.19}$$

where $\{w\} \equiv \frac{1}{2}(w^+ + w^-)$ denotes the average of w at the cell interface. This scheme is actually exactly the same as the scheme (3.3) with the numerical flux (3.4), which is known to be unstable as mentioned above [31, 107]. Notice that the right-hand side of (3.19) is not symmetric with respect to u_h and v_h. We can therefore add another term to symmetrize it, obtaining the following scheme. Find $u_h \in V_h^k$ such that, for all test functions $v_h \in V_h^k$, we have

$$\begin{aligned} \int_a^b (u_h)_t (v_h) dx = &-\sum_{j=1}^N \int_{I_j} (u_h)_x (v_h)_x dx \\ &- \sum_{j=1}^N \{(u_h)_x\}_{j+\frac{1}{2}} [v_h]_{j+\frac{1}{2}} - \sum_{j=1}^N \{(v_h)_x\}_{j+\frac{1}{2}} [u_h]_{j+\frac{1}{2}}. \end{aligned} \tag{3.20}$$

Notice that, since the exact solution is continuous, the additional term $-\sum_{j=1}^N \{(v_h)_x\}_{j+\frac{1}{2}} [u_h]_{j+\frac{1}{2}}$ is zero if the numerical solution u_h is replaced by the exact solution u, hence the scheme is consistent. Scheme (3.20) is symmetric, unfortunately it is still unconditionally unstable. In order to

stabilize the scheme, a further penalty term must be added, resulting in the following symmetric internal penalty discontinuous Galerkin (SIPG) method [6, 84]

$$\int_a^b (u_h)_t (v_h) dx = -\sum_{j=1}^{N} \int_{I_j} (u_h)_x (v_h)_x dx - \sum_{j=1}^{N} \{(u_h)_x\}_{j+\frac{1}{2}} [v_h]_{j+\frac{1}{2}} - \sum_{j=1}^{N} \{(v_h)_x\}_{j+\frac{1}{2}} [u_h]_{j+\frac{1}{2}} - \sum_{j=1}^{N} \frac{\alpha}{h} [u_h]_{j+\frac{1}{2}} [v_h]_{j+\frac{1}{2}}. \tag{3.21}$$

Clearly, the scheme (3.21) is still symmetric, and it can be proved [6, 84] that, for sufficiently large α, it is stable and has optimal $O(h^{k+1})$ order convergence in L^2. The disadvantage of this scheme is that it involves a parameter α which has to be chosen adequately to ensure stability. Another possible way to obtain a stable scheme is to change the sign of the last term in the unstable scheme (3.20), resulting in the following non-symmetric internal penalty discontinuous Galerkin (NIPG) method [8, 66] of Baumann and Oden

$$\int_a^b (u_h)_t (v_h) dx = -\sum_{j=1}^{N} \int_{I_j} (u_h)_x (v_h)_x dx - \sum_{j=1}^{N} \{(u_h)_x\}_{j+\frac{1}{2}} [v_h]_{j+\frac{1}{2}} + \sum_{j=1}^{N} \{(v_h)_x\}_{j+\frac{1}{2}} [u_h]_{j+\frac{1}{2}}. \tag{3.22}$$

This scheme is not symmetric; however, it is L^2 stable and convergent, although it has a suboptimal $O(h^k)$ order of L^2 errors for even k [8, 66, 107].

There are other types of DG methods involving the internal penalty methodology, for example the direct discontinuous Galerkin (DDG) methods [60, 61].

3.3. Ultra Weak Discontinuous Galerkin Methods. Ultra weak discontinuous Galerkin methods are designed in [17]. Let us again use the simple heat equation (3.2) to demonstrate the idea. If we multiply both sides of (3.2) by a test function v and integrate over the cell I_j, and integrate by parts twice for the right-hand side, we obtain the equality

$$\int_{I_j} u_t v dx = \int_{I_j} u v_{xx} dx + (u_x)_{j+\frac{1}{2}} v_{j+\frac{1}{2}} - (u_x)_{j-\frac{1}{2}} v_{j-\frac{1}{2}} - u_{j+\frac{1}{2}} (v_x)_{j+\frac{1}{2}} + u_{j-\frac{1}{2}} (v_x)_{j-\frac{1}{2}}. \tag{3.23}$$

We can then follow the general principle of designing DG schemes, namely converting the solution u and its derivatives at the cell boundary into numerical fluxes, and taking values of the test function v and its derivatives

at the cell boundary by values inside the cell I_j, to obtain the following scheme. Find $u_h \in V_h^k$ such that, for all test functions $v_h \in V_h^k$ and all $1 \le j \le N$, we have

$$\int_{I_j} (u_h)_t \, v_h dx = \int_{I_j} u_h \, (v_h)_{xx} dx + \widehat{u_x}_{j+\frac{1}{2}} (v_h)^-_{j+\frac{1}{2}} - \widehat{u_x}_{j-\frac{1}{2}} (v_h)^+_{j-\frac{1}{2}} - \hat{u}_{j+\frac{1}{2}} ((v_h)_x)^-_{j+\frac{1}{2}} + \hat{u}_{j-\frac{1}{2}} ((v_h)_x)^+_{j-\frac{1}{2}}. \tag{3.24}$$

The crucial ingredient for the stability of the scheme (3.24) is still the choice of numerical fluxes. It is proved in [17] that the following choice of numerical fluxes

$$\hat{u}_{j+\frac{1}{2}} = (u_h)^-_{j+\frac{1}{2}}, \qquad \widehat{u_x}_{j+\frac{1}{2}} = ((u_h)_x)^+_{j+\frac{1}{2}} + \frac{\alpha}{h} [u_h]_{j+\frac{1}{2}} \tag{3.25}$$

would yield a stable DG scheme if the constant $\alpha > 0$ is sufficiently large. Notice that the choice in (3.25) is a combination of alternating fluxes and internal penalty. The following choice of alternating fluxes would also work

$$\hat{u}_{j+\frac{1}{2}} = (u_h)^+_{j+\frac{1}{2}}, \qquad \widehat{u_x}_{j+\frac{1}{2}} = ((u_h)_x)^-_{j+\frac{1}{2}} + \frac{\alpha}{h} [u_h]_{j+\frac{1}{2}}.$$

Suboptimal L^2 error estimates are given in [17] for the scheme (3.24) with the fluxes (3.25) for $k \ge 1$. In numerical experiments, optimal L^2 convergence rate of $O(h^{k+1})$ is observed for all $k \ge 1$. The scheme can be easily generalized to the general nonlinear convection-diffusion equation (3.9) with the same stability property [17].

3.4. Superconvergence. Results for superconvergence of DG methods, similar to those for hyperbolic equations discussed in Sect. 2.2, have been obtained for convection-diffusion equations in the literature. Superconvergence of the DG solution to the exact smooth solution in negative norms for convection-diffusion equations is studied in [49]. The superconvergence of the DG solution to a special projection of the exact smooth solution is addressed in [19, 20].

4. DG Methods for Third Order Convection-Dispersion Equations. In this section we study convection-dispersion equations which are wave equations involving third spatial derivatives. We study the following general KdV type equations

$$u_t + \sum_{i=1}^{d} f_i(u)_{x_i} + \sum_{i=1}^{d} \left(r_i'(u) \sum_{j=1}^{d} g_{ij} (r_i(u)_{x_i})_{x_j} \right)_{x_i} = 0, \tag{4.1}$$

where $f_i(u)$, $r_i(u)$, and $g_{ij}(q)$ are arbitrary (smooth) nonlinear functions. The one-dimensional KdV equation

$$u_t + (\alpha u + \beta u^2)_x + \sigma u_{xxx} = 0, \tag{4.2}$$

where α, β, and σ are constants, is a special case of the general class (4.1).

It is important to realize that third order dispersive equations are wave equations, sharing similarity with first order hyperbolic conservation laws and being quite different from diffusion equations. For example, the third order linear equation

$$u_t + u_{xxx} = 0 \tag{4.3}$$

admits the following simple wave solution

$$u(x,t) = \sin(x+t),$$

that is, information propagates from right to left. Therefore, upwinding is a relevant concept for the design of DG schemes for convection-dispersion equations.

We will discuss two classes of DG schemes for third order dispersive PDEs.

4.1. Local Discontinuous Galerkin Methods. We can again design LDG methods for third order dispersive PDEs, by rewriting such a PDE into a first order system and then applying the discontinuous Galerkin method on the system. Of course, a key ingredient for the success of such methods is still the correct design of interface numerical fluxes. These fluxes must be designed to guarantee stability and local solvability of all the auxiliary variables introduced to approximate the derivatives of the solution, thus justifying the terminology "local" DG. As mentioned above, upwinding should participate in the guiding principles for the design of numerical fluxes for dispersive PDEs.

We will use the simple linear equation (4.3) to demonstrate the idea of LDG schemes. We rewrite Eq. (4.3) as the following system

$$u_t + p_x = 0, \qquad p - q_x = 0, \qquad q - u_x = 0, \tag{4.4}$$

which "looks like" a system of conservation laws, except that the second and third equations do not have time derivatives. We can then formally write down the DG scheme (2.8) for each equation in (4.4), resulting in the following scheme. Find $u_h, p_h, q_h \in V_h^k$ such that, for all test functions $v_h, w_h, z_h \in V_h^k$ and all $1 \le j \le N$, we have

$$\begin{aligned}
&\int_{I_j} (u_h)_t \, v_h dx - \int_{I_j} p_h \, (v_h)_x dx + \hat{p}_{j+\frac{1}{2}} (v_h)^-_{j+\frac{1}{2}} - \hat{p}_{j-\frac{1}{2}} (v_h)^+_{j-\frac{1}{2}} = 0; \\
&\int_{I_j} p_h w_h dx + \int_{I_j} q_h \, (w_h)_x dx - \hat{q}_{j+\frac{1}{2}} (w_h)^-_{j+\frac{1}{2}} + \hat{q}_{j-\frac{1}{2}} (w_h)^+_{j-\frac{1}{2}} = 0; \\
&\int_{I_j} q_h z_h dx + \int_{I_j} u_h \, (z_h)_x dx - \hat{u}_{j+\frac{1}{2}} (z_h)^-_{j+\frac{1}{2}} + \hat{u}_{j-\frac{1}{2}} (z_h)^+_{j-\frac{1}{2}} = 0.
\end{aligned} \tag{4.5}$$

Of course, we would still need to define the numerical fluxes $\hat{u}_{j+\frac{1}{2}}$, $\hat{q}_{j+\frac{1}{2}}$ and $\hat{p}_{j+\frac{1}{2}}$. Since the wind blows from right to left, intuitively, we should choose more information from the right. If we take all three fluxes from the right

$$\hat{u} = (u_h)^+, \qquad \hat{q} = (q_h)^+, \qquad \hat{p} = (p_h)^+,$$

we obtain an unstable scheme. It is often the case that we should not attempt to be completely upwind, only upwind-biased. The next logical thing to try is to take two numerical fluxes from the right and one from the left. Analysis helps us to pinpoint the following upwind-biased alternating fluxes

$$\hat{u} = (u_h)^-, \qquad \hat{q} = (q_h)^+, \qquad \hat{p} = (p_h)^+. \tag{4.6}$$

In fact, it turns out that the most important thing is to take $\hat{q}$, which approximates u_x, from the upwind side $(q_h)^+$. The fluxes $\hat{u}$ and $\hat{p}$, approximating u and u_{xx}, respectively, can be taken in alternating sides. Therefore, the following choice of upwind-biased alternating fluxes

$$\hat{u} = (u_h)^+, \qquad \hat{q} = (q_h)^+, \qquad \hat{p} = (p_h)^-$$

is also fine.

It is proved in [103] that the LDG scheme (4.5) with the upwind-biased alternating fluxes (4.6) is L^2 stable

$$\frac{d}{dt}\int_a^b (u_h)^2 dx \le 0, \tag{4.7}$$

or

$$\|u_h(\cdot,t)\| \le \|u_h(\cdot,0)\|. \tag{4.8}$$

A suboptimal L^2 error estimate of order $O(h^{k+1/2})$ is also proved in [103]. In a more recent work [101], Xu and Shu proved optimal L^2 error estimate of order $O(h^{k+1})$ for this scheme. This extra half order turns out to be difficult to obtain, mainly because of the wave nature of the Eq. (4.3) and hence a lack of control of the derivatives. The approach in [101] is to establish stability not only for u_h as in (4.7), but also for q_h and p_h approximating u_x and u_{xx}.

The LDG scheme can be designed for the general nonlinear convection-dispersion equation (4.1). Let us use the one-dimensional case to describe the scheme

$$u_t + f(u)_x + (r'(u)g(r(u)_x)_x)_x = 0, \tag{4.9}$$

where $f(u)$, $r(u)$, and $g(q)$ are arbitrary (smooth) nonlinear functions. The LDG method is based on rewriting it as the following system

$$u_t + (f(u) + r'(u)p)_x = 0, \quad p - g(q)_x = 0, \quad q - r(u)_x = 0. \tag{4.10}$$

The semi-discrete LDG scheme is defined as follows. Find $u_h, p_h, q_h \in V_h^k$ such that, for all test functions $v_h, w_h, z_h \in V_h^k$ and all $1 \le i \le N$, we have

$$\begin{aligned}
&\int_{I_j} (u_h)_t \, v_h dx - \int_{I_j} (f(u_h) + r'(u_h) p_h)(v_h)_x dx \\
&\quad + (\hat{f} + \widehat{r'}\hat{p})_{j+\frac{1}{2}} (v_h)^-_{j+\frac{1}{2}} - (\hat{f} + \widehat{r'}\hat{p})_{j-\frac{1}{2}} (v_h)^+_{j-\frac{1}{2}} = 0; \qquad (4.11) \\
&\int_{I_j} p_h w_h dx + \int_{I_j} g(q_h) \, (w_h)_x dx - \hat{g}_{j+\frac{1}{2}} (w_h)^-_{j+\frac{1}{2}} + \hat{g}_{j-\frac{1}{2}} (w_h)^+_{j-\frac{1}{2}} = 0; \\
&\int_{I_j} q_h z_h dx + \int_{I_j} r(u_h) \, (z_h)_x dx - \hat{r}_{j+\frac{1}{2}} (z_h)^-_{j+\frac{1}{2}} + \hat{r}_{j-\frac{1}{2}} (z_h)^+_{j-\frac{1}{2}} = 0.
\end{aligned}$$

By our experience with linear equations discussed above, we would like to use the following upwind-biased alternating fluxes

$$\widehat{r'} = \frac{r((u_h)^+) - r((u_h)^-)}{(u_h)^+ - (u_h)^-}, \; \hat{r} = r((u_h)^-), \; \hat{g} = \hat{g}((q_h)^-, (q_h)^+), \; \hat{p} = (p_h)^+. \qquad (4.12)$$

Here, $-\hat{g}((q_h)^-, (q_h)^+)$ is a monotone flux for $-g(q)$, namely $\hat{g}$ is a non-increasing function in the first argument and a non-decreasing function in the second argument. The important point is again the "alternating fluxes," namely $\hat{r}$ and $\hat{p}$ should come from opposite sides. Thus

$$\widehat{r'} = \frac{r((u_h)^+) - r((u_h)^-)}{(u_h)^+ - (u_h)^-}, \; \hat{r} = r((u_h)^+), \; \hat{g} = \hat{g}((q_h)^-, (q_h)^+), \; \hat{p} = (p_h)^-$$

would also work.

It is quite interesting to observe that monotone fluxes, which are originally designed for hyperbolic conservation laws, can be used also for nonlinear dispersive equations to obtain stability. Also notice that, from the third equation in the scheme (4.11), we can solve q_h explicitly and locally (in cell I_j) in terms of u_h, by inverting the small mass matrix inside the cell I_j. Then, from the second equation in the scheme (4.11), we can solve p_h explicitly and locally (in cell I_j) in terms of q_h. Thus only u_h is the global unknown and the auxiliary variables q_h and p_h can be solved in terms of u_h locally. This justifies again the terminology of "local" discontinuous Galerkin method.

It is proved in [103] that the LDG scheme (4.11) with the upwind-biased alternating fluxes (4.12) is L^2 stable, i.e. (4.7) or (4.8) holds. This is also true for the multi-dimensional case (4.1). A suboptimal L^2 error estimate of order $O(h^{k+1/2})$ is also proved in [94].

4.2. Ultra Weak Discontinuous Galerkin Methods. Ultra weak discontinuous Galerkin methods are designed in [17]. Let us again use the simple linear equation (4.3) to demonstrate the idea. If we multiply both

sides of (4.3) by a test function v and integrate over the cell I_j, and integrate by parts three times for the right-hand side, we obtain the equality

$$\begin{aligned}\int_{I_j} u_t v dx - \int_{I_j} u v_{xxx} dx + (u_{xx})_{j+\frac{1}{2}} v_{j+\frac{1}{2}} - (u_{xx})_{j-\frac{1}{2}} v_{j-\frac{1}{2}} \\ -(u_x)_{j+\frac{1}{2}} (v_x)_{j+\frac{1}{2}} + (u_x)_{j-\frac{1}{2}} (v_x)_{j-\frac{1}{2}} \\ +u_{j+\frac{1}{2}} (v_{xx})_{j+\frac{1}{2}} - u_{j-\frac{1}{2}} (v_{xx})_{j-\frac{1}{2}} = 0.\end{aligned} \tag{4.13}$$

We can then follow the general principle of designing DG schemes, namely converting the solution u and its derivatives at the cell boundary into numerical fluxes, and taking values of the test function v and its derivatives at the cell boundary by values inside the cell I_j, to obtain the following scheme. Find $u_h \in V_h^k$ such that, for all test functions $v_h \in V_h^k$ and all $1 \le j \le N$, we have

$$\begin{aligned}\int_{I_j} (u_h)_t \, v_h dx - \int_{I_j} u_h \, (v_h)_{xxx} dx + \widehat{u_{xx}}_{j+\frac{1}{2}} (v_h)^-_{j+\frac{1}{2}} \\ -\widehat{u_{xx}}_{j-\frac{1}{2}} (v_h)^+_{j-\frac{1}{2}} - \widehat{u_x}_{j+\frac{1}{2}} ((v_h)_x)^-_{j+\frac{1}{2}} + \widehat{u_x}_{j-\frac{1}{2}} ((v_h)_x)^+_{j-\frac{1}{2}} \\ +\hat{u}_{j+\frac{1}{2}} ((v_h)_{xx})^-_{j+\frac{1}{2}} - \hat{u}_{j-\frac{1}{2}} ((v_h)_{xx})^+_{j-\frac{1}{2}} = 0.\end{aligned} \tag{4.14}$$

The crucial ingredient for the stability of the scheme (4.14) is still the choice of numerical fluxes. It is proved in [17] that the following choice of upwind-biased alternating fluxes

$$\hat{u} = (u_h)^-, \qquad \widehat{u_x} = ((u_h)_x)^+, \qquad \widehat{u_{xx}} = ((u_h)_{xx})^+, \tag{4.15}$$

would yield a stable DG scheme. Notice that the choice of numerical fluxes (4.15) is exactly the same as that for the stable LDG scheme (4.6). The most important thing is to take $\widehat{u_x}$ from the upwind side $((u_h)_x)^+$. The fluxes $\hat{u}$ and $\widehat{u_{xx}}$ can be taken in alternating sides. Therefore, the following choice of upwind-biased alternating fluxes

$$\hat{u} = (u_h)^+, \qquad \widehat{u_x} = ((u_h)_x)^+, \qquad \widehat{u_{xx}} = ((u_h)_{xx})^-$$

is also fine.

It is proved in [17] that the ultra weak DG scheme (4.14) with the upwind-biased alternating fluxes (4.15) is L^2 stable, namely (4.7) or (4.8) holds. Suboptimal L^2 error estimates are also given in [17] for this scheme with $k \ge 2$. In numerical experiments, optimal L^2 convergence rate of $O(h^{k+1})$ is observed for all $k \ge 2$. The scheme can be easily generalized to general nonlinear convection-dispersion equations with the same stability property [17].

5. DG Methods for Other Dispersive Wave Equations. DG methods have also been designed for other dispersive wave equations containing higher order (usually odd order) derivatives. We will describe briefly some of these schemes in this section.

5.1. Equations with Fifth Order Spatial Derivatives. An LDG scheme for solving the following fifth order convection-dispersion equation

$$u_t + \sum_{i=1}^{d} f_i(u)_{x_i} + \sum_{i=1}^{d} g_i(u_{x_i x_i})_{x_i x_i x_i} = 0, \tag{5.1}$$

where $f_i(u)$ and $g_i(q)$ are arbitrary functions, was designed in [104]. The numerical fluxes are chosen following the same upwind-biased alternating fluxes principle similar to the third order KdV type Eq. (4.1), namely the flux corresponding to u_{xx} should be chosen according to upwinding, and the flux pairs corresponding to u and u_{xxxx}, and the flux pairs corresponding to u_x and u_{xxx}, should be chosen in an alternating fashion within each pair. A cell entropy inequality and the L^2 stability of the LDG scheme for the nonlinear equation (5.1) can be proved [104], which again do not depend on the smoothness of the solution of (5.1), the order of accuracy of the scheme, or the triangulation. For the linear fifth order equation

$$u_t + u_{xxxxx} = 0, \tag{5.2}$$

optimal $O(h^{k+1})$ order L^2 error estimate is obtained in [101]. Similar results can be obtained for PDEs of higher odd order spatial derivatives.

Ultra weak DG methods for solving (5.2) are designed in [17]. The choice of numerical fluxes are identical to that for the LDG schemes described above. The resulting scheme can be proved to be L^2 stable, and a suboptimal L^2 error estimate for $k \geq 4$ is proved in [17]. Numerical experiments indicate optimal convergence in L^2 for all $k \geq 4$. The scheme as well as the stability analysis can be generalized to certain nonlinear fifth order PDEs [17]. Similar results can also be obtained for PDEs of higher odd order spatial derivatives.

5.2. The $K(m, n)$ Equation. The so-called $K(m, n)$ equation

$$u_t + (u^m)_x + (u^n)_{xxx} = 0 \tag{5.3}$$

arises from mathematical physics and has the *compactons* solutions. In [58], an LDG scheme is designed for (5.3), which is proved to be L^{n+1} stable for the $K(n, n)$ equation with odd n. For all other cases, the LDG scheme is proved to be linearly stable. Computational results including those for compactons indicate excellent performance of these schemes.

This example indicates that we do not always seek to prove L^2 stability for DG schemes. In fact, most time-dependent PDEs arising from physics and applications have certain "energy," which is a positive functional of the solution, and the energy usually does not increase with time. An ideal DG scheme would produce numerical solutions for which the same energy also does not increase with time. In this particular example ($K(n, n)$ equation with odd n), this "energy" is the square of the L^{n+1} norm.

5.3. The KdV-Burgers Type Equations. LDG methods for solving the KdV-Burgers (KdVB) equations

$$u_t + f(u)_x - (a(u)u_x)_x + (r'(u)g(r(u)_x)_x)_x = 0, \tag{5.4}$$

where $f(u)$, $a(u) \geq 0$, $r(u)$, and $g(q)$ are arbitrary functions, are designed in [90]. The design of numerical fluxes follows the same lines as that for the convection-diffusion equation (3.9) for the second derivative term $(a(u)u_x)_x$ and for the KdV type Eq. (4.9) for the third derivative term $(r'(u)g(r(u)_x)_x)_x$. A cell entropy inequality and the L^2 stability of the LDG scheme for the nonlinear equation (5.4) are proved [90], which again do not depend on the smoothness of the solution of (5.4) and the order of accuracy of the scheme. For smooth solutions, a suboptimal $O(h^{k+1/2})$ order L^2 error estimate for the linearized version is proved [90]. The LDG scheme is used in [90] to study different regimes when one of the dissipation and the dispersion mechanisms dominates, and when they have comparable influence on the solution. An advantage of the LDG scheme designed in [90] is that it is stable regardless of which mechanism (convection, diffusion, dispersion) actually dominates.

5.4. The Fifth-Order KdV-Type Equations. An LDG scheme is designed in [90] for the fifth-order KdV type equations

$$u_t + f(u)_x + (r'(u)g(r(u)_x)_x)_x + (s'(u)h(s(u)_{xx})_{xx})_x = 0, \tag{5.5}$$

where $f(u)$, $r(u)$, $g(q)$, $s(u)$, and $h(p)$ are arbitrary functions, and a cell entropy inequality and L^2 stability are proved. A special case is the Kawahara equation

$$u_t + uu_x + u_{xxx} - \delta u_{xxxxx} = 0$$

which has very interesting close-form exact solutions that can be used to test the accuracy of the scheme [90]. Other special cases of (5.5) include the generalized Kawahara equation, Ito's fifth-order KdV equation, and a fifth-order KdV type equations with high nonlinearity, which are also explored in [90].

5.5. The Fully Nonlinear $K(n, n, n)$ Equations. LDG methods for solving the fifth-order fully nonlinear $K(n, n, n)$ equations

$$u_t + (u^n)_x + (u^n)_{xxx} + (u^n)_{xxxxx} = 0, \tag{5.6}$$

where n is a positive integer, have been designed in [90]. The design of numerical fluxes follows the same lines as that for the $K(m, n)$ Eq. (5.3). For odd n, stability in the L^{n+1} norm of the resulting LDG scheme can be proved for the nonlinear equation (5.6) [90]. This scheme is used to simulate compacton propagation in [90].

5.6. The Nonlinear Schrödinger Equations. The nonlinear Schrödinger (NLS) equation

$$i\,u_t + u_{xx} + i\,(g(|u|^2)u)_x + f(|u|^2)u = 0, \tag{5.7}$$

the two-dimensional version

$$i\,u_t + \Delta u + f(|u|^2)u = 0, \tag{5.8}$$

and the coupled nonlinear Schrödinger equation

$$\begin{cases} i\,u_t + i\,\alpha u_x + u_{xx} + \beta\,u + \kappa v + f(|u|^2, |v|^2)u = 0 \\ i\,v_t - i\,\alpha v_x + v_{xx} - \beta\,u + \kappa v + g(|u|^2, |v|^2)v = 0, \end{cases} \tag{5.9}$$

where $f(q)$ and $g(q)$ are arbitrary functions and α, β, and κ are constants, are also dispersive wave equations, even though they involve second order spatial derivatives with $i = \sqrt{-1}$ as the coefficient. In [91], LDG methods are designed for these equations. The cell entropy inequality and L^2 stability are proved for these schemes in [91]. For smooth solutions, an L^2 error estimate of $O(h^{k+1/2})$ for the linearized version is also obtained in [91]. The LDG scheme is used in [91] to simulate the soliton propagation and interaction, and the appearance of singularities. The easiness of $h-p$ adaptivity of the LDG scheme and rigorous stability for the fully nonlinear case make it an ideal choice for the simulation of Schrödinger equations, for which the solutions often have quite localized structures.

5.7. The Ito-Type Coupled KdV Equations. An LDG method is developed in [93] to solve the Ito-type coupled KdV equations

$$u_t + \alpha u u_x + \beta v v_x + \gamma u_{xxx} = 0, v_t + \beta(uv)_x = 0,$$

where α, β, and γ are constants. An L^2 stability is proved for the LDG method. For the Ito's equation

$$\begin{aligned} u_t - (3u^2 + v^2)_x - u_{xxx} &= 0, \\ v_t - 2(uv)_x &= 0, \end{aligned}$$

the result for u behaves like dispersive wave solutions and the result for v behaves like shock wave solutions. Simulation for such solutions is performed in [93] using the LDG scheme.

5.8. The Kadomtsev–Petviashvili (KP) Equations. The two-dimensional Kadomtsev–Petviashvili (KP) equations

$$(u_t + 6uu_x + u_{xxx})_x + 3\sigma^2 u_{yy} = 0, \tag{5.10}$$

where $\sigma^2 = -1$ (referred to as KP-I) or $\sigma^2 = 1$ (referred to as KP-II) are generalizations of the one-dimensional KdV equations and are important models for water waves.

This equation is equivalent to

$$u_t + 6(uu_x) + u_{xxx} + 3\sigma^2 \partial_x^{-1} u_{yy} = 0 \tag{5.11}$$

where the nonlocal operator ∂_x^{-1} makes the equation well posed only in the restricted space

$$\mathcal{V}(R^2) = \left\{ f : \int_{R^2} \left(1 + \xi^2 + \frac{\eta^2}{\xi^2}\right) |\hat{f}(\xi, \eta)|^2 d\xi d\eta < \infty \right\}.$$

It is therefore complicated to design an efficient LDG scheme which relies on local operations. In [92], an LDG scheme for (5.10) is designed by carefully choosing locally supported bases which satisfy the global constraint needed by the solution of (5.10). The LDG scheme satisfies a cell entropy inequality and is L^2 stable for the fully nonlinear equation (5.10). Numerical simulations are performed in [92] for both the KP-I equations and the KP-II equations. Line solitons and lump-type pulse solutions have been simulated.

5.9. The Zakharov–Kuznetsov (ZK) Equation. The two-dimensional Zakharov–Kuznetsov (ZK) equation

$$u_t + (3u^2)_x + u_{xxx} + u_{xyy} = 0 \tag{5.12}$$

is another generalization of the one-dimensional KdV equations.

An LDG scheme is designed for (5.12) in [92]. A cell entropy inequality and the L^2 stability are proved. A suboptimal L^2 error estimate is given in [94]. Various nonlinear waves have been simulated by this scheme in [92].

5.10. The Camassa–Holm (CH) Equation. The Camassa–Holm (CH) equation is given as

$$u_t - u_{xxt} + 2\kappa u_x + 3uu_x = 2u_x u_{xx} + uu_{xxx},$$

where κ is a constant. An LDG scheme is designed in [95]. L^2 stability for general solutions and a suboptimal L^2 error estimate for smooth solutions are provided in [95].

5.11. The Hunter–Saxton (HS) Equation. The Hunter–Saxton (HS) equation is given as

$$u_{xxt} + 2u_x u_{xx} + uu_{xxx} = 0.$$

A regularization with viscosity is given as

$$u_{xxt} + 2u_x u_{xx} + uu_{xxx} - \varepsilon_1 u_{xxxx} = 0,$$

and a regularization with dispersion is given as

$$u_{xxt} + 2u_x u_{xx} + uu_{xxx} - \varepsilon_2 u_{xxxxx} = 0,$$

where $\varepsilon_1 \geq 0$ and ε_2 are small constants.

In [96, 98], we design LDG schemes for these equations and prove their energy stability.

5.12. The Generalized Zakharov System. The following system

$$iE_t + \Delta E - Nf(|E|^2)E + g(|E|^2)E = 0,$$
$$\epsilon^2 N_{tt} - \Delta(N + F(|E|^2)) = 0$$

is referred to as the generalized Zakharov system and is originally introduced to describe the Langmuir turbulence in a plasma. In [87], we design an LDG scheme for this system and prove two energy conservations for this scheme. Numerical experiments for the Zakharov system are presented to illustrate the accuracy and capability of the methods, including accuracy tests, plane waves, soliton–soliton collisions of the standard and generalized Zakharov system and a two-dimensional problem.

5.13. The Degasperis–Procesi (DP) Equation. The Degasperis–Procesi (DP) equation is given as

$$u_t - u_{txx} + 4f(u)_x = f(u)_{xxx},$$

where $f(u) = \frac{1}{2}u^2$. The solution may be discontinuous regardless of the smoothness of the initial conditions.

In [100], we develop LDG methods and prove L^2 stability for the general polynomial spaces and total variation stability for P^0 elements. The numerical simulation results for different types of solutions of the nonlinear Degasperis–Procesi equation are provided to illustrate the accuracy and capability of the LDG method in [100].

6. DG Methods for Other Dissipative Equations. DG methods have also been designed for other dissipative equations containing higher even order derivatives. We will describe briefly some of these schemes in this section.

6.1. The Bi-harmonic Type Equations. An LDG scheme for solving the time-dependent convection bi-harmonic equation

$$u_t + \sum_{i=1}^{d} f_i(u)_{x_i} + \sum_{i=1}^{d} (a_i(u_{x_i})u_{x_i x_i})_{x_i x_i} = 0, \tag{6.1}$$

where $f_i(u)$ and $a_i(q) \geq 0$ are arbitrary functions, was designed in [104]. The numerical fluxes are chosen following the same "alternating fluxes" principle similar to the second order convection-diffusion equation (3.1), namely the flux pairs corresponding to u and u_{xxx}, and the flux pairs corresponding to u_x and u_{xx}, should be chosen in an alternating fashion within each pair. A cell entropy inequality and the L^2 stability of the LDG scheme for the nonlinear equation (6.1) can be proved [104], which do not depend on the smoothness of the solution of (6.1), the order of accuracy of the scheme, or the triangulation. Optimal L^2 error estimates can be proved for the linear biharmonic equation

$$u_t + \triangle^2 u = 0, \tag{6.2}$$

for both structured and unstructured meshes, see [36]. In [63], superconvergence of the LDG method for linear fourth order equations is studied.

Ultra weak DG methods for solving (6.2) are designed in [17]. The choice of numerical fluxes are similar to that for the LDG schemes described above, with additional internal penalty term on the flux corresponding to u_{xxx}. The resulting scheme can be proved to be L^2 stable, and a suboptimal L^2 error estimate for $k \geq 3$ is proved in [17]. Numerical experiments indicate optimal convergence in L^2 for all $k \geq 3$. The scheme as well as the stability analysis can be generalized to certain nonlinear fourth order PDEs [17].

Both the LDG schemes and the ultra weak DG methods can be generalized to PDEs of higher even order spatial derivatives. For example, [36] contains optimal L^2 error estimates for linear diffusion PDEs with higher even orders.

6.2. The Kuramoto–Sivashinsky Type Equations. LDG methods are developed in [93] to solve the Kuramoto–Sivashinsky type equations

$$u_t + f(u)_x - (a(u)u_x)_x + (r'(u)g(r(u)_x)_x)_x + (s(u_x)u_{xx})_{xx} = 0, \quad (6.3)$$

where $f(u)$, $a(u)$, $r(u)$, $g(q)$, and $s(p) \geq 0$ are arbitrary functions. The Kuramoto–Sivashinsky equation

$$u_t + uu_x + \alpha u_{xx} + \beta u_{xxxx} = 0, \quad (6.4)$$

where α and $\beta \geq 0$ are constants, which is a special case of (6.3), is a canonical evolution equation which has attracted considerable attention over the last decades. When the coefficients α and β are both positive, its linear terms describe a balance between long-wave instability and short-wave stability, with the nonlinear term providing a mechanism for energy transfer between wave modes. The LDG method developed in [93] can be proved to satisfy a cell entropy inequality and is therefore L^2 stable, for the general nonlinear equation (6.3). The LDG scheme is used in [93] to simulate chaotic solutions of (6.4).

6.3. Semi-conductor Device Simulations. Device simulation models in semi-conductor device simulations include drift-diffusion, hydrodynamic, energy transport, high field, kinetic and Boltzmann–Poisson models. DG or LDG methods have been designed for these models, many of them with stability analysis and error estimates.

In [44, 45], an LDG method is designed to solve time-dependent and steady state moment models including the hydrodynamic (HD) models and the energy transport (ET) models, for semiconductor device simulations, in which both the first derivative convection terms and second derivative diffusion (heat conduction) terms exist and are discretized by the DG method and the LDG method, respectively. The potential equation for the electric field is also discretized by the LDG method, thus the numerical tool is based

on a unified discontinuous Galerkin methodology for different components and is hence potentially viable for efficient $h-p$ adaptivity and parallel implementation. One-dimensional n^+-n-n^+ diode and two-dimensional MESFET device are simulated by the DG methods using the HD and ET models and comparison is made with earlier finite difference essentially non-oscillatory (ENO) simulation results. In [46], we obtain L^2 error estimates for smooth solutions of the drift-diffusion (DD) and high-field (HF) models using the LDG method.

In [14, 15], a discontinuous Galerkin scheme applied to deterministic computations of the transients for the Boltzmann–Poisson system describing electron transport in semiconductor devices is developed and applied to simulate hot electron transport in bulk silicon, in a silicon n^+-n-n^+ diode and in a double gated 12nm MOSFET. Additionally, the obtained results are compared to those of a high order WENO scheme simulation.

6.4. Cahn–Hilliard Equations. An important class of high order nonlinear diffusion equations is the class of the Cahn–Hilliard equation

$$u_t = \nabla \cdot (b(u)\nabla(-\gamma\Delta u + \Psi'(u))), \tag{6.5}$$

and the Cahn–Hilliard system

$$\mathbf{u}_t = \nabla \cdot (\mathbf{B}(\mathbf{u})\nabla\omega), \qquad \omega = -\gamma\Delta\mathbf{u} + D\Psi(\mathbf{u}), \tag{6.6}$$

where $\{D\Psi(\mathbf{u})\}_l = \frac{\partial\Psi(\mathbf{u})}{\partial u_l}$ and γ is a positive constant. Here $b(u)$ is the nonnegative diffusion mobility and $\Psi(u)$ is the homogeneous free energy density for the scalar case (6.5). For the system case (6.6), $\mathbf{B}(\mathbf{u})$ is the symmetric positive semi-definite mobility matrix and $\Psi(\mathbf{u})$ is the homogeneous free energy density.

In [85, 86], LDG methods are designed for the Cahn–Hilliard equation (6.5) and the Cahn–Hilliard system (6.6), respectively. The proof of the energy stability for the LDG schemes is given for the general nonlinear solutions. Many simulation results are given. In [36], optimal L^2 error estimate is obtained for the LDG method for solving the linearized Cahn–Hilliard equation.

6.5. The Surface Diffusion Equations. The surface diffusion equation is given as

$$u_t + \nabla \cdot \left(Q \left(\mathbf{I} - \frac{\nabla u \otimes \nabla u}{Q^2} \right) \nabla H \right) = 0 \tag{6.7}$$

where Q is the area element

$$Q = \sqrt{1 + |\nabla u|^2}$$

and H is the mean curvature of the domain boundary Γ

$$H = \nabla \cdot \left(\frac{\nabla u}{Q} \right).$$

The Willmore flow equation is given as

$$u_t + Q\nabla \cdot \left(\frac{1}{Q} \left(\mathbf{I} - \frac{\nabla u \otimes \nabla u}{Q^2} \right) \nabla (QH) \right) - \frac{1}{2} Q \nabla \cdot \left(\frac{H^2}{Q} \nabla u \right) = 0. \tag{6.8}$$

In [97], LDG methods are designed for both the surface diffusion equation (6.7) and the Willmore flow equation (6.8). Energy stability is proved. In [47, 48], L^2 error estimates are given for these LDG methods.

7. Concluding Remarks. In these lectures we have given a brief summary of discontinuous Galerkin (DG) methods for time-dependent PDEs. Clearly, DG schemes can be designed for a wide class of PDEs, both of the dispersive type and of the dissipative type, and often energy stability similar to those for the exact solution of the PDEs can be obtained. Among the current and future research topics for discontinuous Galerkin method, we would like to point out the following.

First, it is worthwhile to study efficient time discretization techniques. While the explicit TVD Runge–Kutta time discretization might be suitable for hyperbolic equations or strongly convection dominated problems, for other equations, the time step restriction is too severe to use explicit Runge–Kutta time discretization for the semi-discrete DG schemes. Suitable time discretization techniques, such as exponential time stepping, preconditioning and multigrid techniques, are being investigated. It is particularly challenging to design efficient time discretization techniques for PDEs with high and odd leading order of spatial derivatives (dispersive type PDEs), especially when the leading term is nonlinear.

Second, it is worthwhile to study effective and efficient error indicators and *a posteriori* error estimates, to guide the design of both h and p adaptivity. DG methods have the flexibility in h-p adaptivity; however, this potential can only be fully realized if we have reliable error indicators to tell us where to refine or coarsen the mesh and where to increase or decrease the polynomial degree. Again, it is particularly challenging to design reliable error indicators for PDEs with high and odd leading order of spatial derivatives (dispersive type PDEs).

Finally, it is worthwhile to study the design and stability analysis of DG schemes for more demanding nonlinear PDEs from applications.

Acknowledgement. The work of the author was supported in part by NSF grant DMS-1112700 and DOE grant DE-FG02-08ER25863.

REFERENCES

[1] S. Adjerid and M. Baccouch, *Asymptotically exact a posteriori error estimates for a one-dimensional linear hyperbolic problem*, Applied Numerical Mathematics, **60** (2010), pp. 903–914.

[2] S. Adjerid, K. Devine, J. Flaherty and L. Krivodonova, *A posteriori error estimation for discontinuous Galerkin solutions of hyperbolic problems*, Computational Methods in Applied Mechanics and Engineering, **191** (2002), pp. 1097–1112.

[3] S. Adjerid and T. Massey, *Superconvergence of discontinuous Galerkin solutions for a nonlinear scalar hyperbolic problem*, Computer Methods in Applied Mechanics and Engineering, **195** (2006), pp. 3331–3346.

[4] S. Adjerid and T. Weinhart, *Discontinuous Galerkin error estimation for linear symmetric hyperbolic systems*, Computer Methods in Applied Mechanics and Engineering, **198** (2009), pp. 3113–3129.

[5] S. Adjerid and T. Weinhart, *Discontinuous Galerkin error estimation for linear symmetrizable hyperbolic systems*, Mathematics of Computations, **80** (2011), pp. 1335–1367.

[6] D.N. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM Journal on Numerical Analysis, **39** (1982), pp. 742–760.

[7] F. Bassi and S. Rebay, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations*, Journal of Computational Physics, **131** (1997), pp. 267–279.

[8] C.E. Baumann and J.T. Oden, *A discontinuous hp finite element method for convection-diffusion problems*, Computer Methods in Applied Mechanics and Engineering, **175** (1999), pp. 311–341.

[9] R. Biswas, K.D. Devine and J. Flaherty, *Parallel, adaptive finite element methods for conservation laws*, Applied Numerical Mathematics, **14** (1994), pp. 255–283.

[10] O. Bokanowski, Y. Cheng and C.-W. Shu, *A discontinuous Galerkin solver for front propagation*, SIAM Journal on Scientific Computing, **33** (2011), pp. 923–938.

[11] O. Bokanowski, Y. Cheng and C.-W. Shu, *A discontinuous Galerkin scheme for front propagation with obstacles*, Numerische Mathematik, to appear. DOI: 10.1007/s00211-013-0555-3

[12] J.H. Bramble and A.H. Schatz, *High order local accuracy by averaging in the finite element method*, Mathematics of Computation, **31** (1977), pp. 94–111.

[13] A. Burbeau, P. Sagaut and Ch.H. Bruneau, *A problem-independent limiter for high-order Runge-Kutta discontinuous Galerkin methods*, Journal of Computational Physics, **169** (2001), pp. 111–150.

[14] Y. Cheng, I.M. Gamba, A. Majorana and C.-W. Shu, *Discontinuous Galerkin solver for Boltzmann-Poisson transients*, Journal of Computational Electronics, **7** (2008), pp. 119–123.

[15] Y. Cheng, I.M. Gamba, A. Majorana and C.-W. Shu, *A discontinuous Galerkin solver for Boltzmann Poisson systems in nano devices*, Computer Methods in Applied Mechanics and Engineering, **198** (2009), pp. 3130–3150.

[16] Y. Cheng and C.-W. Shu, *A discontinuous Galerkin finite element method for directly solving the Hamilton-Jacobi equations*, Journal of Computational Physics, **223** (2007), pp. 398–415.

[17] Y. Cheng and C.-W. Shu, *A discontinuous Galerkin finite element method for time dependent partial differential equations with higher order derivatives*, Mathematics of Computation, **77** (2008), pp. 699–730.

[18] Y. Cheng and C.-W. Shu, *Superconvergence and time evolution of discontinuous Galerkin finite element solutions*, Journal of Computational Physics, **227** (2008), pp. 9612–9627.

[19] Y. Cheng and C.-W. Shu, *Superconvergence of local discontinuous Galerkin methods for one-dimensional convection-diffusion equations*, Computers & Structures, **87** (2009), pp. 630–641.

[20] Y. Cheng and C.-W. Shu, *Superconvergence of discontinuous Galerkin and local discontinuous Galerkin schemes for linear hyperbolic and convection diffusion equations in one space dimension*, SIAM Journal on Numerical Analysis, **47** (2010), pp. 4044–4072.

[21] B. Cockburn, *Discontinuous Galerkin methods for convection-dominated problems*, in High-Order Methods for Computational Physics, T.J. Barth and H. Deconinck (eds.), Lecture Notes in Computational Science and Engineering, volume 9, Springer, 1999, pp. 69–224.

[22] B. Cockburn, B. Dong and J. Guzmán, *Optimal convergence of the original DG method for the transport-reaction equation on special meshes*, SIAM Journal on Numerical Analysis, **46** (2008), pp. 1250–1265.

[23] 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*, Mathematics of Computation, **54** (1990), pp. 545–581.

[24] B. Cockburn, G. Karniadakis and C.-W. Shu, *The development of discontinuous Galerkin methods*, in Discontinuous Galerkin Methods: Theory, Computation and Applications, B. Cockburn, G. Karniadakis and C.-W. Shu (eds.), Lecture Notes in Computational Science and Engineering, volume 11, Springer, 2000, Part I: Overview, pp. 3–50.

[25] 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*, Journal of Computational Physics, **84** (1989), pp. 90–113.

[26] B. Cockburn, M. Luskin, C.-W. Shu and E. Süli, *Enhanced accuracy by post-processing for finite element methods for hyperbolic equations*, Mathematics of Computation, **72** (2003), pp. 577–606.

[27] B. Cockburn and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws II: general framework*, Mathematics of Computation, **52** (1989), pp. 411–435.

[28] B. Cockburn and C.-W. Shu, *The Runge-Kutta local projection P^1-discontinuous-Galerkin finite element method for scalar conservation laws*, Mathematical Modelling and Numerical Analysis (M^2AN), **25** (1991), pp. 337–361.

[29] B. Cockburn and C.-W. Shu, *The Runge-Kutta discontinuous Galerkin method for conservation laws V: multidimensional systems*, Journal of Computational Physics, **141** (1998), pp. 199–224.

[30] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection diffusion systems*, SIAM Journal on Numerical Analysis, **35** (1998), pp. 2440–2463.

[31] B. Cockburn and C.-W. Shu, *Runge-Kutta Discontinuous Galerkin methods for convection-dominated problems*, Journal of Scientific Computing, **16** (2001), pp. 173–261.

[32] B. Cockburn and C.-W. Shu, *Foreword for the special issue on discontinuous Galerkin method*, Journal of Scientific Computing, **22–23** (2005), pp. 1–3.

[33] B. Cockburn and C.-W. Shu, *Foreword for the special issue on discontinuous Galerkin method*, Journal of Scientific Computing, **40** (2009), pp. 1–3.

[34] S. Curtis, R.M. Kirby, J.K. Ryan and C.-W. Shu, *Post-processing for the discontinuous Galerkin method over non-uniform meshes*, SIAM Journal on Scientific Computing, **30** (2007), pp. 272–289.

[35] C. Dawson, *Foreword for the special issue on discontinuous Galerkin method*, Computer Methods in Applied Mechanics and Engineering, **195** (2006), p. 3183.

[36] B. Dong and C.-W. Shu, *Analysis of a local discontinuous Galerkin method for linear time-dependent fourth-order problems*, SIAM Journal on Numerical Analysis, **47** (2009), pp. 3240–3268.

[37] A. Harten, *High resolution schemes for hyperbolic conservation laws*, Journal of Computational Physics, **49** (1983), pp. 357–393.

[38] J. Hesthaven and T. Warburton, *Nodal Discontinuous Galerkin Methods*, Springer, New York, 2008.

[39] S. Hou and X.-D. Liu, *Solutions of multi-dimensional hyperbolic systems of conservation laws by square entropy condition satisfying discontinuous Galerkin method*, Journal of Scientific Computing, **31** (2007), pp. 127–151.

[40] C. Hu and C.-W. Shu, *Weighted essentially non-oscillatory schemes on triangular meshes*, Journal of Computational Physics, **150** (1999), pp. 97–127.

[41] C. Hu and C.-W. Shu, *A discontinuous Galerkin finite element method for Hamilton-Jacobi equations*, SIAM Journal on Scientific Computing, **21** (1999), pp. 666–690.

[42] O. Lepsky, C. Hu and C.-W. Shu, *Analysis of the discontinuous Galerkin method for Hamilton-Jacobi equations*, Applied Numerical Mathematics, **33** (2000), pp. 423–434.

[43] F. Li and C.-W. Shu, *Reinterpretation and simplified implementation of a discontinuous Galerkin method for Hamilton-Jacobi equations*, Applied Mathematics Letters, **18** (2005), pp. 1204–1209.

[44] Y.-X. Liu and C.-W. Shu, *Local discontinuous Galerkin methods for moment models in device simulations: formulation and one dimensional results*, Journal of Computational Electronics, **3** (2004), pp. 263–267.

[45] Y.-X. Liu and C.-W. Shu, *Local discontinuous Galerkin methods for moment models in device simulations: Performance assessment and two dimensional results*, Applied Numerical Mathematics, **57** (2007), pp. 629–645.

[46] Y.-X. Liu and C.-W. Shu, *Error analysis of the semi-discrete local discontinuous Galerkin method for semiconductor device simulation models*, Science China Mathematics, **53** (2010), pp. 3255–3278.

[47] L. Ji and Y. Xu, *Optimal error estimates of the local discontinuous Galerkin method for Willmore flow of graphs on Cartesian meshes*, International Journal of Numerical Analysis and Modeling, **8** (2011), pp. 252–283.

[48] L. Ji and Y. Xu, *Optimal error estimates of the local discontinuous Galerkin method for surface diffusion of graphs on Cartesian meshes*, Journal of Scientific Computing, **51** (2012), pp. 1–27.

[49] L. Ji, Y. Xu and J. Ryan, *Accuracy-enhancement of discontinuous Galerkin solutions for convection-diffusion equations in multiple-dimensions*, Mathematics of Computation, **81** (2012), pp. 1929–1950.

[50] L. Ji, Y. Xu and J. Ryan, *Negative order norm estimates for nonlinear hyperbolic conservation laws*, Journal of Scientific Computing, **54** (2013), pp. 531–548.

[51] G.-S. Jiang and C.-W. Shu, *On cell entropy inequality for discontinuous Galerkin methods*, Mathematics of Computation, **62** (1994), pp. 531–538.

[52] G.-S. Jiang and C.-W. Shu, *Efficient implementation of weighted ENO schemes*, Journal of Computational Physics, **126** (1996), pp. 202–228.

[53] C. Johnson and J. Pitkäranta, *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation*, Mathematics of Computation, **46** (1986), pp. 1–26.

[54] G. Kanschat, *Discontinuous Galerkin Methods for Viscous Flow*, Deutscher Universitätsverlag, Wiesbaden, 2007.

[55] L. Krivodonova, J. Xin, J.-F. Remacle, N. Chevaugeon and J.E. Flaherty, *Shock detection and limiting with discontinuous Galerkin methods for hyperbolic conservation laws*, Applied Numerical Mathematics, **48** (2004), pp. 323–338.

[56] P. Lesaint and P.A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical aspects of finite elements in partial differential equations, C. de Boor (ed.), Academic Press, 1974, pp. 89–145.
[57] R.J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhauser Verlag, Basel, 1990.
[58] D. Levy, C.-W. Shu and J. Yan, *Local discontinuous Galerkin methods for nonlinear dispersive equations*, Journal of Computational Physics, **196** (2004), pp. 751–772.
[59] B. Li, *Discontinuous Finite Elements in Fluid Dynamics and Heat Transfer*, Birkhauser, Basel, 2006.
[60] H. Liu and J. Yan, *The direct discontinuous Galerkin (DDG) methods for diffusion problems*, SIAM Journal on Numerical Analysis, **47** (2009), pp. 675–698.
[61] H. Liu and J. Yan, *The direct discontinuous Galerkin (DDG) methods for diffusion with interface corrections*, Communications in Computational Physics, **8** (2010), pp. 541–564.
[62] X. Liu, S. Osher and T. Chan, *Weighted essentially non-oscillatory schemes*, Journal of Computational Physics, **115** (1994), pp. 200–212.
[63] X. Meng, C.-W. Shu and B. Wu, *Superconvergence of the local discontinuous Galerkin method for linear fourth order time dependent problems in one space dimension*, IMA Journal of Numerical Analysis, **32** (2012), pp. 1294–1328.
[64] X. Meng, C.-W. Shu, Q. Zhang and B. Wu, *Superconvergence of discontinuous Galerkin method for scalar nonlinear conservation laws in one space dimension*, SIAM Journal on Numerical Analysis, **50** (2012), pp. 2336–2356.
[65] H. Mirzaee, L. Ji, J. Ryan and R.M. Kirby, *Smoothness-increasing accuracy-conserving (SIAC) post-processing for discontinuous Galerkin solutions over structured triangular meshes*, SIAM Journal on Numerical Analysis, **49** (2011), pp. 1899–1920.
[66] J.T. Oden, I. Babuvska and C.E. Baumann, *A discontinuous hp finite element method for diffusion problems*, Journal of Computational Physics, **146** (1998), pp. 491–519.
[67] T. Peterson, *A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation*, SIAM Journal on Numerical Analysis, **28** (1991), pp. 133–140.
[68] J. Qiu and C.-W. Shu, *Hermite WENO schemes and their application as limiters for Runge-Kutta discontinuous Galerkin method: one dimensional case*, Journal of Computational Physics, **193** (2003), pp. 115–135.
[69] J. Qiu and C.-W. Shu, *A comparison of troubled-cell indicators for Runge-Kutta discontinuous Galerkin methods using weighted essentially nonoscillatory limiters*, SIAM Journal on Scientific Computing, **27** (2005), pp. 995–1013.
[70] J. Qiu and C.-W. Shu, *Runge-Kutta discontinuous Galerkin method using WENO limiters*, SIAM Journal on Scientific Computing, **26** (2005), pp. 907–929.
[71] J. Qiu and C.-W. Shu, *Hermite WENO schemes and their application as limiters for Runge-Kutta discontinuous Galerkin method II: two dimensional case*, Computers & Fluids, **34** (2005), pp. 642–663.
[72] W.H. Reed and T.R. Hill, *Triangular mesh methods for the neutron transport equation*, Tech. Report LA-UR-73-479, Los Alamos Scientific Laboratory, 1973.
[73] J.-F. Remacle, J. Flaherty and M. Shephard, *An adaptive discontinuous Galerkin technique with an orthogonal basis applied to Rayleigh-Taylor flow instabilities*, SIAM Review, **45** (2003), pp. 53–72.
[74] G.R. Richter, *An optimal-order error estimate for the discontinuous Galerkin method*, Mathematics of Computation, **50** (1988), pp. 75–88.
[75] B. Rivière, *Discontinuous Galerkin methods for solving elliptic and parabolic equations. Theory and implementation*, SIAM, Philadelphia, 2008.

[76] J. Ryan and C.-W. Shu, *On a one-sided post-processing technique for the discontinuous Galerkin methods*, Methods and Applications of Analysis, **10** (2003), pp. 295–308.

[77] J. Ryan, C.-W. Shu and H. Atkins, *Extension of a postprocessing technique for the discontinuous Galerkin method for hyperbolic equations with application to an aeroacoustic problem*, SIAM Journal on Scientific Computing, **26** (2005), pp. 821–843.

[78] J. Shi, C. Hu and C.-W. Shu, *A technique of treating negative weights in WENO schemes*, Journal of Computational Physics, **175** (2002), pp. 108–127.

[79] C.-W. Shu, *TVB uniformly high-order schemes for conservation laws*, Mathematics of Computation, **49** (1987), pp. 105–121.

[80] C.-W. Shu, *Discontinuous Galerkin methods: general approach and stability*, Numerical Solutions of Partial Differential Equations, S. Bertoluzza, S. Falletta, G. Russo and C.-W. Shu, Advanced Courses in Mathematics CRM Barcelona, Birkhäuser, Basel, 2009, pp. 149–201.

[81] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, Journal of Computational Physics, **77** (1988), pp. 439–471.

[82] M. Steffan, S. Curtis, R.M. Kirby and J. Ryan, *Investigation of smoothness enhancing accuracy-conserving filters for improving streamline integration through discontinuous fields*, IEEE-TVCG, **14** (2008), pp. 680–692.

[83] C. Wang, X. Zhang, C.-W. Shu and J. Ning, *Robust high order discontinuous Galerkin schemes for two-dimensional gaseous detonations*, Journal of Computational Physics, **231** (2012), pp. 653–665.

[84] M.F. Wheeler, *An elliptic collocation-finite element method with interior penalties*, SIAM Journal on Numerical Analysis, **15** (1978), pp. 152–161.

[85] Y. Xia, Y. Xu and C.-W. Shu, *Local discontinuous Galerkin methods for the Cahn-Hilliard type equations*, Journal of Computational Physics, **227** (2007), pp. 472–491.

[86] Y. Xia, Y. Xu and C.-W. Shu, *Application of the local discontinuous Galerkin method for the Allen-Cahn/Cahn-Hilliard system*, Communications in Computational Physics, **5** (2009), pp. 821–835.

[87] Y. Xia, Y. Xu and C.-W. Shu, *Local discontinuous Galerkin methods for the generalized Zakharov system*, Journal of Computational Physics, **229** (2010), pp. 1238–1259.

[88] Y. Xing, X. Zhang and C.-W. Shu, *Positivity preserving high order well balanced discontinuous Galerkin methods for the shallow water equations*, Advances in Water Resources, **33** (2010), pp. 1476–1493.

[89] T. Xiong, C.-W. Shu and M. Zhang, *A priori error estimates for semi-discrete discontinuous Galerkin methods solving nonlinear Hamilton-Jacobi equations with smooth solutions*, International Journal of Numerical Analysis and Modeling, **10** (2013), pp. 154–177.

[90] Y. Xu and C.-W. Shu, *Local discontinuous Galerkin methods for three classes of nonlinear wave equations*, Journal of Computational Mathematics, **22** (2004), pp. 250–274.

[91] Y. Xu and C.-W. Shu, *Local discontinuous Galerkin methods for nonlinear Schrödinger equations*, Journal of Computational Physics, **205** (2005), pp. 72–97.

[92] Y. Xu and C.-W. Shu, *Local discontinuous Galerkin methods for two classes of two dimensional nonlinear wave equations*, Physica D, **208** (2005), pp. 21–58.

[93] Y. Xu and C.-W. Shu, *Local discontinuous Galerkin methods for the Kuramoto-Sivashinsky equations and the Ito-type coupled KdV equations*, Computer Methods in Applied Mechanics and Engineering, **195** (2006), pp. 3430–3447.

[94] Y. Xu and C.-W. Shu, *Error estimates of the semi-discrete local discontinuous Galerkin method for nonlinear convection-diffusion and KdV equations*, Computer Methods in Applied Mechanics and Engineering, **196** (2007), pp. 3805–3822.

[95] Y. Xu and C.-W. Shu, *A local discontinuous Galerkin method for the Camassa-Holm equation*, SIAM Journal on Numerical Analysis, **46** (2008), pp. 1998–2021.

[96] Y. Xu and C.-W. Shu, *Local discontinuous Galerkin method for the Hunter-Saxton equation and its zero-viscosity and zero-dispersion limit*, SIAM Journal on Scientific Computing, **31** (2008), pp. 1249–1268.

[97] Y. Xu and C.-W. Shu, *Local discontinuous Galerkin method for surface diffusion and Willmore flow of graphs*, Journal of Scientific Computing, **40** (2009), pp. 375–390.

[98] Y. Xu and C.-W. Shu, *Dissipative numerical methods for the Hunter-Saxton equation*, Journal of Computational Mathematics, **28** (2010), pp. 606–620.

[99] Y. Xu and C.-W. Shu, *Local discontinuous Galerkin methods for high-order time-dependent partial differential equations*, Communications in Computational Physics, **7** (2010), pp. 1–46.

[100] Y. Xu and C.-W. Shu, *Local discontinuous Galerkin methods for the Degasperis-Procesi equation*, Communications in Computational Physics, **10** (2011), pp. 474–508.

[101] Y. Xu and C.-W. Shu, *Optimal error estimates of the semi-discrete local discontinuous Galerkin methods for high order wave equations*, SIAM Journal on Numerical Analysis, **50** (2012), pp. 79–104.

[102] J. Yan and S. Osher, *A local discontinuous Galerkin method for directly solving HamiltonJacobi equations*, Journal of Computational Physics, **230** (2011), pp. 232–244.

[103] J. Yan and C.-W. Shu, *A local discontinuous Galerkin method for KdV type equations*, SIAM Journal on Numerical Analysis, **40** (2002), pp. 769–791.

[104] J. Yan and C.-W. Shu, *Local discontinuous Galerkin methods for partial differential equations with higher order derivatives*, Journal of Scientific Computing, **17** (2002), pp. 27–47.

[105] Y. Yang and C.-W. Shu, *Analysis of optimal superconvergence of discontinuous Galerkin method for linear hyperbolic equations*, SIAM Journal on Numerical Analysis, **50** (2012), pp. 3110–3133.

[106] Y. Yang and C.-W. Shu, *Discontinuous Galerkin method for hyperbolic equations involving δ-singularities: negative-order norm error estimates and applications*, Numerische Mathematik, **124** (2013), pp. 753–781.

[107] M. Zhang and C.-W. Shu, *An analysis of three different formulations of the discontinuous Galerkin method for diffusion equations*, Mathematical Models and Methods in Applied Sciences (M^3AS), **13** (2003), pp. 395–413.

[108] Q. Zhang and C.-W. Shu, *Error estimates to smooth solutions of Runge-Kutta discontinuous Galerkin methods for scalar conservation laws*, SIAM Journal on Numerical Analysis, **42** (2004), pp. 641–666.

[109] Q. Zhang and C.-W. Shu, *Error estimates to smooth solutions of Runge-Kutta discontinuous Galerkin method for symmetrizable systems of conservation laws*, SIAM Journal on Numerical Analysis, **44** (2006), pp. 1703–1720.

[110] Q. Zhang and C.-W. Shu, *Stability analysis and a priori error estimates to the third order explicit Runge-Kutta discontinuous Galerkin method for scalar conservation laws*, SIAM Journal on Numerical Analysis, **48** (2010), pp. 1038–1063.

[111] X. Zhang and C.-W. Shu, *On maximum-principle-satisfying high order schemes for scalar conservation laws*, Journal of Computational Physics, **229** (2010), pp. 3091–3120.

[112] X. Zhang and C.-W. Shu, *On positivity preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes*, Journal of Computational Physics, **229** (2010), pp. 8918–8934.

[113] X. Zhang and C.-W. Shu, *Positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations with source terms*, Journal of Computational Physics, **230** (2011), pp. 1238–1248.

[114] X. Zhang and C.-W. Shu, *Maximum-principle-satisfying and positivity-preserving high order schemes for conservation laws: Survey and new developments*, Proceedings of the Royal Society A, **467** (2011), pp. 2752–2776.

[115] X. Zhang, Y. Xia and C.-W. Shu, *Maximum-principle-satisfying and positivity-preserving high order discontinuous Galerkin schemes for conservation laws on triangular meshes*, Journal of Scientific Computing, **50** (2012), pp. 29–62.

[116] Y.-T. Zhang and C.-W. Shu, *Third order WENO scheme on three dimensional tetrahedral meshes*, Communications in Computational Physics, **5** (2009), pp. 836–848.

[117] X. Zhong and C.-W. Shu, *Numerical resolution of discontinuous Galerkin methods for time dependent wave equations*, Computer Methods in Applied Mechanics and Engineering, **200** (2011), pp. 2814–2827.

[118] X. Zhong and C.-W. Shu, *A simple weighted essentially nonoscillatory limiter for Runge-Kutta discontinuous Galerkin methods*, Journal of Computational Physics, **232** (2012), pp. 397–415.

[119] J. Zhu, J.-X. Qiu, C.-W. Shu and M. Dumbser, *Runge-Kutta discontinuous Galerkin method using WENO limiters II: unstructured meshes*, Journal of Computational Physics, **227** (2008), pp. 4330–4353.

[120] J. Zhu, X. Zhong, C.-W. Shu and J.-X. Qiu, *Runge-Kutta discontinuous Galerkin method using a new type of WENO limiters on unstructured mesh*, Journal of Computational Physics, Numerische Mathematik, **124** (2013), pp. 753–781.

ADAPTIVITY AND ERROR ESTIMATION FOR DISCONTINUOUS GALERKIN METHODS

SLIMANE ADJERID* AND MAHBOUB BACCOUCH†

Abstract. We test the a posteriori error estimates of discontinuous Galerkin (DG) discretization errors (Adjerid and Baccouch, J. Sci. Comput. 33(1):75–113, 2007; Adjerid and Baccouch, J. Sci. Comput. 38(1):15–49, 2008; Adjerid and Baccouch Comput. Methods Appl. Mech. Eng. 200:162–177, 2011) for hyperbolic problems on adaptively refined unstructured triangular meshes. A local error analysis allows us to construct asymptotically correct a posteriori error estimates by solving local hyperbolic problems on each element. The Taylor-expansion-based error analysis (Adjerid and Baccouch, J. Sci. Comput. 33(1):75–113, 2007; Adjerid and Baccouch, J. Sci. Comput. 38(1):15–49, 2008; Adjerid and Baccouch Comput. Methods Appl. Mech. Eng. 200:162–177, 2011) does not apply near discontinuities and shocks and lead to inaccurate estimates under uniform mesh refinement. Here, we present several computational results obtained from adaptive refinement computations that suggest that even in the presence of shocks our error estimates converge to the true error under adaptive mesh refinement. We also show the performance of several adaptive strategies for hyperbolic problems with discontinuous solutions.

Key words. Adaptive discontinuous Galerkin method, Hyperbolic problems, A posteriori error estimation, Unstructured meshes

AMS(MOS) subject classifications. Primary 65N30, 65N50.

1. Introduction.

The DG method was first developed for the neutron equation [20]. Since then, DG methods have been used to solve hyperbolic [11, 13–15, 17], parabolic [16], and elliptic [10] partial differential equations.

The DG methods are a family of locally conservative, stable, and high-order accurate methods that are easily coupled with other well-known methods and are well suited to adaptive strategies. For these reasons, they have attracted the attention of many researchers working in computational mechanics, computational mathematics, and computer science. They provide an appealing approach to address problems having discontinuities, such as those arising in hyperbolic conservation laws. The DG method does not require the approximate solutions to be continuous across element boundaries, it instead involves a flux term to account for the discontinuities. For a more complete list of citations on DG methods and its applications, consult [12]. A main advantage of using discontinuous finite element basis is to simplify adaptive p- and h-refinement with hanging nodes.

The DG method has a simple communication pattern between elements with a common face that makes it useful for parallel computation.

*Department of Mathematics, Virginia Tech, Blacksburg, VA 24061, USA, adjerids@vt.edu

†Department of Mathematics, University of Nebraska, Omaha, NE 68182, USA, mbaccouch@unomaha.edu

X. Feng et al. (eds.), *Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations*, The IMA Volumes in Mathematics and its Applications 157, DOI 10.1007/978-3-319-01818-8_3,

Furthermore, it can handle problems with complex geometries to high order. Regardless of the type of DG method, we need to know how well our computed solution approximates the exact solution. In practice, the exact solution of the problem is not available and a method to estimate the discretization error is needed. For these reasons, a posteriori error estimates have been developed for DG methods and provide some initial guidance for deciding on the degree of the approximation and the size of the mesh that guarantee a prescribed level of accuracy. Furthermore, error estimates may be used to guide *hp*-adaptive refinement.

The first superconvergence result for DG solutions of hyperbolic partial differential equations appeared in Adjerid et al. [4]. The authors showed that DG solutions of one-dimensional linear and nonlinear hyperbolic problems using p-degree polynomial approximations exhibit an $O(h^{p+2})$ superconvergence rate at the roots of $(p+1)$-degree Radau polynomial. This lead to the conclusion that the leading term of the DG error on each element is proportional to $(p+1)$-degree Radau polynomial which was used to construct asymptotically correct a posteriori error estimates. They further established a strong $O(h^{2p+1})$ superconvergence at the downwind end of every element. Later, Krivodonova and Flaherty [19] showed that the leading term of the local discretization error on triangles having one *outflow* edge is spanned by a suboptimal set of orthogonal polynomials of degree p and $p+1$ and computed DG error estimates by solving local problems involving numerical fluxes, thus requiring information from neighboring *inflow* elements. Adjerid and Massey [5] extended these results to multi-dimensional problems using rectangular meshes and presented an error analysis for linear and nonlinear hyperbolic problems. They showed that the leading term in the true local error is spanned by two $(p+1)$-degree Radau polynomials in the x and y directions, respectively. They further discovered that a p-degree discontinuous finite element solution exhibits an $O(h^{p+2})$ superconvergence at Radau points obtained as a tensor product of the roots of Radau polynomial of degree $p+1$. Using these results, they were able to compute asymptotically exact a posteriori error estimates for linear and nonlinear hyperbolic problems on Cartesian meshes.

Adjerid and Baccouch [1, 2] investigated the superconvergence properties of discontinuous Galerkin solutions of a scalar first-order hyperbolic problem on triangular meshes. They presented a detailed discussion on the superconvergence properties versus the choice of finite element polynomial spaces. First, they classified triangular elements into three types: (i) type I with one *inflow* edge and two *outflow* edges, (ii) type II with two *inflow* edges and one *outflow* edge, and (iii) type III with one *inflow* edge, one *outflow* edge, one edge parallel to the characteristics. Through computations, they showed that the local superconvergence results [1] extend to global DG solutions on general meshes with a corrected *inflow* boundary condition. In particular, they showed that the discontinuous finite element solution is $O(h^{p+2})$ superconvergent at the Legendre points on the outflow edge for triangles having one outflow edge. For triangles having two outflow

edges the finite element error is $O(h^{p+2})$ superconvergent at the end points of the inflow edge.

Adjerid and Weinhart [7–9] studied the asymptotic behavior of the local DG error for multi-dimensional first-order linear symmetric and symmetrizable hyperbolic systems of partial differential equations. They performed a local error analysis by writing the local error as a series and showing that its leading term can be expressed as a linear combination of Legendre polynomials of degree p and $p+1$. They were able to compute efficient and asymptotically exact estimates of the discontinuous finite element error.

In this manuscript we consider the modified discontinuous Galerkin method [2] with a corrected inflow flux and an enriched polynomial space $\mathcal{U}_p$ with adaptive mesh refinement. We consider several mesh refinement strategies guided by both discretization error estimates and local residuals. Since $\mathcal{L}^2$ a posteriori error estimates based on Taylor expansions fail to be asymptotically exact under uniform mesh refinement in the presence of shocks [3, 5, 6], we present several numerical results which suggest that such error estimates converge to the true errors under adaptive mesh refinement on general unstructured triangular meshes in the presence of discontinuities. Numerical results further suggest that using local residuals to guide the adaptive mesh refinement yield more efficient algorithms when compared to using the error estimate itself in the presence of discontinuities. Thus, we recommend an adaptive strategy that combines the local residuals or any other explicit estimators to guide mesh refinement and the proposed error estimate to assess solution accuracy and terminate the adaptive refinement process.

This paper is organized as follows: In Sect. 2 we state the modified DG formulation for linear and nonlinear hyperbolic problem and present our a posteriori error estimation procedures for linear and nonlinear problems. In Sect. 3 we describe several adaptive mesh refinement strategies that will be used to test the performance of our error estimates. Finally, in Sect. 4 we present numerical results for several linear and nonlinear hyperbolic problems with discontinuous solutions and conclude with a few remarks in Sect. 5.

2. Discontinuous Galerkin Formulation and Error Estimation. In this section we present an adaptive modified discontinuous Galerkin method [1–3] combined with a posteriori error estimation procedure. In addition to being used to steer the adaptive process, the a posteriori error estimate is also used to correct the numerical flux needed to compute the DG solution on downwind elements. This modified DG method maintains the structure of the local discretization error on each element of the mesh which makes the error estimation both efficient and accurate.

In order to further simplify the error estimation procedure we use the augmented polynomial space $\mathcal{U}_p$ given by

$$\mathcal{U}_p = \mathcal{P}_p \cup \text{span}\{x^{p+1}, y^{p+1}\},\ p \geq 1,$$

where

$$\mathcal{P}_k = \left\{ q \,|\, q = \sum_{m=0}^{k} \sum_{i=0}^{m} c_i^m x^i y^{m-i} \right\}, \qquad k = 0, 1, \ldots, p. \tag{2.1a}$$

The finite element space $\mathcal{U}_p$ is suboptimal, *i.e.*, it contains $p+1$-degree terms that do not contribute to global convergence rate but simplifies the a posteriori error estimation procedures described later in this manuscript.

We consider a reference triangle Δ_0 defined by the vertices $(0,0)$, $(1,0)$, and $(0,1)$ and define the following orthogonal polynomials [18]

$$\varphi_k^l(\xi,\eta) = 2^k \hat{L}_k\left(\frac{2\xi}{1-\eta} - 1\right)(1-\eta)^k \hat{P}_l^{2k+1,0}(2\eta-1),\ k,l \geq 0\ k+l = p \geq 0, \tag{2.2a}$$

where $\hat{P}_n^{\alpha,\beta}(x),\ -1 \leq x \leq 1$, is the Jacobi polynomial

$$\hat{P}_n^{\alpha,\beta}(x) = \frac{(-1)^n}{2^n n!}(1-x)^{-\alpha}(1+x)^{-\beta}\frac{d^n}{dx^n}[(1-x)^{\alpha+n}(1+x)^{\beta+n}], \quad \alpha,\beta > -1, \tag{2.2b}$$

and $\hat{L}_n(x) = \hat{P}_n^{0,0}(x),\ -1 \leq x \leq 1$ is the nth-degree Legendre polynomial.

We note that these polynomials satisfy the $\mathcal{L}^2$ orthogonality

$$\int_0^1 \int_0^{1-\eta} \varphi_k^l \varphi_p^q d\xi d\eta = c_{kp}^{lq} \delta_{kp} \delta_{lq}. \tag{2.3}$$

Radau polynomials are defined by

$$\hat{R}_{p+1}(x) = (1-x)\hat{P}_p^{1,0}(x) = C(\hat{L}_{p+1}(x) - \hat{L}_p(x)),\ -1 \leq x \leq 1. \tag{2.4}$$

We drop the hat and let L_p, $P_p^{\alpha,\beta}$ and R_p, respectively, denote the Jacobi, Legendre and Radau polynomials shifted to $[0,1]$.

Let us note that U on the physical element is in the modified polynomial space

$$\mathcal{U}_p = \mathcal{P}_p + \text{span}\{\xi(x,y)^{p+1}, \eta(x,y)^{p+1}\},$$

where $(x,y) \to (\xi(x,y), \eta(x,y))$ is the standard affine mapping from the physical element Δ to the reference element Δ_0.

First, let us consider a divergence-free linear stationary hyperbolic problem on an open bounded convex polygonal domain $\Omega \subseteq R^2$. Let $\boldsymbol{a} = [a_1(x,y), a_2(x,y)]^T$ denote a nonzero velocity vector. If $\boldsymbol{n}$ denotes the outward unit normal vector, the domain boundary $\partial\Omega = \partial\Omega^+ \cup \partial\Omega^- \cup \partial\Omega^0$,

where the inflow, outflow, and characteristic boundaries, respectively, are $\partial\Omega^- = \{(x,y) \in \partial\Omega \mid \boldsymbol{a}\cdot\boldsymbol{n} < 0\}$, $\partial\Omega^+ = \{(x,y) \in \partial\Omega \mid \boldsymbol{a}\cdot\boldsymbol{n} > 0\}$, and $\partial\Omega^0 = \{(x,y) \in \partial\Omega \mid \boldsymbol{a}\cdot\boldsymbol{n} = 0\}$.
Let $u(x,y)$ denote a smooth function on Ω and consider the following hyperbolic boundary value problem

$$L(u) = \boldsymbol{a}.\nabla u + cu = f, \quad (x,y) \in \Omega = (0,1)^2, \tag{2.5a}$$

$$\nabla.\boldsymbol{a} = \frac{\partial a_1}{\partial x} + \frac{\partial a_2}{\partial y} = 0, \tag{2.5b}$$

subject to the boundary conditions

$$u(x,0) = g_0(x), \quad u(0,y) = g_1(y), \tag{2.5c}$$

where the functions $\boldsymbol{a}(x,y)$, $c(x,y)$, $f(x,y)$, $g_0(x)$, and $g_1(y)$ are selected such that the exact solution $u(x,y) \in C^\infty(\Omega)$.

In order to obtain the weak discontinuous Galerkin formulation for (2.5), we partition the domain Ω into N triangular elements Δ_j, $j = 1,\ldots,N$, such that on every edge $\boldsymbol{a}\cdot\boldsymbol{n}$ does not change sign. Thus, every edge is either *inflow*, *outflow*, or *characteristic*, respectively, if $\boldsymbol{a}\cdot\boldsymbol{n} < 0$, $\boldsymbol{a}\cdot\boldsymbol{n} > 0$ or $\boldsymbol{a}\cdot\boldsymbol{n} = 0$. Using this mesh orientation, a triangle can be classified into type I having one *inflow* edge and 2 *outflow* edges, type II having two *outflow* and one *inflow* edges, or type III having one *inflow*, one *outflow* and one *characteristic* edges. The problem is solved on each element starting from the upwind elements and proceeding to the neighboring elements in the downwind direction, i.e., we order the elements such that the inflow boundary Γ_j^- of an element Δ_j is contained in the inflow boundary $\partial\Omega^-$ of the domain or in the outflow boundary Γ_i^+ of Δ_i, $i < j$.

In the remainder of this paper we omit the element index and refer to an arbitrary element by Δ whenever confusion is unlikely. Note that Γ^+ and Γ^-, respectively, denote the outflow and inflow boundaries of Δ.

Thus, our discontinuous Galerkin method consists of finding $U \in \mathcal{U}_p$ such that

$$\int_{\Gamma^-} \boldsymbol{a}\cdot\boldsymbol{n}\hat{U}V\,ds + \int_{\Gamma^+} \boldsymbol{a}\cdot\boldsymbol{n}UV\,ds - \iint_\Delta (\boldsymbol{a}\cdot\nabla V - cV)U\,dxdy = \iint_\Delta fV\,dxdy, \quad \forall\, V \in \mathcal{U}_p. \tag{2.6a}$$

In order to complete the definition of the DG method we need to select the corrected upwind numerical flux $\hat{U}$ on Γ^- as

$$\hat{U} = \begin{cases} u, & \text{on } \Gamma^- \bigcap \partial\Omega^- \\ U^- + E^-, & \text{elsewhere,} \end{cases} \tag{2.6b}$$

where U^- and E^-, respectively, are the limit of U and E from the *inflow* element sharing Γ^- , i.e., if $(x,y) \in \Gamma^-$, then

$$U^-(x,y) = \lim_{s\to 0^+} U((x,y) + s\boldsymbol{n}), E^-(x,y) = \lim_{s\to 0^+} E((x,y) + s\boldsymbol{n}). \tag{2.6c}$$

Here, E is an a posteriori error estimate that will be defined by following [2, 3] to write the leading term of the local DG error on each triangle as a linear combination of the following p error basis functions

$$(u-U)(x,y) \approx E(x,y) = \sum_{i=1}^{p} d^i \chi^i_{p+1-i}(\xi(x,y),\eta(x,y)), \tag{2.7a}$$

where χ_j are given in [2]. For the sake of completeness we include them in Tables 1 and 2 for types I and II and $[\alpha,\beta]$ as given in (2.11) while for type III the leading term of the error can be written as

$$Q_{p+1} = \sum_{\substack{i,j\geq 0\\ i+j=p}} c_i^j\,(1-\xi-\eta)P_i^{2j+2,0}(2\eta-1)(1-\eta)^j P_j^{1,0}\left(\frac{2\xi}{1-\eta}-1\right), \tag{2.7b}$$

with

$$c_0^p = \frac{1}{p+1}\sum_{i=1}^{p}(-1)^{i+1}(p+1-i)c_i^{p-i}. \tag{2.7c}$$

After computing the finite element solution U on an element Δ, we compute an error estimate by solving the problem for $i = 1, 2, \ldots, p$,

$$\iint_{\Delta} \left(\boldsymbol{a}\cdot\nabla(U+E) + c(U+E)\right)\chi^i_{p+1-i}dxdy = \iint_{\Delta} f\chi^i_{p+1-i}dxdy. \tag{2.7d}$$

Next we consider nonlinear problems of the form

$$L(u) = \nabla\cdot\mathbf{F}(u) = h(u)_x + g(u)_y = f(x,y), \quad (x,y)\in\Omega \;=\; (0,1)^2, \tag{2.8a}$$

subject to the boundary conditions

$$u(x,0) = g_0(x), \quad u(0,y) = g_1(y). \tag{2.8b}$$

In our analysis [3] we assume $\mathbf{F}:\mathbb{R}\to\mathbb{R}^2$, $u:\mathbb{R}^2\to\mathbb{R}$, f, g_0 and g_1 to be analytic functions such that $g'(u) > 0$ and $h'(u) > 0$ over the domain Ω. The *inflow*, *outflow*, and *characteristic* boundaries, respectively, are defined by $\partial\Omega^- = \{(x,y)\in\partial\Omega \mid \mathbf{F}'(u)\cdot\boldsymbol{n} = [h'(u),g'(u)]^t\cdot\boldsymbol{n} < 0\}$, $\partial\Omega^+ = \{(x,y)\in\partial\Omega \mid \mathbf{F}'(u)\cdot\boldsymbol{n} > 0\}$, and $\partial\Omega^0 = \{(x,y)\in\partial\Omega \mid \mathbf{F}'(u)\cdot\boldsymbol{n} = 0\}$, such that $\partial\Omega = \partial\Omega^-\cup\partial\Omega^+\cup\partial\Omega^0$ and $\boldsymbol{n}$ is the outward unit normal to $\partial\Omega$. We further assume that the unstructured triangular mesh is such that $\mathbf{F}'(u).\boldsymbol{n}$ does not change sign on all edges, i.e., every edge is either *inflow*, *outflow*, or *characteristic*.

The discrete DG formulation consists of determining $U\in\mathcal{U}_p$ such that

$$\begin{aligned}\int_{\Gamma^-}\boldsymbol{n}\cdot\mathbf{F}(\hat{U})Vds + \int_{\Gamma^+}\boldsymbol{n}\cdot\mathbf{F}(U)Vds - \iint_{\Delta}\mathbf{F}(U)\cdot\nabla V dxdy\\ = \iint_{\Delta} fVdxdy, \quad \forall\, V\in\mathcal{U}_p,\end{aligned} \tag{2.9a}$$

TABLE 1
Error basis functions for the spaces $\mathcal{U}_p$ for $p = 1$ to 3 on the reference triangle of type I where $s = \alpha/\beta$

$p = 1$
$\chi_1^1 = (12(\varphi_1^0 + \varphi_1^1)s^2 + 2(10\varphi_0^0 - 2\varphi_1^1 - 3\varphi_1^0 - 2\varphi_0^2 + 12\varphi_1^1 + 5\varphi_2^0)s + 12\varphi_1^1 + 6\varphi_1^0 - 8\varphi_0^2 + 6\varphi_1^1 + 5\varphi_2^0)/10s + 5$
$p = 2$
$\chi_2^1 = (75(2\varphi_1^1 + \varphi_2^0 + 2\varphi_1^2 + \varphi_2^1)s^4 + 5(28\varphi_0^0 + 28\varphi_0^1 - 42\varphi_1^0 - 12\varphi_0^2 + 54\varphi_1^1 + 51\varphi_2^0 - 12\varphi_0^3 + 96\varphi_1^2 + 51\varphi_2^1)s^3 + (-140\varphi_0^0 + 532\varphi_0^1 + 126\varphi_1^0 - 68\varphi_0^2 + 36\varphi_1^1 + 140\varphi_2^0 - 180\varphi_0^3 + 540\varphi_1^2 + 315\varphi_2^1)s^2 + (-140\varphi_0^0 + 280\varphi_0^1 + 168\varphi_1^0 + 100\varphi_0^2 - 12\varphi_1^1 - 10\varphi_2^0 - 180\varphi_0^3 + 240\varphi_1^2 + 165\varphi_2^1)s + 5(16\varphi_0^2 + 6\varphi_1^1 - \varphi_2^0 - 12\varphi_0^3 + 6\varphi_1^2 + 6\varphi_2^1))/15(s+1)^3(5s+2)$ $\chi_1^2 = (-240(\varphi_1^1 + \varphi_1^2)s^5 + (-448\varphi_0^0 - 112\varphi_0^1 + 840\varphi_1^0 + 128\varphi_0^2 - 780\varphi_1^1 - 350\varphi_2^0 + 72\varphi_0^3 - 1200\varphi_1^2 + 105\varphi_3^0)s^4 - 21(24\varphi_0^0 + 56\varphi_0^1 - 68\varphi_1^0 - 24\varphi_0^2 + 12\varphi_1^1 + 30\varphi_2^0 - 16\varphi_0^3 + 108\varphi_1^2 - 17\varphi_3^0)s^3 + (616\varphi_0^0 - 2072\varphi_0^1 - 84\varphi_1^0 + 408\varphi_0^2 + 936\varphi_1^1 + 576\varphi_0^3 - 2004\varphi_1^2 + 441\varphi_3^0)s^2 + (392\varphi_0^0 - 784\varphi_0^1 - 336\varphi_1^0 - 184\varphi_0^2 + 624\varphi_1^1 + 280\varphi_2^0 + 432\varphi_0^3 - 804\varphi_1^2 + 231\varphi_3^0)s + 2(-80\varphi_0^2 + 30\varphi_1^1 + 35\varphi_2^0 + 60\varphi_0^3 - 54\varphi_1^2 + 21\varphi_3^0))/21(s+1)^3(5s+2)$
$p = 3$
$\chi_3^1 = (420(2\varphi_1^3 + \varphi_2^2 + 2\varphi_1^2 + \varphi_2^1)s^5 + 4(84\varphi_0^0 + 228\varphi_0^1 - 80\varphi_0^4 + 870\varphi_1^3 + 455\varphi_2^2 - 54\varphi_1^0 + 108\varphi_0^2 - 324\varphi_1^1 - 116\varphi_0^3 + 600\varphi_1^2 + 455\varphi_2^1)s^4 + (-2352\varphi_0^0 + 2976\varphi_0^1 - 1280\varphi_0^4 + 5520\varphi_1^3 + 3080\varphi_2^2 + 4608\varphi_1^0 + 2736\varphi_0^2 - 2592\varphi_1^1 - 2520\varphi_2^0 - 1352\varphi_0^3 + 1596\varphi_1^2 + 2450\varphi_2^1 + 441\varphi_3^0)s^3 + 6(8\varphi_0^0 - 232\varphi_0^1 - 320\varphi_0^4 + 680\varphi_1^3 + 420\varphi_2^2 + 156\varphi_1^0 + 888\varphi_0^2 + 216\varphi_1^1 - 240\varphi_2^0 - 152\varphi_0^3 - 112\varphi_1^2 + 120\varphi_2^1 + 63\varphi_3^0)s^2 + (720\varphi_0^0 - 1440\varphi_0^1 - 1280\varphi_0^4 + 1320\varphi_1^3 + 980\varphi_2^2 - 864\varphi_1^0 + 2160\varphi_0^2 + 1296\varphi_1^1 + 520\varphi_2^0 - 372\varphi_0^3 - 190\varphi_1^2 + 63\varphi_3^0)s - 20(16\varphi_0^4 - 6\varphi_1^3 - 7\varphi_2^2 - 20\varphi_0^3 - 6\varphi_1^2 + 2\varphi_2^1))/140(s+1)^4(3s+1)$ $\chi_2^2 = (-168(4\varphi_1^3 - \varphi_3^1 + 4\varphi_1^2 - \varphi_3^0)s^5 - 4(168\varphi_0^0 + 240\varphi_0^1 - 40\varphi_0^4 + 708\varphi_1^3 - 182\varphi_3^1 - 216\varphi_1^0 - 486\varphi_1^1 + 135\varphi_2^0 - 112\varphi_0^3 + 438\varphi_1^2 + 135\varphi_2^1 - 182\varphi_3^0)s^4 + 2(1512\varphi_0^0 - 2088\varphi_0^1 + 320\varphi_0^4 - 2304\varphi_1^3 + 616\varphi_3^1 - 2916\varphi_1^0 - 648\varphi_0^2 + 2484\varphi_1^1 + 1350\varphi_2^0 + 752\varphi_0^3 - 180\varphi_1^2 - 540\varphi_2^1 + 175\varphi_3^0)s^3 + 3(48\varphi_0^0 + 288\varphi_0^1 + 320\varphi_0^4 - 1184\varphi_1^3 + 336\varphi_3^1 - 576\varphi_1^0 - 1392\varphi_0^2 + 144\varphi_1^1 + 600\varphi_2^0 + 488\varphi_0^3 + 724\varphi_1^2 + 30\varphi_2^1 - 21\varphi_3^0)s^2 + (-528\varphi_0^0 + 1056\varphi_0^1 + 640\varphi_0^4 - 1248\varphi_1^3 + 392\varphi_3^1 + 432\varphi_1^0 - 1584\varphi_0^2 - 648\varphi_1^1 + 180\varphi_2^0 - 8\varphi_0^3 + 1236\varphi_1^2 + 450\varphi_2^1 - 49\varphi_3^0)s + 160\varphi_0^4 - 144\varphi_1^3 + 56\varphi_3^1 - 200\varphi_0^3 + 108\varphi_1^2 + 90\varphi_2^1 - 7\varphi_3^0)/56(s+1)^4(3s+1)$ $\chi_1^3 = (3360(\varphi_1^3 + \varphi_1^2)s^6 + 14(384\varphi_0^0 + 240\varphi_0^1 - 64\varphi_0^4 + 1440\varphi_1^3 + 54\varphi_4^0 - 648\varphi_1^0 - 648\varphi_1^1 + 540\varphi_2^0 - 136\varphi_0^3 + 1116\varphi_1^2 + 270\varphi_2^1 - 189\varphi_3^0)s^5 + (-6720\varphi_0^0 + 29280\varphi_0^1 - 5216\varphi_0^4 + 48240\varphi_1^3 + 3276\varphi_4^0 + 13968\varphi_1^0 + 4320\varphi_0^2 - 44712\varphi_1^1 + 3780\varphi_2^0 - 9680\varphi_0^3 + 22320\varphi_1^2 + 17640\varphi_2^1 - 4410\varphi_3^0)s^4 - 6(4984\varphi_0^0 - 6248\varphi_0^1 + 1984\varphi_0^4 - 9760\varphi_1^3 - 924\varphi_4^0 - 10452\varphi_1^0 - 4728\varphi_0^2 + 9108\varphi_1^1 + 3990\varphi_2^0 + 2944\varphi_0^3 - 196\varphi_1^2 - 4200\varphi_2^1 - 441\varphi_3^0)s^3 + (144\varphi_0^0 - 11232\varphi_1^0 - 13376\varphi_0^4 + 37440\varphi_1^3 + 4536\varphi_4^0 + 16416\varphi_1^0 + 44208\varphi_0^2 - 5184\varphi_1^1 - 14400\varphi_2^0 - 10856\varphi_0^3 - 18900\varphi_1^2 + 11610\varphi_2^1 + 5985\varphi_3^0)s^2 + (4848\varphi_0^0 - 9696\varphi_1^0 - 7424\varphi_0^4 + 11520\varphi_1^3 + 1764\varphi_4^0 - 3600\varphi_1^0 + 14544\varphi_0^2 + 5400\varphi_1^1 - 1260\varphi_2^0 + 2008\varphi_0^3 - 9180\varphi_1^2 + 1530\varphi_2^1 + 2835\varphi_3^0)s + 3(-544\varphi_0^4 + 400\varphi_1^3 + 84\varphi_4^0 + 680\varphi_0^3 - 188\varphi_1^2 + 30\varphi_2^1 + 147\varphi_3^0))/252(s+1)^4(3s+1)$

where $\hat{U}$ is as defined in (2.6b). We estimate the error by solving the linearized problem

$$\iint_\Delta [h'(U), g'(U)]^T \cdot \nabla(U+E)\chi_{p+1-i}^i dxdy = \iint_\Delta f\chi_{p+1-i}^i dxdy, \quad i = 1, \ldots, p. \tag{2.9b}$$

TABLE 2
Error basis functions for the spaces $\mathcal{U}_p$ for $p = 1$ to 4 on the reference element of type II, $s = \alpha/\beta$

$p = 1$
$\chi_1^1 = ((-24\varphi_0^1 - 12\varphi_1^0 + 16\varphi_0^2 + 18\varphi_1^1 + 5\varphi_2^0)s^2 - 3(8\varphi_0^1 + 12\varphi_1^0 + 8\varphi_0^2 - 18\varphi_1^1 - 5\varphi_2^0)s - 24\varphi_1^0 + 6\varphi_1^1 + 15\varphi_2^0)5(s^2 + 3s + 3)$
$p = 2$
$\chi_2^1 = -((32\varphi_0^2 + 18\varphi_1^1 + \varphi_2^0 - 24\varphi_0^3 - 24\varphi_2^1 - 6\varphi_3^0)s^2 + 4(8\varphi_0^2 + 18\varphi_1^1 + \varphi_2^0 + 8\varphi_0^3 - 24\varphi_2^1 - 6\varphi_3^0)s + 6(10\varphi_1^1 + \varphi_2^0 - 4\varphi_2^1 - 6\varphi_3^0))6(s^2 + 4s + 6)$ $\chi_1^2 = ((320\varphi_0^2 + 90\varphi_1^1 - 35\varphi_2^0 - 240\varphi_0^3 - 204\varphi_2^1 + 21\varphi_1^2)s^2 + 4(80\varphi_0^2 + 90\varphi_1^1 - 35\varphi_2^0 + 80\varphi_0^3 - 204\varphi_2^1 + 21\varphi_1^2)s + 6(10\varphi_1^1 - 35\varphi_2^0 - 4\varphi_2^1 + 21\varphi_1^2))21(s^2 + 4s + 6)$
$p = 3$
$\chi_3^1 = ((32\varphi_0^4 + 30\varphi_1^3 + 7\varphi_2^2 - 40\varphi_0^3 - 24\varphi_2^1 - 2\varphi_3^0)s^2 - 5(8\varphi_0^4 - 30\varphi_1^3 - 7\varphi_2^2 + 8\varphi_0^3 + 24\varphi_2^1 + 2\varphi_3^0)s + 10(6\varphi_1^3 + 7\varphi_2^2 - 2(6\varphi_2^1 + \varphi_3^0)))7(s^2 + 5s + 10)$ $\chi_2^2 = ((-800\varphi_0^4 - 624\varphi_1^3 + 56\varphi_3^1 + 1000\varphi_0^3 + 348\varphi_2^1 - 90\varphi_3^0 - 7\varphi_1^2)s^2 + 5(200\varphi_0^4 - 624\varphi_1^3 + 56\varphi_3^1 + 200\varphi_0^3 + 348\varphi_2^1 - 90\varphi_3^0 - 7\varphi_1^2)s - 10(24\varphi_1^3 - 56\varphi_3^1 - 48\varphi_2^1 + 90\varphi_3^0 + 7\varphi_1^2))56(s^2 + 5s + 10)$ $\chi_1^3 = ((2656\varphi_0^4 + 2000\varphi_1^3 + 84\varphi_4^0 - 3320\varphi_0^3 - 1012\varphi_2^1 + 30\varphi_3^0 - 147\varphi_1^2)s^2 - 5(664\varphi_0^4 - 2000\varphi_1^3 - 84\varphi_4^0 + 664\varphi_0^3 + 1012\varphi_2^1 - 30\varphi_3^0 + 147\varphi_1^2)s + 10(8\varphi_1^3 + 84\varphi_4^0 - 16\varphi_2^1 + 30\varphi_3^0 - 147\varphi_1^2))84(s^2 + 5s + 10)$
$p = 4$
$\chi_4^1 = ((-48\varphi_0^4 - 30\varphi_1^3 - 3\varphi_2^2 + 40\varphi_0^5 + 36\varphi_1^4 + 8\varphi_2^3)s^2 - 6(8\varphi_0^4 + 30\varphi_1^3 + 3\varphi_2^2 + 8\varphi_0^5 - 36\varphi_1^4 - 8\varphi_2^3)s - 15(14\varphi_1^3 + 3\varphi_2^2 - 8(\varphi_1^4 + \varphi_2^3)))8(s^2 + 6s + 15)$ $\chi_3^2 = ((720\varphi_0^4 + 282\varphi_1^3 - 55\varphi_2^2 - 8\varphi_3^1 - 600\varphi_0^5 - 444\varphi_1^4 + 36\varphi_3^2)s^2 + 6(120\varphi_0^4 + 282\varphi_1^3 - 55\varphi_2^2 - 8\varphi_3^1 + 120\varphi_0^5 - 444\varphi_1^4 + 36\varphi_3^2)s + 15(42\varphi_1^3 - 55\varphi_2^2 - 8\varphi_3^1 - 24\varphi_1^4 + 36\varphi_3^2))36(s^2 + 6s + 15)$ $\chi_2^3 = ((-4176\varphi_0^4 - 1434\varphi_1^3 + 55\varphi_2^2 - 154\varphi_3^1 - 9\varphi_4^0 + 3480\varphi_0^5 + 2460\varphi_1^4 + 90\varphi_4^1)s^2 - 6(696\varphi_0^4 + 1434\varphi_1^3 - 55\varphi_2^2 + 154\varphi_3^1 + 9\varphi_4^0 + 696\varphi_0^5 - 2460\varphi_1^4 - 90\varphi_4^1)s - 15(42\varphi_1^3 - 55\varphi_2^2 + 154\varphi_3^1 + 9\varphi_4^0 - 24\varphi_1^4 - 90\varphi_4^1))90(s^2 + 6s + 15)$ $\chi_1^4 = ((47376\varphi_0^4 + 15834\varphi_1^3 - 55\varphi_2^2 + 154\varphi_3^1 - 891\varphi_4^0 - 39480\varphi_0^5 - 27660\varphi_1^4 + 495\varphi_5^0)s^2 + 6(7896\varphi_0^4 + 15834\varphi_1^3 - 55\varphi_2^2 + 154\varphi_3^1 - 891\varphi_4^0 + 7896\varphi_0^5 - 27660\varphi_1^4 + 495\varphi_5^0)s + 15(42\varphi_1^3 - 55\varphi_2^2 + 154\varphi_3^1 - 891\varphi_4^0 - 24\varphi_1^4 + 495\varphi_5^0))495(s^2 + 6s + 15)$

Next, we consider transient hyperbolic problems of the form

$$L(u) = u_t + \nabla \cdot \mathbf{F}(u) = f(x, y, t), \quad (x, y) \in \Omega \ = \ (0, 1)^2, \ t > 0,$$

subject to initial condition $u_0(x, y)$ and inflow boundary conditions.

The semi-discrete DG formulation consists of determining $U \in \mathcal{U}_p$ such that

$$\iint_{\Delta}(U_t V - \mathbf{F}(U)\cdot\nabla V)dxdy + \int_{\Gamma^-} \boldsymbol{n}\cdot\mathbf{F}(\hat{U})Vds + \int_{\Gamma^+} \boldsymbol{n}\cdot\mathbf{F}(U)Vds = \iint_{\Delta} fVdxdy, \quad \forall\ V \in \mathcal{U}_p, \tag{2.10a}$$

where $\hat{U}$ is as defined in (2.6b). We compute an error estimate by solving the linearized problem

$$\iint_{\Delta}[h'(U), g'(U)]^T \cdot \nabla(U+E)\chi^k_{p+1-k}dxdy = \iint_{\Delta}(f - U_t)\chi^k_{p+1-k}dxdy, \quad k = 1, 2, \ldots, p. \tag{2.10b}$$

In order for the DG error to have the same structure for all times, we approximate the initial conditions u_0 by $U_0 \in \mathcal{U}_p$ computed from the stationary problem

$$\int_{\Gamma^-} \boldsymbol{n}\cdot\mathbf{F}(\hat{U_0})Vds + \int_{\Gamma^+} \boldsymbol{n}\cdot\mathbf{F}(U_0)Vds - \iint_{\Delta}\mathbf{F}(U_0)\cdot\nabla V dxdy = \iint_{\Delta}\nabla\cdot\mathbf{F}(u_0)Vdxdy, \quad \forall\ V \in \mathcal{U}_p, \tag{2.10c}$$

with $\hat{U_0}$ being u_0 on the boundary edges and $U_0 + E_0$ on the interior edges. We estimate the error E_0 at $t = 0$ on Δ by solving the linearized problem

$$\iint_{\Delta}[h'(U_0), g'(U_0)]^T \cdot \nabla(U_0+E_0)\chi^i_{p+1-i}dxdy = \iint_{\Delta}\nabla\cdot\mathbf{F}(u_0)\chi^i_{p+1-i}dxdy, \quad i = 1, 2, \ldots, p. \tag{2.10d}$$

Basic calculus shows that, if $h = \mathrm{diam}(\Delta)$, the Jacobian of the affine transformation from Δ to Δ_0 can be written as

$$\mathbf{J} = \begin{bmatrix}\xi_x & \eta_x \\ \xi_y & \eta_y\end{bmatrix} = \frac{1}{h}\mathbf{J}_0,$$

where $\mathbf{J}_0$ is a 2×2 matrix independent of h.

Applying Taylor's theorem we expand $\mathbf{J}_0\boldsymbol{a}$ as

$$\check{\boldsymbol{a}}(\xi,\eta,h) = \boldsymbol{a}_0 + \sum_{k=1}^{\infty} h^k \boldsymbol{a}_k(\xi,\eta),$$

where $\boldsymbol{a}_k \in [\mathcal{P}_k]^2$, and

$$\boldsymbol{a}_0 = [\alpha,\beta]^T = \begin{cases}\mathbf{J}_0\tilde{\boldsymbol{a}}(1/2,1/2), & \text{if } \Delta \text{ is of type I,}\\ \mathbf{J}_0\tilde{\boldsymbol{a}}(0,0), & \text{if } \Delta \text{ is of type II or III}\end{cases}.$$

The sign of $[\alpha,\beta]^T\cdot\mathbf{n}$ is used to determine inflow and outflow edges.

An accepted efficiency measure of a posteriori error estimates is the effectivity index. In this paper we use the effectivity indices in the $\mathcal{L}^2$ norm defined as

$$\theta \;=\; \frac{||E||_{\mathcal{L}^2(\Omega)}}{||e||_{\mathcal{L}^2(\Omega)}}. \tag{2.11}$$

Ideally, the effectivity indices should approach unity under mesh refinement. We note that for transient problems the effectivity index is denoted by $\theta(t)$.

3. Adaptive Mesh Refinement. In this section we test our error estimation procedures presented in the previous section on adaptively refined meshes. We implement several h-refinement strategies and adaptive algorithms to compute DG solutions and error estimates on successively refined meshes.

Again, we recall that our modified DG method solves steady hyperbolic problems by first creating a mesh and arranging its elements into a list $M = \{\Delta_1, \Delta_2, \ldots, \Delta_j, \ \ldots\}$ such that

- Rule 1: All inflow elements, whose inflow edges are on the domain inflow $\partial\Omega^-$, are put first in the list M.
- Rule 2: The inflow edges of an element Δ_j in M are either on the domain inflow boundary $\partial\Omega^-$ or outflow edges of an element $\Delta_i, i < j$.

The modified DG method starts by computing the solution on the first element Δ_1 and proceeds downwind by computing the solution on elements in Δ_2, $\Delta_3, \ldots$ until the last element in M. Next we discuss several adaptive refinement algorithms that subdivide element having large "errors."

Algorithm 1 solves hyperbolic problems on a succession of locally refined meshes obtained by subdividing elements with errors larger than a specified threshold *delta*.

• **Algorithm 1** consists of the following steps

```
(i)  Set delta and Maxiter and k=0
(ii) construct an initial mesh Omega_0, order its elements
     in a list M of elements satisfying Rules 1 and 2
while k < Maxiter
    a- Solve the DG problem on Omega_k as described above.
    b- Compute errors ||E||_{Delta} for each element Delta in
       Omega_k
    c- For all elements Delta in M
          if ||E||_{Delta} < delta
              accept the DG solution on Delta
          else
             reject the DG solution on Delta
```

```
            subdivide Delta into 4 congruent triangles
          endif
    d- Complete triangulation by eliminating hanging nodes
       to create new mesh Omega_k+1 and order its elements
       in a list M satisfying Rules 1 and 2.
     e- k <-- k+1
endwhile
```

• **Algorithm 2a** solves the whole problem on an initial mesh and then goes back and solves the problem on each element and applies a local refinement algorithm to obtain a more accurate DG solution. It consists of the following steps:

```
(i)   Set omega and Create an initial mesh Omega
(ii)  Solve discrete DG problem on Omega as described above
(iii) Compute errors ||E||_{Delta} for each element Delta
      and compute error Emax = max_{Delta} ||E||_{Delta}
(iv)  For all elements Delta in M satisfying Rules 1 and 2
         if ||E||_{Delta} < omega*Emax
            Accept the DG solution on Delta
         else
            Reject the DG solution
            Subdivide Delta into 4 congruent triangles
            Complete triangulation to eliminate hanging nodes
            Update the list M satisfying Rules 1 and 2
         end
```

• **Algorithm 2b** follows the same steps as Algorithm 2a except for the refinement selection strategy. An element is selected for refinement if the local error exceeds a fraction of the average error Eavg following the steps:

```
(i)   Set omega and create a mesh Omega with N elements
(ii)  Solve discrete DG problem on Omega described above
(iii) Compute errors ||E||_{Delta} for each element Delta
      Compute average error Eavg=sum_{Delta}||E||_{Delta}/N
(iv)  For all elements in M satisfying Rules 1 and 2
        if ||E||_{Delta} < omega*Eavg
           Accept the DG solution on Delta
        else
           Reject the DG solution
           Subdivide Delta into 4 congruent triangles
           Complete triangulation to eliminate hanging nodes
           Update the list M satisfying Rules 1 and 2
        end
```

• **Algorithm 2c** is similar to Algorithm 2a. However, an element is selected for refinement if its residual exceeds a fraction of the maximum residual and follows the steps:

```
(i)   Set omega and Create an initial mesh Omega
(ii)  Solve discrete DG problem on Omega described above
(iii) Compute residual ||r||_{Delta}=||L(U)-f||_{Delta}
      on each element Delta and compute maximum residual
      rmax= max_{Delta} ||r||_{Delta}
(iv)  For all elements in M satisfying Rules 1 and 2
         if ||r||_{Delta} < omega*rmax
            Accept the DG solution on Delta
         else
            Reject the DG solution
            Subdivide Delta into 4 congruent triangles
            Complete triangulation to eliminate hanging nodes
            Update the list M satisfying Rules 1 and 2
         end
```

• **Algorithm 2d** follows the same steps as Algorithm 2a with one exception: an element is selected for refinement if the element residual exceeds a fraction of the average residual and follows the steps:

```
(i)   Set omega and Create a mesh Omega with N elements
(ii)  Solve discrete DG problem on Omega described above
(iii) Compute residual ||r||_{Delta}=||L(U)-f||_{Delta}
       on each element Delta and compute average residual
       ravg= sum_{Delta} ||r||_{Delta}/N
(iv)  For all elements in M satisfying Rules 1 and 2
          if ||r||_{Delta} < omega*ravg
            Accept the DG solution on Delta
          else
            Reject the DG solution
            Subdivide Delta into 4 congruent triangles
            Complete triangulation to eliminate hanging nodes
            Update the list M satisfying Rules 1 and 2
          end
```

• **Algorithm 3** prevents the errors on elements near the discontinuity from polluting elements having smooth solutions. Each element having a large error is immediately refined after computing its DG solution, while the refinement in Algorithm 1 is performed after computing the solution on all elements. Thus, this algorithm will reduce the errors as they appear and is expected to reduce the pollution errors observed with Algorithm 1. This algorithm consists of the following steps:

```
(i)   Set delta
(ii)  Construct an initial mesh
(iii) Order elements in a list M satisfying Rules 1 and 2
```

```
(iv)  For all elements Delta in M
          a- Compute  DG solution on Delta
          b- Compute error ||E||_{Delta}

          c- If ||E||_{Delta} < delta
                Accept the DG solution on Delta
             else
                Reject DG solution on Delta
                Subdivide Delta into 4 congruent triangles
                Eliminate hanging nodes
                Update the list M satisfying Rules 1 and 2
             endif
```

• **Algorithm 4** is similar to Algorithm 3 except for the refinement selection criteria. Here an element Delta is selected for refinement if the maximum residual over all elements before Delta in the list M exceeds a user specified tolerance delta and follows the following steps:

```
(i)   Set delta
(ii)  Construct an initial mesh
(iii) Order elements in a list M satisfying Rules 1 and 2
(iv)  For all elements Delta in M
         a- Compute  DG solution on Delta
         b- Compute maximum residual rmax over all elements
            in M before element Delta
         c- If rmax < delta
                Accept the DG solution on Delta
            else
                Reject DG solution on Delta
                Subdivide Delta into 4 congruent triangles
                Eliminate hanging nodes
                Update the list M satisfying Rules 1 and 2
            endif
```

• **Algorithm 5** is similar to Algorithm 4 except for the refinement selection criteria. Here an element Delta is selected for refinement if the average residual over all elements before Delta exceeds a user specified tolerance delta and follows the following steps:

```
(i)   Set delta
(ii)  Construct an initial mesh
(iii) Order elements in a list M satisfying Rules 1 and 2
(iv)  For all elements Delta in M
          a- Compute  DG solution on Delta
          b- Compute average residual ravg over all elements
             in M and before element Delta
```

```
          c- If ravg < delta
                Accept the DG solution on Delta
             else
                Reject DG solution on Delta
                Subdivide Delta into 4 congruent triangles
                Eliminate hanging nodes
                Update the list M satisfying Rules 1 and 2
             endif
```

• **Algorithm 6** is used with a time-marching scheme that converges to a steady state solution and consists of the following steps:

```
(i) Set delta, NStep, k=0 and construct a mesh Omega_k
    Set dt and t_k = k*dt
while k < Nstep
   (ii)  Integrate from t_k to t_{k+1} on Omega_k
   (iii) Compute errors ||E||_{Delta} on each element Delta
   (iv)  For all elements Delta
            If ||E||_{Delta} < delta
               Accept the DG solution on Delta
            Else
               Subdivide Delta into 4 congruent triangles
            endif

    (v)   Complete triangulation to eliminate hanging nodes
           Create a new mesh Omega_{k+1}

    (vi)  Increment k <--k+1 and return to step (ii)
endwhile
```

4. Computational Examples. In this section, we present numerical results for several hyperbolic problems showing the convergence properties of DG solutions and a posteriori error estimates in the presence of discontinuities on unstructured meshes. The error estimates are tested on linear and nonlinear problems with discontinuous solutions to show their efficiency and accuracy under adaptive mesh refinement. For all examples we use exact boundary conditions at the *inflow* boundary. In all our examples we use the space $\mathcal{U}_p$ on unstructured triangular meshes. Furthermore, for the transient problem we apply the MATLAB ode45 to perform the temporal integration and assume the temporal discretization errors to be negligible.

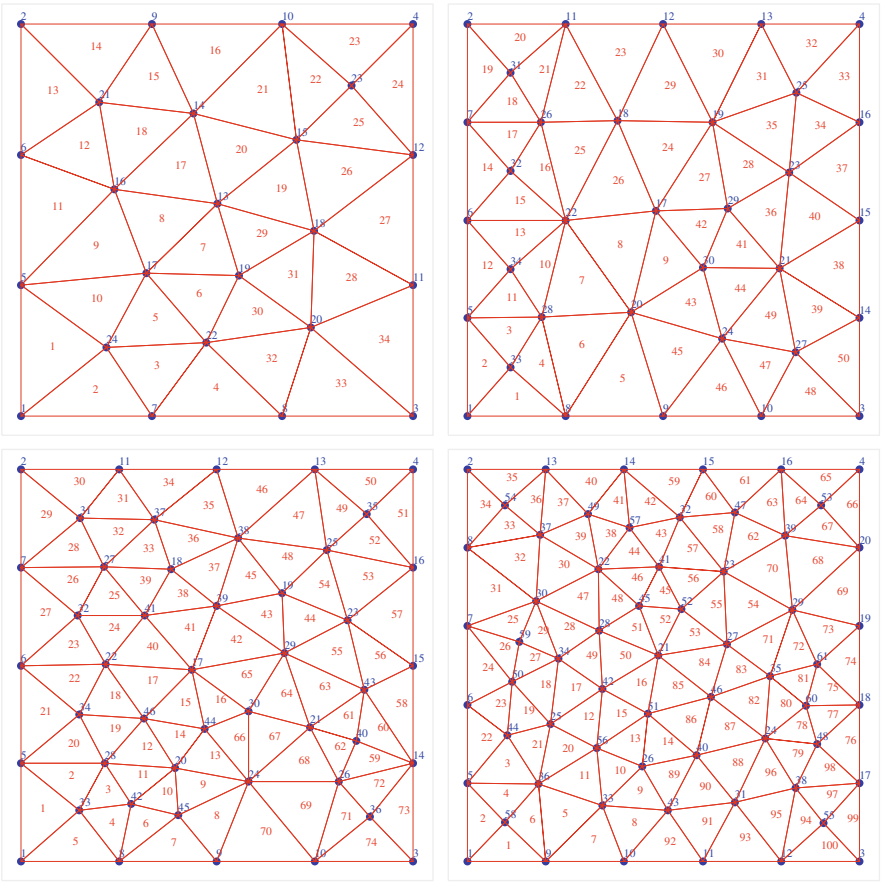


FIG. 1. Unstructured meshes having $N = 34, 50, 74, 100$ triangles

EXAMPLE 1. Let us consider the initial-boundary value problem for the inviscid Burgers' equation

$$u_y + \left(\frac{u^2}{2}\right)_x = u_y + uu_x = 0, \ (x, y) \in [0, 1] \times [0, 0.999], \tag{4.1a}$$

subject to the initial conditions

$$u(x, 0) = g_0(x) = 1 + \frac{1}{2}\sin(2\pi x). \tag{4.1b}$$

and select $u(0, y) = g_1(y)$ such that the true solution is periodic and forms a shock discontinuity at the point $(\frac{1}{\pi}, \frac{1}{\pi})$ which propagates along $y = x$. We perform several tests on this example to study both the accuracy and efficiency of our a posteriori error estimates for the modified DG method on adaptively refined unstructured meshes.

First, we apply the modified DG method (2.10) and (2.6b) to solve problem (4.1) on the unstructured meshes shown in Fig. 1 having $N = 34, 50, 74, 100$ triangular elements with $\mathcal{U}_p$, $p = 1, 2, 3, 4$. We observe that the proposed error estimates do not converge to the true error under regular mesh refinement. The global error is underestimated is due to the fact that errors on elements near the discontinuity, which have an important contribution to the global error, are underestimated (Table 3).

TABLE 3
$\mathcal{L}^2$ errors and global effectivity indices on for problem (4.1) on unstructured meshes having $N = 34, 50, 74, 100$ elements using $p = 1, 2, 3, 4$

N	$p=1$		$p=2$	
	$\|\|e\|\|_{\mathcal{L}^2}$	θ	$\|\|e\|\|_{\mathcal{L}^2}$	θ
34	8.0543e−1	0.87505	1.2714e−1	0.67312
50	5.1083e−1	0.88783	4.1497e−2	0.68294
74	2.6779e−1	0.89174	2.0175e−2	0.68595
100	1.2956e−1	0.89585	1.1439e−2	0.68912
	$p=3$		$p=4$	
34	1.5127e−1	0.36461	1.1168e−1	0.25002
50	4.6004e−2	0.36993	2.2925e−2	0.25366
74	1.6925e−2	0.37156	6.9803e−3	0.25478
100	8.0424e−3	0.37327	2.8216e−3	0.25596

We apply the modified DG method (2.9) and (2.6b) to solve the inviscid Burgers' equation (4.1) on $[0, 1] \times [0, 0.999]$ on unstructured meshes having $N = 432$ elements shown in Fig. 2 with $\mathcal{U}_p$ and with no special treatment at the shock such as stabilization or limiting. Then, we apply Algorithm 1 to locally refine the mesh by performing 6 iterations to generate a sequence of refined meshes. Algorithm 1 subdivides triangles for which the local error in the $\mathcal{L}^2$ norm is larger than a prescribed tolerance δ. In Figs. 3 and 4 we show the sequence of meshes obtained by applying Algorithm 1 with $\delta = 0.001$ for $p = 1, 2$ where elements near the shock discontinuity are refined. However, we notice that a large portion of elements away from the discontinuity are refined which suggest that this algorithm is not efficient.

In Table 4 we present the number of elements in each mesh obtained at every refinement iteration of Algorithm 1, the true $\mathcal{L}^2$ errors and the effectivity indices for $p = 1, 2$ for each refinement iteration. These results suggest that the global $\mathcal{L}^2$ error estimates converge to the true error under a "crude" adaptive mesh refinement of Algorithm 1 in the presence of shock discontinuities.

We now consider the same problem (4.1) and use Algorithm 2a with the modified DG method (2.9) and (2.6b). In Figs. 5 and 6 we plot the meshes obtained from Algorithm 2a with $\omega = 0.85$, $\omega = 0.75$, $\omega = 0.5$, $\omega = 0.25$,

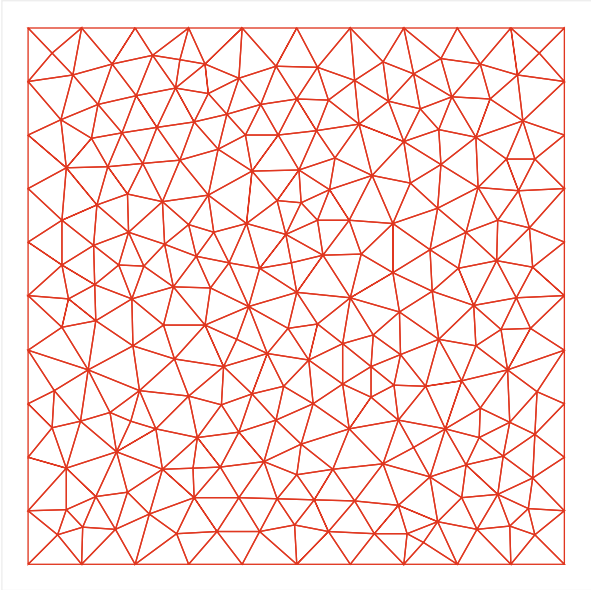


FIG. 2. An unstructured mesh having $N = 432$ triangles

TABLE 4
$\mathcal{L}^2$ errors and global effectivity indices for problem (4.1) on unstructured meshes having N elements using Algorithm 1 with 6 refinement iterations

	$p=1$			$p=2$		
Iteration	N	$\|\|e\|\|_{\mathcal{L}^2}$	θ	N	$\|\|e\|\|_{\mathcal{L}^2}$	θ
1	432	2.5104e−2	0.5462	432	3.0244e−3	0.4189
2	841	1.4230e−2	0.6253	711	1.3184e−3	0.5491
3	1,778	9.2502e−3	0.7555	1,323	6.9683e−4	0.6393
4	3,012	6.4997e−3	0.8508	2,544	4.0806e−4	0.7595
5	5,514	4.2522e−4	0.9448	3,970	2.5819e−5	0.8698
6	8,148	2.0173e−4	0.9747	5,783	1.5164e−5	0.9499

and $p = 1, 2$. In Table 5 we present the number of elements N, the true $\mathcal{L}^2$ errors, and the global $\mathcal{L}^2$ effectivity indices with $p = 1, 2$ which suggest that the proposed error estimates converge to the true error under adaptive refinement of unstructured triangular meshes.

Now, we use Algorithm 2b to solve (4.1) with the modified DG method (2.9) and (2.6b). In Fig. 7 we plot the meshes obtained by applying the adaptive algorithm with $\omega =$ 2, 1.75, 1.5, 1.25, 1, 0.75, 0.5, 0.25 and $p = 1, 2$. In Table 6 we present the number of elements N, the true $\mathcal{L}^2$ errors, and the global $\mathcal{L}^2$ effectivity indices with $p = 1, 2$ for $\omega = 2$, 1.75,

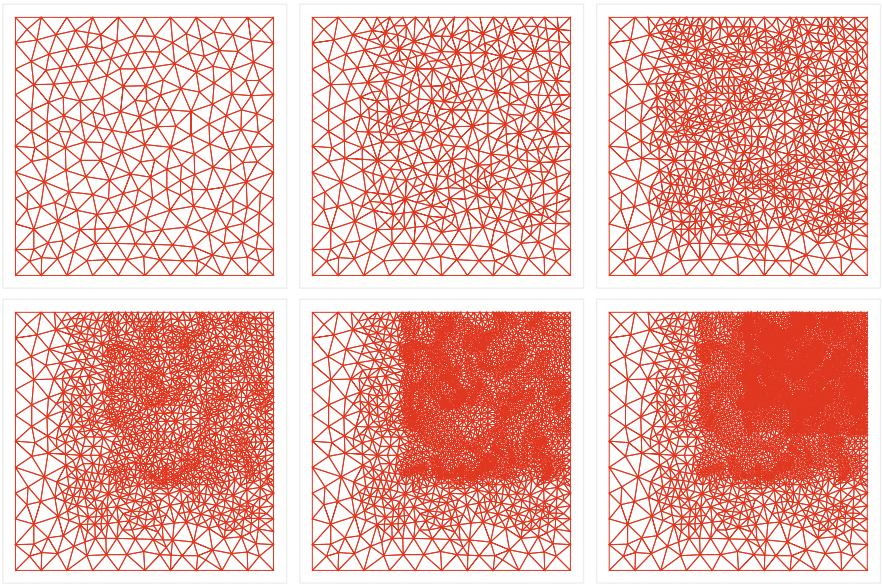


FIG. 3. Adaptive meshes obtained by Algorithm 1 for problem (4.1) and $p = 1$

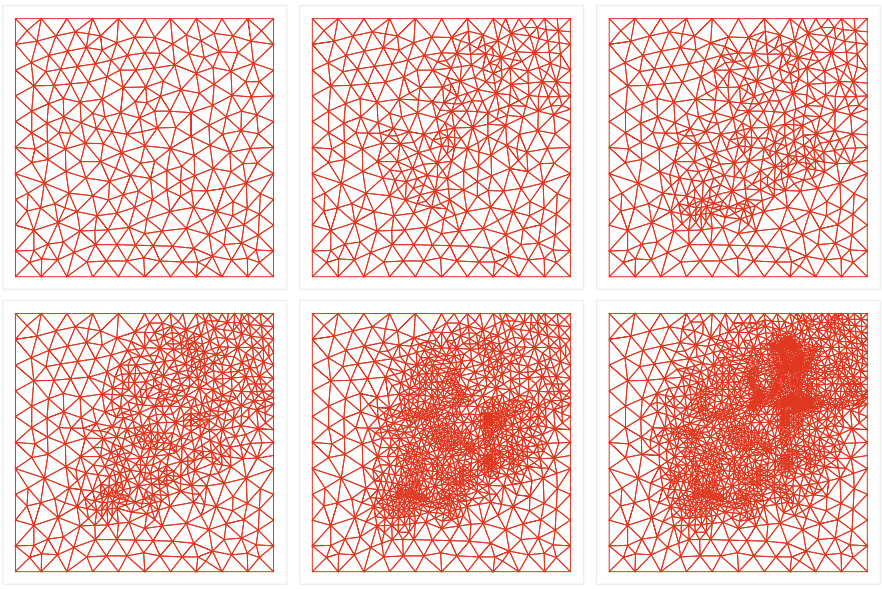


FIG. 4. Adaptive meshes obtained by Algorithm 1 for problem (4.1) and $p = 2$

TABLE 5
$\mathcal{L}^2$ errors and global effectivity indices for problem (4.1) on unstructured meshes having N elements using Algorithm 2a with $\omega = 0.85, 0.75, 0.5, 0.25, 0.15, 0.1$ and $p = 1, 2$

p	$p=1$			$p=2$		
ω	N	$\|\|e\|\|_{\mathcal{L}^2}$	θ	N	$\|\|e\|\|_{\mathcal{L}^2}$	θ
0.85	627	2.3903e−2	0.6528	842	2.9155e−3	0.5561
0.75	917	1.3549e−2	0.7744	1,262	6.7175e−4	0.6898
0.5	1,433	8.8076e−3	0.8704	2,156	3.9337e−4	0.8674
0.25	2,094	4.0488e−4	0.9761	5,011	4.5249e−5	0.9686
0.15	3,450	1.9546e−4	0.9798	7,213	2.2231e−5	0.9752
0.1	4,343	1.5432e−4	0.9812	8,982	1.1856e−5	0.9765

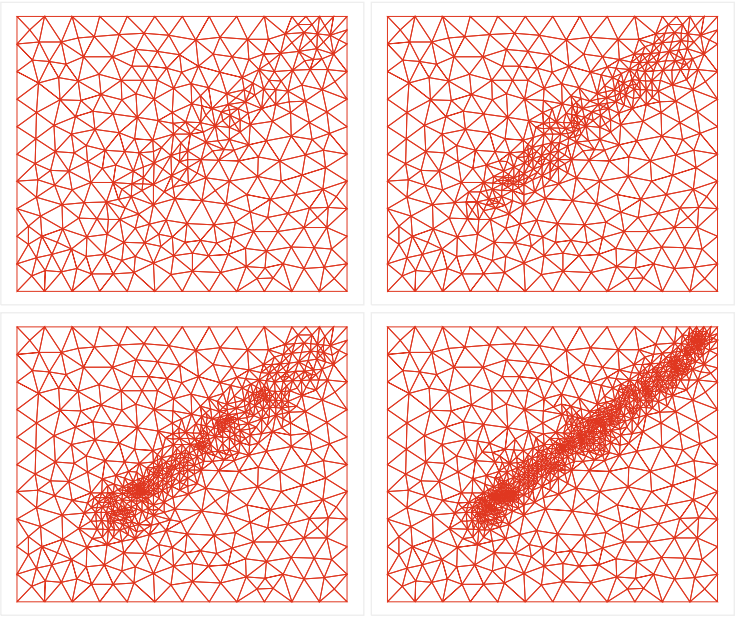


FIG. 5. Meshes generated by Algorithm 2a for the problem (4.1) with $\omega = 0.85$, $\omega = 0.75$, $\omega = 0.5$, $\omega = 0.25$ (*upper left to lower right*) and $p = 1$

1.5, 1.25, 1, 0.75, 0.5, 0.25 which show that our error estimates converge to the true error under adaptive refinement of unstructured triangular meshes.

This adaptive mesh-refinement strategy also yields an efficient adaptive algorithm.

Next, we solve problem (4.2) using Algorithm 2c with the modified DG method (2.9) and (2.6b). In Figs. 8 and 9 we plot the meshes obtained by Algorithm 2c with $\omega = 0.85$, $\omega = 0.75$, $\omega = 0.5$, $\omega = 0.25$ and $p = 1, 2$ to the problem (4.1). In Table 7 we present the number of elements N, the global $\mathcal{L}^2$ norm of the error, and the global $\mathcal{L}^2$ effectivity indices with $p = 1, 2$

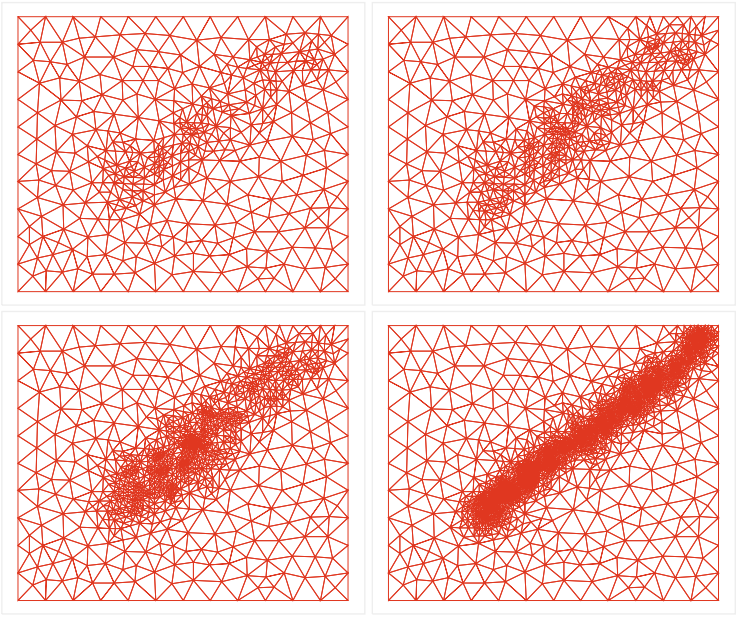


FIG. 6. Meshes generated by Algorithm 2a for problem (4.1) with $\omega = 0.85$, $\omega = 0.75$, $\omega = 0.5$, $\omega = 0.25$ (*upper left to lower right*) and $p = 2$

which show that the error estimates converge to the true error under adaptive mesh refinement. They further show that the proposed error estimates are accurate on adaptively refined unstructured triangular meshes.

TABLE 6
$\mathcal{L}^2$ errors and global effectivity indices for problem (4.1) on unstructured meshes having N elements using Algorithm 2d with $\omega = 2, 1.75, 1.5, 1.25, 1, 0.75, 0.5, 0.25$ and $p = 1, 2$

	$p=1$			$p=2$		
ω	N	$\|\|e\|\|_{\mathcal{L}^2}$	θ	N	$\|\|e\|\|_{\mathcal{L}^2}$	θ
2.00	705	1.8498e−2	0.5351	803	1.7897e−3	0.5448
1.75	865	1.0485e−2	0.5819	967	1.0145e−3	0.5938
1.50	1,426	6.8160e−3	0.6959	1,756	6.5947e−4	0.7167
1.25	1,941	5.7147e−3	0.7813	2,305	4.6337e−4	0.7978
1.00	3,140	4.7892e−3	0.8918	3,922	3.3465e−4	0.9186
0.75	5,462	3.1332e−4	0.9631	6,431	3.0315e−5	0.9435
0.50	11,142	2.0637e−4	0.9722	8,670	2.0813e−5	0.9627
0.25	19,280	1.2973e−4	0.9828	14,548	1.0637e−5	0.9749

Next, we apply Algorithm 2d with the modified DG method (2.9) and (2.6b) to solve (4.1). In Fig. 10 we plot the meshes obtained by applying Algorithm 2d with $\omega = 2, 1.75, 1.5, 1.25, 1, 0.75, 0.5, 0.25$ for $p = 1, 2$. In

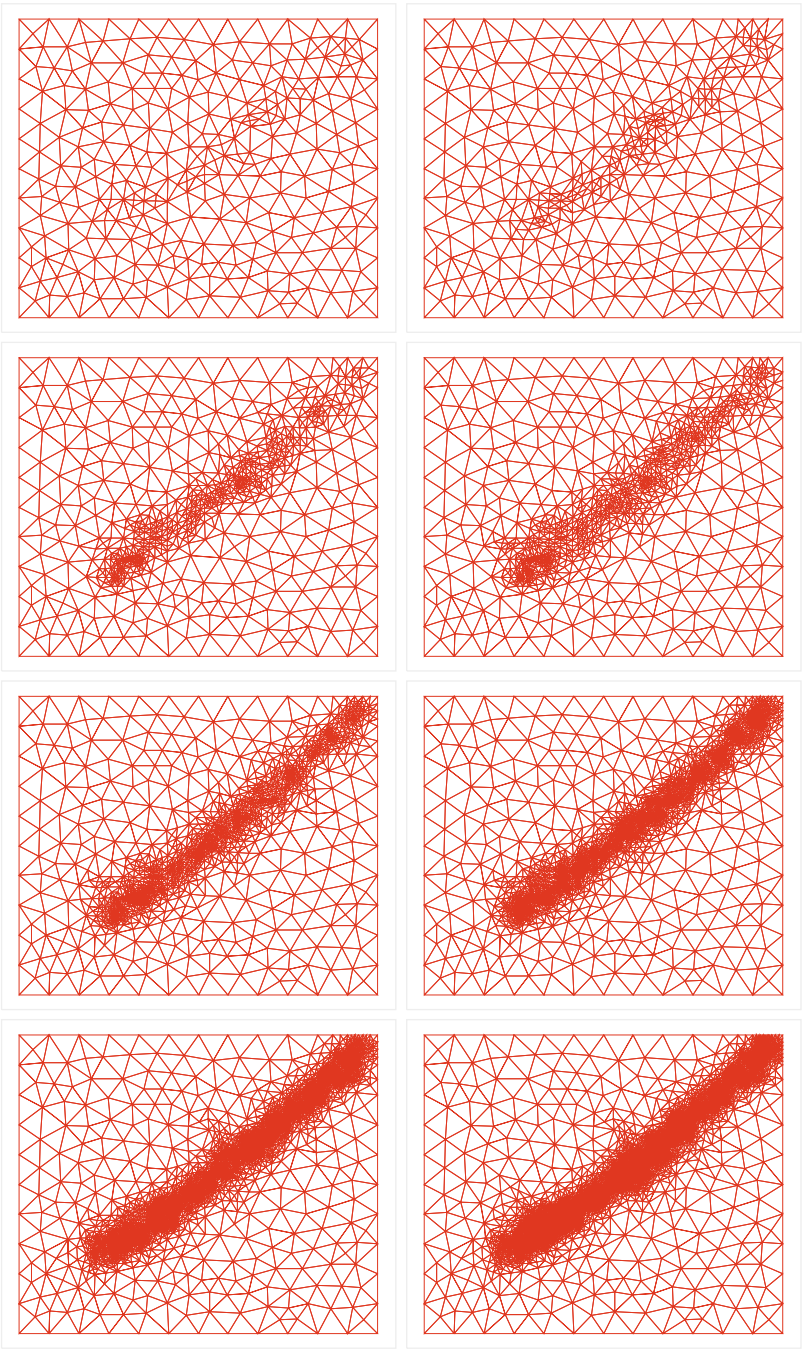


FIG. 7. Meshes obtained by Algorithm 2b for problem (4.1) with $\omega = 2,\ 1.75,\ 1.5,\ 1.25,\ 1,\ 0.75,\ 0.5,\ 0.25$ (*upper left to lower right*) and $p = 1$

TABLE 7
$\mathcal{L}^2$ errors and global effectivity indices for problem (4.1) on unstructured meshes having N elements using Algorithm 2c with $\omega = 0.85, 0.75, 0.5, 0.25, 0.15, 0.1$ and $p = 1, 2$

Degree	$p=1$			$p=2$		
ω	N	$\|\|e\|\|_{\mathcal{L}^2}$	θ	N	$\|\|e\|\|_{\mathcal{L}^2}$	θ
0.85	677	2.3062e−2	0.5942	817	3.0276e−3	0.5860
0.75	937	1.3072e−2	0.7368	1,362	1.3199e−3	0.7395
0.50	1,505	8.4978e−3	0.8477	2,130	4.0850e−4	0.8453
0.25	2,543	3.9064e−4	0.9664	3,385	3.5191e−5	0.9599
0.15	3,143	2.1354e−4	0.9713	4,328	1.6104e−5	0.9674
0.10	3,951	1.4983e−4	0.9784	5,235	1.1467e−5	0.9733

Table 8 we present the number of elements N, the $\mathcal{L}^2$ errors, and the global $\mathcal{L}^2$ effectivity indices for $p = 1, 2$ and $\omega = 2$, 1.75, 1.5, 1.25, 1, 0.75, 0.5, 0.25 which suggest that the proposed error estimates converge to the true error under adaptive mesh refinement on unstructured triangular meshes.

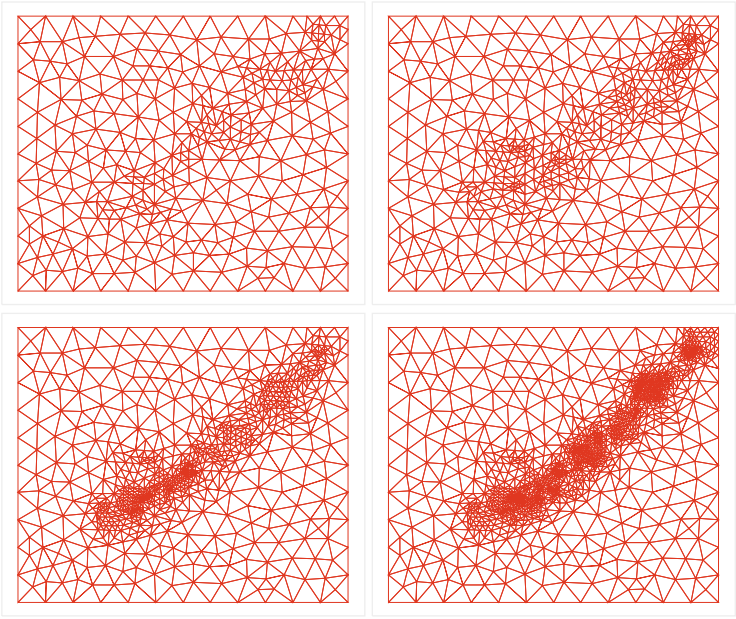


FIG. 8. Meshes generated by Algorithm 2c for problem (4.1) with $\omega = 0.85$, $\omega = 0.75$, $\omega = 0.5$, $\omega = 0.25$ (*upper left to lower right*) and $p = 1$

Now we apply the modified DG method (2.9) and (2.6b) with Algorithm 3 to solve (4.1). Again, the final mesh is constructed through a sequence of successively refined meshes by refining triangles whose local

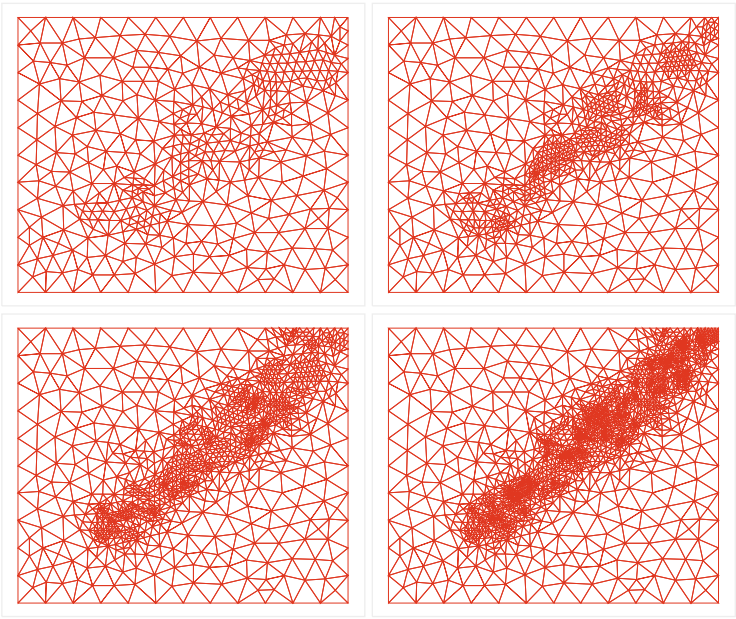


FIG. 9. Meshes generated by Algorithm 2c for problem (4.1) with $\omega = 0.85$, $\omega = 0.75$, $\omega = 0.5$, $\omega = 0.25$ (*upper left to lower right*) and $p = 2$

TABLE 8
$\mathcal{L}^2$ errors and global effectivity indices for problem (4.1) on unstructured meshes having N elements using Algorithm 2b with $\omega = 2,\ 1.75,\ 1.5,\ 1.25,\ 1,\ 0.75,\ 0.5,\ 0.25$ and $p = 1, 2$

Degree	$p = 1$			$p = 2$		
ω	N	$\|\|e\|\|_{\mathcal{L}^2}$	θ	N	$\|\|e\|\|_{\mathcal{L}^2}$	θ
2.00	580	2.0179e−2	0.5263	678	2.0660e−3	0.4536
1.75	749	2.7182e−2	0.5631	957	1.1710e−3	0.5138
1.50	1,399	1.1438e−2	0.6860	1,832	7.6126e−4	0.6313
1.25	1,836	7.4356e−3	0.7539	2,450	5.3490e−4	0.6868
1.00	3,094	5.2246e−3	0.8962	3,985	3.6675e−4	0.8171
0.75	5,430	3.4181e−4	0.9648	5,043	8.3722e−5	0.8791
0.50	10,272	2.2616e−4	0.9781	8,086	3.1995e−5	0.9409
0.25	17,841	1.4217e−4	0.9787	15,479	1.1250e−5	0.9629

error in the $\mathcal{L}^2$ norm is larger than some prescribed δ. For instance, in Figs. 11 and 12 we show the final meshes obtained by applying Algorithm 3 with $\delta = 0.01,\ 0.001$, for $p = 1, 2$. The $\mathcal{L}^2$ errors and effectivity indices for $\delta = 0.05,\ 0.01,\ 0.005,\ 0.001,\ 0.0005$ and $p = 1, 2$ shown in Table 9 suggest that the proposed error estimates converge to the true error under adaptive mesh refinement. Furthermore, we observe that the adaptive

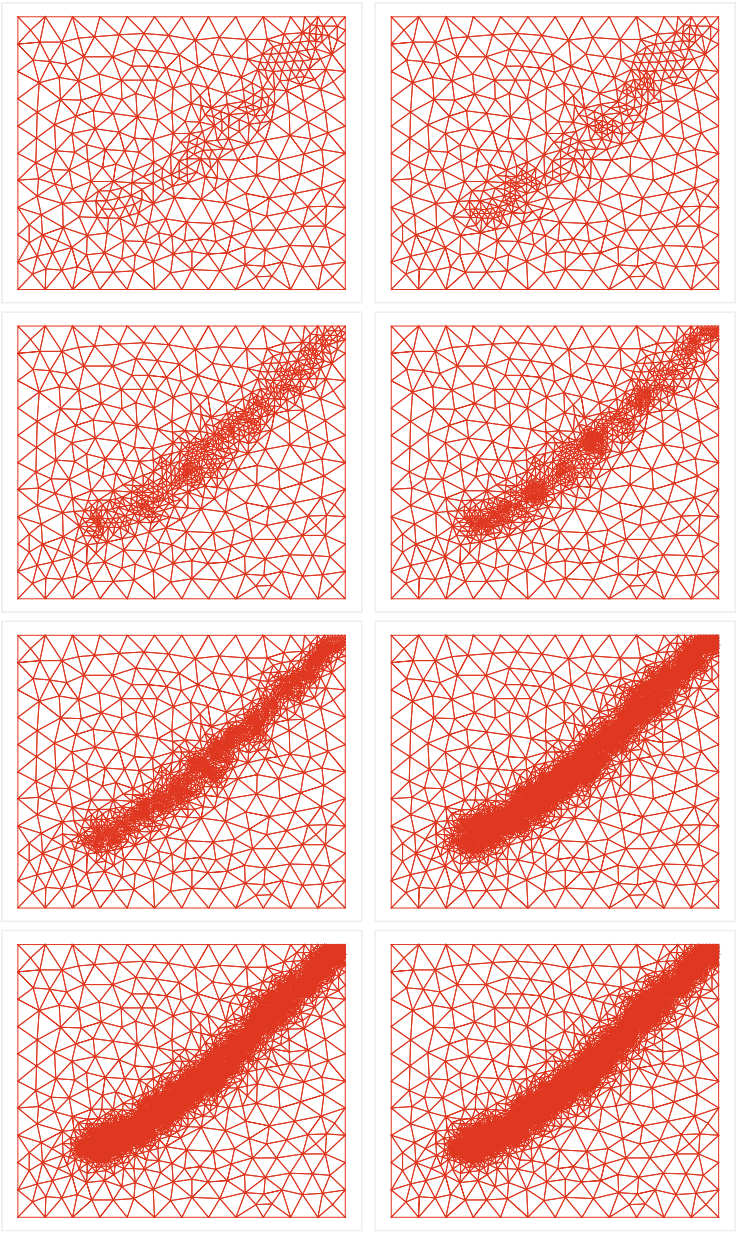


FIG. 10. Meshes obtained by Algorithm 2d for problem (4.1) with $\omega = 2,\ 1.75,\ 1.5,\ 1.25,\ 1,\ 0.75,\ 0.5,\ 0.25$ (*upper left to lower right*) and $p = 1$

TABLE 9
$\mathcal{L}^2$ errors and global effectivity indices for problem (4.1) on unstructured meshes having N elements using Algorithm 3 with $p = 1, 2$

p	$p = 1$			$p = 2$		
δ	N	$\|\|e\|\|_{\mathcal{L}^2}$	θ	N	$\|\|e\|\|_{\mathcal{L}^2}$	θ
0.05	759	9.8076e−3	0.9541	1,167	9.7476e−4	0.9569
0.01	904	8.3650e−3	0.9628	2,092	6.2409e−4	0.9644
0.005	1,532	7.2128e−4	0.9718	3,945	7.0488e−5	0.9718
0.001	2,267	3.8453e−4	0.9737	5,444	2.4167e−5	0.9754
0.0005	5,890	1.4675e−4	0.9813	8,412	1.1247e−5	0.9841

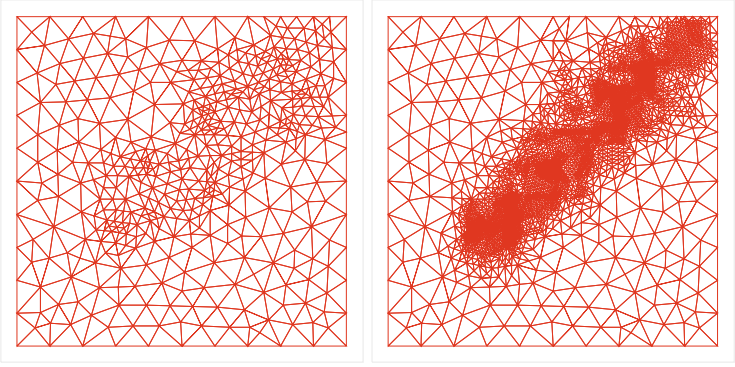


FIG. 11. Meshes obtained by Algorithm 3 for problem (4.1) with tolerance $\delta = 0.01$ and $p = 1$ (*left*), $p = 2$ (*right*)

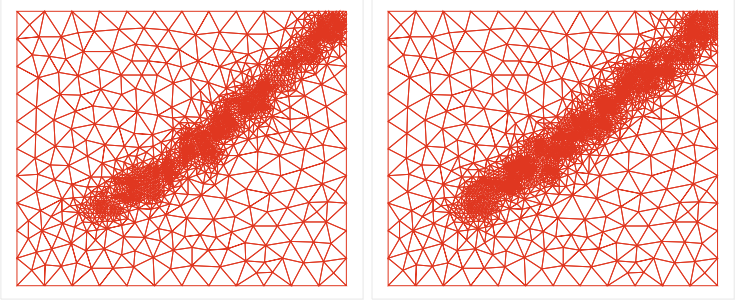


FIG. 12. Meshes obtained by Algorithm 3 for problem (4.1) with tolerance $\delta = 0.001$ and $p = 1$ (*left*), $p = 2$ (*right*)

TABLE 10
$\mathcal{L}^2$ errors and global effectivity indices for problem (4.1) on unstructured meshes having N elements using Algorithm 4 with $p = 1, 2$

	$p = 1$			$p = 2$		
δ	N	$\|\|e\|\|_{\mathcal{L}^2}$	θ	N	$\|\|e\|\|_{\mathcal{L}^2}$	θ
0.0500	732	9.3422e−3	0.6447	1,087	8.8284e−4	0.5558
0.0100	1,023	6.3276e−3	0.7864	1,983	5.2330e−4	0.6574
0.0050	1,465	6.6306e−4	0.8771	3,829	5.5168e−5	0.7681
0.0010	2,105	3.0643e−4	0.9372	5,356	2.1162e−5	0.8882
0.0005	5,459	1.2974e−4	0.9800	8,221	1.0933e−5	0.9588

mesh-refinement strategy of Algorithm 3 yields a more efficient adaptive algorithm for the modified DG method applied to hyperbolic problems on general unstructured triangular meshes.

Now, we use Algorithm 4 and the modified DG method (2.9) and (2.6b) to solve (4.1). The final mesh is constructed through a sequence of successively refined meshes where triangles for which the $\mathcal{L}^2$ norm of the residual on each element $||r_i||$ is larger than $\delta = 0.001$ are refined. In Fig. 13 we plot the final meshes obtained by applying Algorithm 4 for $p = 1, 2$. In Table 10 we present the number of elements, $\mathcal{L}^2$ errors, and the global $\mathcal{L}^2$ effectivity indices for the tolerances $\delta = 0.05, 0.01, 0.005, 0.001, 0.0005$, $p = 1, 2$ which suggest that the error estimates converge to the true error during under adaptive mesh refinement.

As a final test, we solve (4.1) using Algorithm 5 with modified DG method (2.9) and (2.6b). The final mesh is constructed through a sequence of successively refined meshes where triangles for which the average residual exceeds the specified threshold δ are refined. In Fig. 14 we plot the final meshes from Algorithm 5 with $\delta = 0.001$ and $p = 1, 2$. In Table 11 we present the number of elements, $\mathcal{L}^2$ errors, and the global $\mathcal{L}^2$ effectivity indices for tolerances $\delta = 0.05, 0.01, 0.005, 0.001, 0.0005$, and degrees $p = 1, 2$. These results suggest that our a posteriori error estimates converge to the true error under adaptive mesh refinement.

TABLE 11
$\mathcal{L}^2$ errors and global effectivity indices for problem (4.1) on unstructured meshes having N elements using Algorithm 5 with $p = 1, 2$

	$p = 1$			$p = 2$		
δ	N	$\|\|e\|\|_{\mathcal{L}^2}$	θ	N	$\|\|e\|\|_{\mathcal{L}^2}$	θ
0.0500	712	9.9638e−3	0.6257	1,054	8.9863e−4	0.5578
0.0100	914	3.2468e−3	0.7474	1,623	4.8685e−4	0.7594
0.0050	1,296	5.4606e−4	0.8881	3,298	5.2624e−5	0.8601
0.0010	1,949	2.9263e−4	0.9482	4,846	2.0400e−5	0.9202
0.0005	4,787	1.3041e−4	0.9788	6,759	1.0845e−5	0.9608

We observe that Algorithms 2–5 generate adaptive meshes with fine elements only in the vicinity of the shock. The results from Algorithms 2a–d of Tables 12 and 13 suggest that Algorithms 2a–d have comparable efficiency. In Fig. 15 we plot the true $\mathcal{L}^2$ errors versus the number of elements needed to satisfy the specified tolerance for all Algorithms 1–5 applied to the Burger's equation with a shock (4.1). Our computations suggest that Algorithm 1 is the least efficient adaptive method, while Algorithms 3, 4, and 5 are the most efficient adaptive procedures. We further note that using the local residuals to refine elements in Algorithm 5 yields a slightly more efficient algorithm. Thus, we recommend the space-time DG method that uses local residuals to select elements for refinement and the a posteriori error estimates to assess the solution accuracy and terminate the adaptive process.

TABLE 12
$\mathcal{L}^2$ errors and number of elements for Algorithms 2a and 2c applied to problem (4.1)

$p=1$					$p=2$			
Algorithm 2a			Algorithm 2c		Algorithm 2a		Algorithm 2c	
ω	N	$\|\|e\|\|_{\mathcal{L}^2}$	N	$\|\|e\|\|_{\mathcal{L}^2}$	N	$\|\|e\|\|_{\mathcal{L}^2}$	N	$\|\|e\|\|_{\mathcal{L}^2}$
0.85	627	2.3903e−2	677	2.3062e−2	842	2.9155e−3	817	3.0276e−3
0.75	917	1.3549e−2	937	1.3072e−2	1,262	6.7175e−4	1,362	1.3199e−3
0.50	1,433	8.8076e−3	1,505	8.4978e−3	2,156	3.9337e−4	2,130	4.0850e−4
0.25	2,094	4.0488e−4	2,543	3.9064e−4	5,011	4.5249e−5	3,385	3.5191e−5
0.15	3,450	1.9546e−4	3,143	2.1354e−4	7,213	2.2231e−5	4,328	1.6104e−5
0.10	4,343	1.5432e−4	3,951	1.4983e−4	8,982	1.1856e−5	5,235	1.1467e−5

EXAMPLE 2 (Transient Burger's equation). Let us consider the initial-boundary value problem for the inviscid Burgers' equation

$$\epsilon u_t + u_y + \left(\frac{u^2}{2}\right)_x = 0, \ (x,y,t) \in [0,1] \times [0,0.999] \times [0,T], \tag{4.2a}$$

subject to the boundary conditions

$$u(x,0,t) = u_2(x) = 1 + \frac{1}{2}\sin(2\pi x), \qquad u(0,y,t) = u(1,y,t) = u_1(y). \tag{4.2b}$$

The initial conditions $u(x,y,0) = u_0(x,y)$ are selected such that $u_0(x,0) = u_2(x)$ and $u_0(0,y) = u_1(y)$ as follows

$$\begin{aligned} u_0(x,y) &= \overline{N}_1(x)u_1(y) + \overline{N}_2(x)\tilde{u}_1(y) + \overline{N}_1(y)u_2(x) + \overline{N}_2(y)\tilde{u}_2(x) \\ &\quad - \overline{N}_1(x)\overline{N}_1(y)u_1(0) - \overline{N}_2(x)\overline{N}_2(y)\tilde{u}_1(1) - \overline{N}_1(x)\overline{N}_2(y)u_1(1) \\ &\quad - \overline{N}_2(x)\overline{N}_1(y)u_2(1). \end{aligned} \tag{4.2c}$$

where $\overline{N}_1(x) = 1 - x$, $\overline{N}_2(x) = x$, $\tilde{u}_1(y) = (1-y)u_2(1)$ and $\tilde{u}_2(x) = (1-x)u_1(1)$.

We apply the modified DG method (2.10) to solve this problem on $[0,1] \times [0,0.999] \times [0,T]$ with a smooth solution on initial unstructured

TABLE 13
$\mathcal{L}^2$ errors and number of elements for Algorithms 2b and 2d applied to problem (4.1)

Algorithm 2b			Algorithm 2d	
ω	N	$\|\|e\|\|_{\mathcal{L}^2}$	N	$\|\|e\|\|_{\mathcal{L}^2}$
$p=1$				
2.00	580	2.0179e−2	705	1.8498e−2
1.75	749	2.7182e−2	865	1.0485e−2
1.5	1,399	1.1438e−2	1,426	6.8160e−3
1.25	1,836	7.4356e−3	1,941	5.7147e−3
1.00	3,094	5.2246e−3	3,140	4.7892e−3
0.75	5,430	3.4181e−4	5,462	3.1332e−4
0.50	10,272	2.2616e−4	11,142	2.0637e−4
0.25	17,841	1.4217e−4	19,280	1.2973e−4
$p=2$				
2.00	678	2.0660e−3	803	1.7897e−3
1.75	957	1.1710e−3	967	1.0145e−3
1.5	1,832	7.6126e−4	1,756	6.5947e−4
1.25	2,450	5.3490e−4	2,305	4.6337e−4
1.00	3,985	3.6675e−4	3,922	3.3465e−4
0.75	5,043	8.3722e−5	6,431	3.0315e−5
0.50	8,086	3.1995e−5	8,670	2.0813e−5
0.25	15,479	1.1250e−5	14,548	1.0637e−5

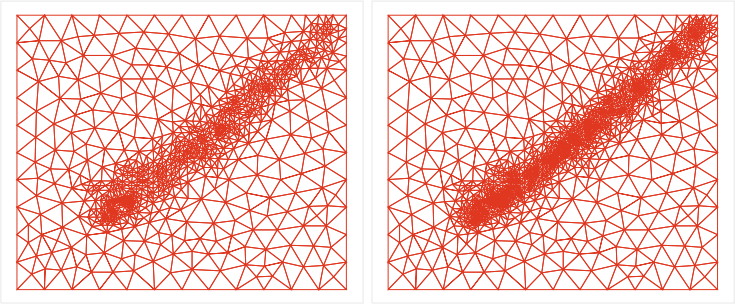


FIG. 13. Meshes generated by Algorithm 4 for problem (4.1) with tolerance $\delta = 0.001$ and $p = 1$ (*left*), $p = 2$ (*right*)

meshes having $N = 500$ triangular elements of type I, II, and III with $\mathcal{U}_p$, $p = 1, 2$, and $\epsilon = 10^{-2}$.

We use the adaptive mesh-refinement procedure given in Algorithm 6. The final mesh at $t = T = 1$ is constructed through a sequence of successively refined meshes where triangles for which the $\mathcal{L}^2$ error exceeds δ are refined. In Figs. 16 and 17 we plot the sequence of meshes obtained by applying the adaptive method with $\delta = 0.001$ and $t = 0, 0.25, 0.5, 0.75, 1$ to the problem (4.2).

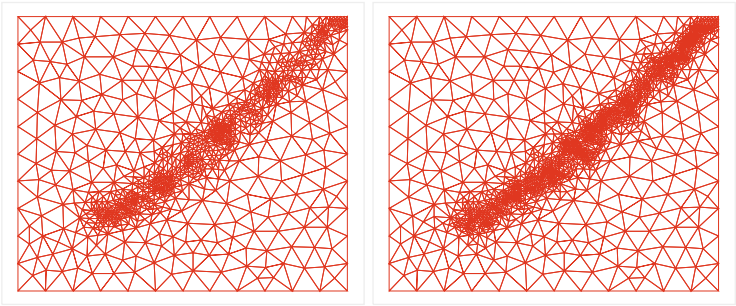


FIG. 14. Meshes generated by Algorithm 5 for problem (4.1) with $\delta = 0.001$ and $p = 1$ (*left*), $p = 2$ (*right*)

In Table 14 we present the number of elements, the true $\mathcal{L}^2$ errors, and the global $\mathcal{L}^2$ effectivity indices at $t = 0, 0.25, 0.5, 0.75, 1$ and $p = 1, 2$ which show that the error estimates converge to the true error during the simulation. These computational results indicate that our estimators are accurate on adaptively refined unstructured triangular meshes and further suggest that they converge to the true error under adaptive mesh refinement of Algorithm 6.

EXAMPLE 3. We consider the following linear problem

$$2u_x + u_y = 0, \qquad (x, y) \in [0, 1]^2, \tag{4.3a}$$

subject to the boundary conditions

$$u(x, 0) = e^{-x}, \qquad 0 \le x \le 1, \tag{4.3b}$$

$$u(0, y) = e^{2y} + .25, \qquad 0 < y \le 1, \tag{4.3c}$$

with the true solution having a contact discontinuity along $y = x/2$

$$u(x, y) = \begin{cases} e^{2y-x} & \text{if } x \ge y \\ e^{2y-x} + .25 & \text{if } x < y \end{cases}. \tag{4.3d}$$

We solve (4.3) on the unstructured mesh having 100 elements shown in Fig. 1 with the spaces $\mathcal{U}_p$, $p = 1, 2, 3, 4$ and apply the modified DG method (2.6) and (2.7) with the adaptive refinement strategy described in Algorithm 3 for $\delta = 0.01$, 0.005, and 0.001. We present the mesh in Fig. 18 for $\delta = 0.001$ and show in Table 15 the number of elements N, the global $\mathcal{L}^2$ norm of the error, and the global $\mathcal{L}^2$ effectivity indices with $\delta = 0.01$, 0.005, and 0.001 and $p = 1, 2, 3, 4$. These results suggest that the error estimates converge to the true error under local adaptive mesh refinement algorithm that refines elements near the discontinuity and whose errors are underestimated.

TABLE 14
$\mathcal{L}^2$ errors and global effectivity indices for problem (4.2) on unstructured meshes having N elements with $p = 1, 2$ at $t = 0, 0.25, 0.5, 0.75, 1$

	$p=1$			$p=2$		
t	N	$\|\|e\|\|_{\mathcal{L}^2}$	θ	N	$\|\|e\|\|_{\mathcal{L}^2}$	θ
0.00	432	1.4402e−2	0.5860	432	2.7326e−3	0.4879
0.25	834	8.1633e−3	0.6743	984	1.1912e−3	0.5679
0.50	1,895	5.3067e−3	0.7654	2,348	6.2960e−4	0.6974
0.75	3,469	3.7288e−3	0.8904	4,467	3.6869e−4	0.8279
1.00	5,737	2.4394e−4	0.9747	7,404	1.4292e−5	0.9372

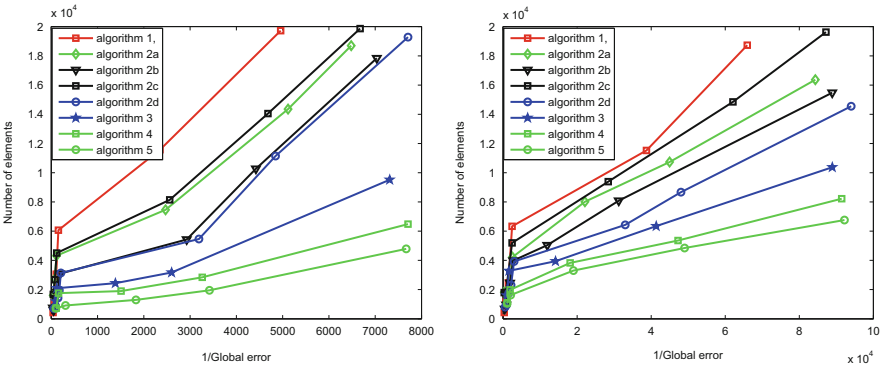


FIG. 15. True $\mathcal{L}^2$ errors versus the number of elements for Algorithms 1–6 with $p = 1$ (*left*) and $p = 2$ (*right*) for Problem (4.1)

TABLE 15
Global effectivity indices and final number of elements N for problem (4.3) with $p = 1, 2, 3, 4$ and error tolerances of $\delta = 0.01, 0.005, 0.001$

	$p=1$		$p=2$		$p=3$		$p=4$	
δ	θ	N	θ	N	θ	N	θ	N
0.01	0.9280	100	0.9255	100	0.9242	100	0.9348	100
0.005	0.9628	139	0.9456	121	0.9399	116	0.9482	116
0.001	0.9941	738	0.9928	523	0.9864	492	0.9858	426

5. Conclusions. We tested the residual-based a posteriori DG error estimates of Adjerid and Baccouch [3] for hyperbolic problems on unstructured triangular meshes. Several computational examples suggest that the Taylor's series residual-based error estimates proposed in this manuscript converge to the true error under local mesh refinement and in the presence of discontinuities. Similarly, we expect that the error estimates proposed by Adjerid and Mechai [6] on tetrahedral meshes will also converge to the true error under adaptive mesh refinement. Our future work will focus on

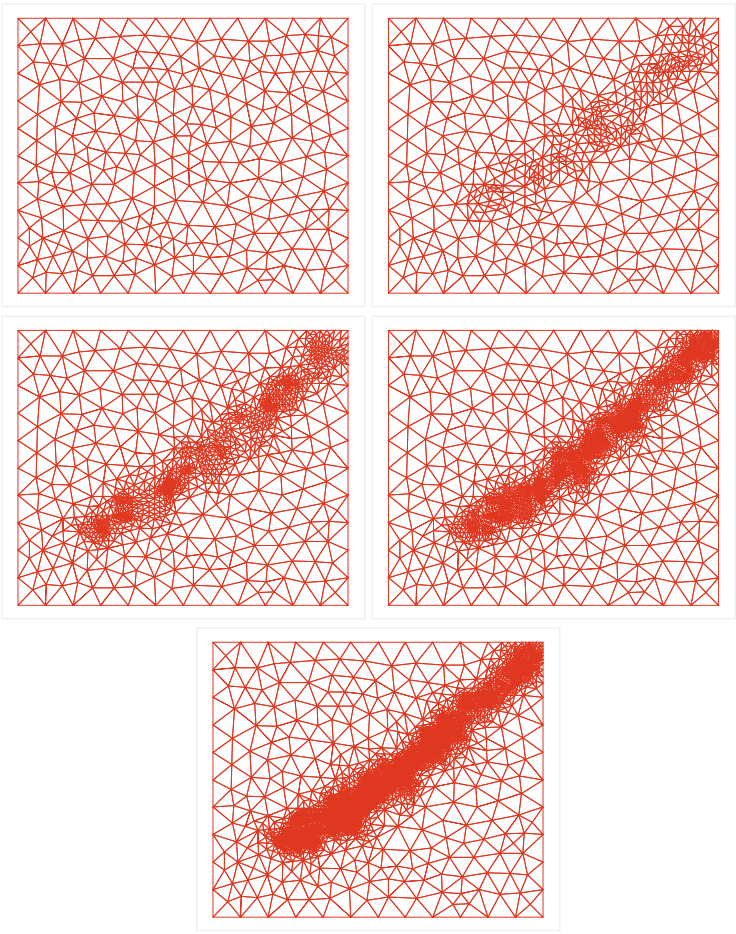


FIG. 16. Adaptive meshes for problem (4.2) with $p = 1$ at $t = 0.00, 0.25, 0.50, 0.75, 1.00$ (*upper left to lower right*)

applying adaptive refinement strategies to multi-dimensional hyperbolic systems of conservation laws and reach similar conclusions for general unstructured tetrahedral meshes.

Acknowledgments. The authors are grateful to Bryan Johnson (undergraduate student at the University of Nebraska at Omaha) for applying the adaptive algorithms to the contact problem to generate the results for Example 3.

The work of the Slimane Adjerid author was supported in part by NSF grant DMS-0809262. The work of the Mahboub Baccouch author was supported by the NASA Nebraska Space Grant Program and UCRCA at the University of Nebraska at Omaha.

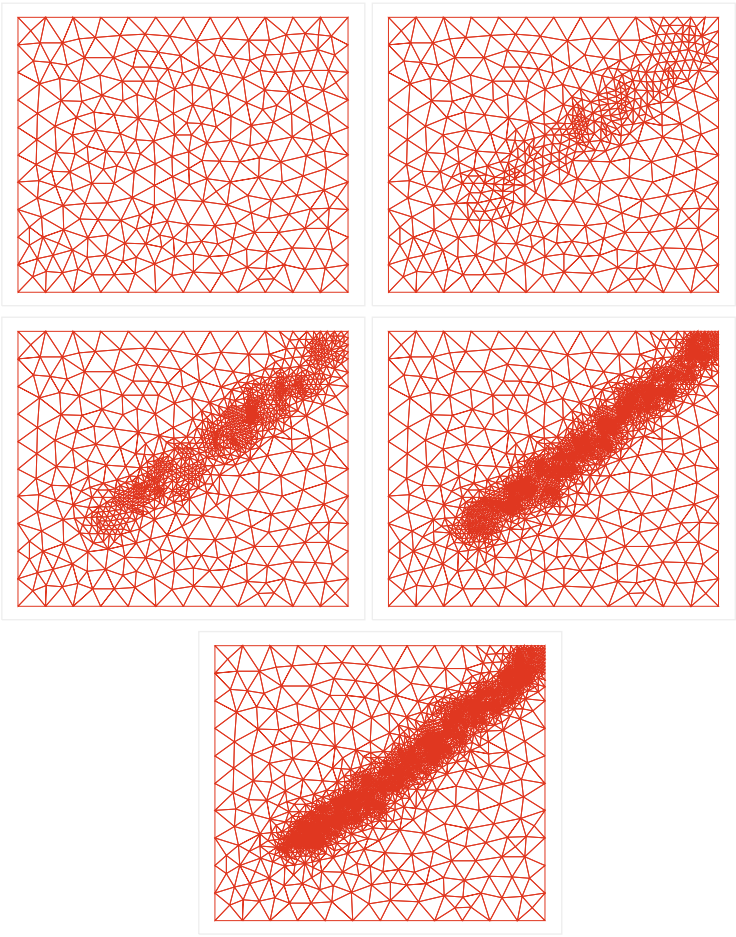


FIG. 17. Adaptive meshes for problem (4.2) with $p = 2$ at $t = 0.00, 0.25, 0.50, 0.75, 1.00$ (*upper left to lower right*)

REFERENCES

[1] S. Adjerid and M. Baccouch. The discontinuous Galerkin method for two-dimensional hyperbolic problems part I: Superconvergence error analysis. *J. Sci. Comput.*, 33(1):75–113, 2007.

[2] S. Adjerid and M. Baccouch. The discontinuous Galerkin method for two-dimensional hyperbolic problems part II: A posteriori error estimation. *J. Sci. Comput.*, 38(1):15–49, 2008.

[3] S. Adjerid and M. Baccouch. *A Posteriori* error analysis of the discontinuous Galerkin method for two-dimensional hyperbolic problems on unstructured meshes. *Computer Methods in Applied Mechanics and Engineering*, 200: 162–177, 2011.

[4] S. Adjerid, K. Devine, J. Flaherty, and L. Krivodonova. *A posteriori* error estimation for discontinuous Galerkin solutions of hyperbolic problems. *Computer Methods in Applied Mechanics and Engineering*, 191:1097–1112, 2002.

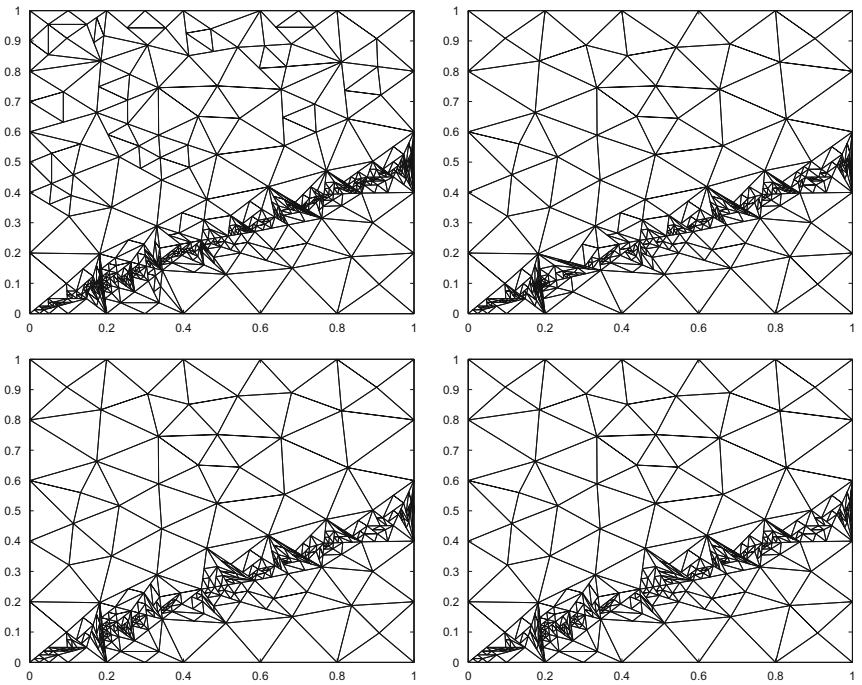


FIG. 18. Meshes generated by adaptive Algorithm 3 for problem (4.3) with $\delta = 0.001$ and $p = 1, 2, 3, 4$ (*upper left to lower right*)

[5] S. Adjerid and T. C. Massey. *A posteriori* discontinuous finite element error estimation for two-dimensional hyperbolic problems. *Computer Methods in Applied Mechanics and Engineering*, 191:5877–5897, 2002.

[6] S. Adjerid and I. Mechai. *A posteriori* discontinuous Galerkin error estimation on tetrahedral meshes. *Computer Methods in Applied Mechanics and Engineering*, 201–204:157–178, 2012.

[7] S. Adjerid and T. Weinhart. Discontinuous Galerkin error estimation for linear symmetric hyperbolic systems. *Computer Methods in Applied Mechanics and Engineering*, 198:3113–3129, 2009.

[8] S. Adjerid and T. Weinhart. Asymptotically exact discontinuous Galerkin error estimates for linear symmetric hyperbolic systems. *Applied Numerical Mathematics*, in press, 2011.

[9] S. Adjerid and T. Weinhart. Discontinuous Galerkin error estimation for linear symmetrizable hyperbolic systems. *Mathematics of Computation*, 80: 1335–1367, 2011.

[10] 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.*, 131:267–279, 1997.

[11] R. Biswas, K. Devine, and J. E. Flaherty. Parallel adaptive finite element methods for conservation laws. *Applied Numerical Mathematics*, 14:255–284, 1994.

[12] B. Cockburn, G. E. Karniadakis, and C. W. Shu, editors. *Discontinuous Galerkin Methods Theory, Computation and Applications, Lectures Notes in Computational Science and Engineering*, volume 11. Springer, Berlin, 2000.

[13] B. Cockburn, S. Y. Lin, and C. W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin methods of scalar conservation laws III: One dimensional systems. *Journal of Computational Physics*, 84:90–113, 1989.

[14] B. Cockburn and C. W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin methods for scalar conservation laws II: General framework. *Mathematics of Computation*, 52:411–435, 1989.

[15] K. D. Devine and J. E. Flaherty. Parallel adaptive hp-refinement techniques for conservation laws. *Computer Methods in Applied Mechanics and Engineering*, 20:367–386, 1996.

[16] K. Ericksson and C. Johnson. Adaptive finite element methods for parabolic problems I: A linear model problem. *SIAM Journal on Numerical Analysis*, 28: 12–23, 1991.

[17] J. E. Flaherty, R. 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. *Journal of Parallel and Distributed Computing*, 47:139–152, 1997.

[18] G. E. Karniadakis and S. J. Sherwin. *Spectral/hp Element Methods for CFD*. Oxford University Press, New York, 1999.

[19] L. Krivodonova and J. E. Flaherty. Error estimation for discontinuous Galerkin solutions of two-dimensional hyperbolic problems. *Advances in Computational Mathematics*, 19:57–71, 2003.

[20] 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, Los Alamos, 1973.

A QUADRATIC C^0 INTERIOR PENALTY METHOD FOR AN ELLIPTIC OPTIMAL CONTROL PROBLEM WITH STATE CONSTRAINTS

S.C. BRENNER, L.-Y. SUNG, AND Y. ZHANG*

Abstract. We consider an elliptic distributed optimal control problem on convex polygonal domains with pointwise state constraints and solve it as a fourth order variational inequality for the state by a quadratic C^0 interior penalty method. The error for the state in an H^2-like energy norm is $O(h^\alpha)$ on quasi-uniform meshes (where $\alpha \in (0,1]$ is determined by the interior angles of the domain) and $O(h)$ on graded meshes. The error for the control in the L_2 norm has the same behavior. Numerical results that illustrate the performance of the method are also presented.

Key words. Elliptic distributed optimal control problem, Pointwise state constraints, Simply supported plate, Fourth order variational inequality, Finite element, C^0 interior penalty method, Discontinuous Galerkin, Graded meshes

AMS(MOS) subject classifications. 65K15, 65N30.

1. Introduction.

Let Ω be a bounded convex polygonal domain in $\mathbb{R}^2$, $y_d \in L_2(\Omega)$, $\gamma \geq 0$ and $\beta > 0$ be constants. The following problem [33] is a model elliptic distributed optimal control problem with pointwise state constraints:

Find the minimizer of the functional

$$J(y,u) = \frac{\gamma}{2}\int_\Omega (y-y_d)^2\,dx + \frac{\beta}{2}\int_\Omega u^2\,dx, \tag{1.1}$$

where $(y,u) \in H_0^1(\Omega)\times L_2(\Omega)$ are subjected to the constraints

$$\int_\Omega \nabla y\cdot\nabla v\,dx = \int_\Omega uv\,dx \qquad \forall\, v\in H_0^1(\Omega), \tag{1.2}$$

$$\psi_1 \leq y \leq \psi_2 \qquad \text{a.e. in } \Omega. \tag{1.3}$$

Here the functions $\psi_1(x), \psi_2(x) \in C^2(\Omega)\cap C(\bar\Omega)$ satisfy

$$\psi_1 < \psi_2 \qquad \text{in } \Omega, \tag{1.4a}$$

$$\psi_1 < 0 < \psi_2 \qquad \text{on } \partial\Omega. \tag{1.4b}$$

Since Ω is convex, elliptic regularity [36, 45, 58] implies that (1.2) is equivalent to $y \in H^2(\Omega)\cap H_0^1(\Omega)$ and $u = -\Delta y$. Note that [46, Theorem 2.2.1]

$$\int_\Omega (\Delta v)(\Delta w)\,dx = \int_\Omega (D^2v : D^2w)\,dx \qquad \forall\, v,w\in H^2(\Omega)\cap H_0^1(\Omega), \tag{1.5}$$

*Department of Mathematics and Center for Computation and Technology, Louisiana State University, Baton Rouge, LA 70803, USA, brenner@math.lsu.edu; sung@math.lsu.edu; yzhan112@utk.edu

X. Feng et al. (eds.), *Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations*, The IMA Volumes in Mathematics and its Applications 157, DOI 10.1007/978-3-319-01818-8_4,

where

$$D^2v : D^2w = \sum_{1\le i,j\le 2} \left(\frac{\partial^2 v}{\partial x_i \partial x_j}\right)\left(\frac{\partial^2 w}{\partial x_i \partial x_j}\right)$$

is the (Frobenius) inner product of the Hessian matrices of v and w. Therefore we can solve the optimal control problem (1.1)–(1.3) by looking for the minimizer of the reduced functional

$$\hat{J}(y) = \frac{\gamma}{2}\int_\Omega (y - y_d)^2\,dx + \frac{\beta}{2}\int_\Omega (D^2y : D^2y)\,dx$$

in the set

$$K = \{v \in H^2(\Omega) \cap H_0^1(\Omega) : \psi_1 \le v \le \psi_2 \text{ in } \Omega\}. \tag{1.6}$$

A simple calculation shows that this is equivalent to the following problem:

$$\text{Find} \quad \bar{y} = \operatorname*{argmin}_{y\in K}\left[\frac{1}{2}\mathcal{A}(y,y) - (f,y)\right], \tag{1.7}$$

where $f = \gamma y_d$, $(\cdot,\cdot)$ is the inner product of $L_2(\Omega)$, and

$$\mathcal{A}(v,w) = \int_\Omega \left[\beta(D^2v : D^2w) + \gamma vw\right] dx. \tag{1.8}$$

Since (1.4) implies that K is a nonempty closed convex subset of $H^2(\Omega)\cap H_0^1(\Omega)$ and the bilinear form $\mathcal{A}(\cdot,\cdot)$ is symmetric, bounded, and coercive on $H^2(\Omega)\cap H_0^1(\Omega)$, we can apply the standard theory [43, 52, 54, 59] to conclude that the problem (1.7) has a unique solution $\bar{y} \in K$ characterized by the variational inequality

$$\mathcal{A}(\bar{y}, y - \bar{y}) \ge (f, y - \bar{y}) \qquad \forall\, y \in K. \tag{1.9}$$

The solution of the optimal control problem is then given by $(\bar{y}, \bar{u})$, where $\bar{u} = -\Delta\bar{y}$. Note that (1.7) becomes the displacement obstacle problem for simply supported Kirchhoff plates if we take γ to be 0. For this reason we will also refer to (1.7) as an obstacle problem.

According to the regularity results in [32, 41, 42] for fourth order obstacle problems, the solution $\bar{y}$ of (1.7) belongs to $H^3_{\text{loc}}(\Omega)\cap C^2(\Omega)$ under our assumptions on the functions y_d, ψ_1, and ψ_2. Note that (1.4b) implies that the constraints are inactive near $\partial\Omega$ and hence

$$\beta\Delta^2\bar{y} + \gamma\bar{y} = f$$

near $\partial\Omega$. It then follows from the elliptic regularity theory for the biharmonic equation (cf. [8] and Appendix A) that there exists $\alpha \in (0,1]$ (determined by the interior angles of Ω) such that $\bar{y} \in H^{2+\alpha}(\mathcal{N})$ in a

neighborhood $\mathcal{N}$ of $\partial\Omega$ disjoint from the active set. Thus globally $\bar{y}$ belongs to $H^{2+\alpha}(\Omega)$. We shall refer to α as the index of elliptic regularity for the obstacle problem (1.7).

A main difficulty in the analysis of finite element methods for fourth order obstacle problems is that the solutions in general do not belong to $H^4_{\text{loc}}(\Omega)$ even for smooth data, which means that the complementarity form of the variational inequality (1.9) in general only exists in a weak sense. In contrast, the solutions of second order obstacle problems belong to $H^2(\Omega)$ under appropriate assumptions on the data (cf. [29, 53]). Hence the complementarity forms of the variational inequalities arising from second order obstacle problems exist in the strong sense, which is a crucial ingredient for the derivations of optimal error estimates in [30, 31, 40].

A new approach to the obstacle problem for clamped Kirchhoff plates on convex polygonal domains was introduced in [25], where optimal error estimates were obtained for C^1 finite element methods, classical nonconforming finite element methods, and discontinuous Galerkin methods. The results were later extended to general domains and general Dirichlet boundary conditions in [15, 23, 24]. This new approach does not rely on the complementarity forms of the variational inequalities and hence can bypass the aforementioned difficulty. The goal of this paper is to extend the results in [23] to (1.7)/(1.9), which covers both obstacle problems for simply supported plates and optimal control problems with pointwise state constraints. We will show that the magnitude of the error in the energy norm is $O(h^\alpha)$ on quasi-uniform meshes and $O(h)$ on graded meshes.

Finite element methods for state constrained elliptic optimal control problems were investigated in [37, 56], where the finite element approximation $(\bar{y}_h, \bar{u}_h)$ of $(\bar{y}, \bar{u})$ is obtained from discrete versions of the optimal control problems. In this approach the error analysis for the state and the error analysis for the control are coupled and hence the estimates for $|\bar{y}-\bar{y}_h|_{H^1(\Omega)}$ and $\|\bar{u}-\bar{u}_h\|_{L_2(\Omega)}$ have the same magnitude, which in the case of a rectangle with quasi-uniform meshes is $O(h^{1-\epsilon})$. In our approach we obtain instead an error estimate for the approximation $\bar{y}_h$ of $\bar{y}$ in an H^2-like energy norm, which then implies an error estimate in the L_2 norm for the approximation $\bar{u}_h$ of $\bar{u}$ (generated from $\bar{y}_h$ by a postprocessing procedure) with the same magnitude. In the case of a rectangle with quasi-uniform meshes, the magnitudes of these errors are $O(h)$. On the other hand, the convergence of $\bar{y}_h$ in the $H^1(\Omega)$ norm and the $L_\infty(\Omega)$ norm, which are weaker than the energy norm, can be expected to be of higher order. This is indeed observed in our numerical experiments, where the magnitudes of the errors of $\bar{y}_h$ in the $H^1(\Omega)$ norm and the $L_\infty(\Omega)$ norm are $O(h^2)$ for a rectangle.

The optimal control problem defined by (1.1)–(1.3) is solved as a fourth order variational inequality in [55] by a Morley finite element method and in [44] by a mixed finite element method. However the analyses in [44, 55] rely on additional assumptions on the active set first introduced in [7]. Our new approach for fourth order obstacle problems may provide an error

analysis for the finite element methods in [44, 55] without the additional assumptions on the active set.

Other numerical methods for (1.1)–(1.3) are investigated, for example, in [5, 6, 34, 48–51, 57, 60].

The rest of the paper is organized as follows. We introduce a quadratic C^0 interior penalty method for (1.7) in Sect. 2 and an intermediate obstacle problem that connects the continuous and discrete obstacle problems in Sect. 3. Section 4 contains several preliminary estimates which are useful for the convergence analysis carried out in Sect. 5. Numerical results that illustrate the performance of our method are presented in Sect. 6, followed by some concluding remarks in Sect. 7. Elliptic regularity results for simply supported plates, which play an important role in the error analysis, are summarized in Appendix A. Some technical results concerning an enriching operator that connects the discrete and continuous spaces are given in Appendix B.

We will follow the notation for Sobolev spaces and norms in [20, 35]. Throughout the paper we will denote by C a generic positive constant independent of mesh sizes that can take different values at different occurrences. To avoid the proliferation of constants, we will also use $A \lesssim B$ (or $B \gtrsim A$) to denote the statement that $A \leq (\text{constant})B$, where the positive constant is independent of mesh sizes. The statement $A \approx B$ is equivalent to $A \lesssim B$ and $B \lesssim A$.

2. A Quadratic C^0 Interior Penalty Method. C^0 interior penalty methods were introduced in [39] for fourth order elliptic boundary value problems. They were further studied in [13, 16, 18, 21] and fast solvers for C^0 interior methods were developed in [22, 26, 27]. Adaptive [17] and isoparametric [19] versions of C^0 interior penalty methods are also available. Below we will recall the notation for C^0 interior penalty methods and introduce the discrete obstacle problem for (1.7).

2.1. Triangulation. Let $\mathcal{T}_h$ be a simplicial triangulation of Ω that is regular (i.e., $\mathcal{T}_h$ satisfies a minimum angle condition). We will use the following notation throughout the paper.

- h_T is the diameter of the triangle T.
- h is a mesh parameter proportional to $\max_{T\in\mathcal{T}_h} h_T$.
- v_T is the restriction of the function v to the triangle T.
- $\mathcal{E}_h$ is the set of the edges of the triangles in $\mathcal{T}_h$.
- $\mathcal{E}_h^i$ is the subset of $\mathcal{E}_h$ consisting of edges interior to Ω.
- $\mathcal{E}_h^b$ is the subset of $\mathcal{E}_h$ consisting of edges along $\partial\Omega$.
- $|e|$ is the length of an edge e.
- $\mathcal{V}_h$ is the set of the vertices of the triangles in $\mathcal{T}_h$.
- $\mathcal{V}_T$ is the set of the three vertices of T.
- $\mathcal{E}_{\mathcal{V}_T}^i$ is the set of the edges in $\mathcal{E}_h^i$ emanating from the vertices of T.
- $\mathcal{T}_T$ is the set of triangles sharing a vertex with T.
- $\mathcal{S}_T$ is the interior of the closure of $\cup_{T'\in\mathcal{T}_T} T'$.
- $\mathcal{T}_p$ is the set of the triangles in $\mathcal{T}_h$ that share the common vertex p.

- $\mathcal{T}_e$ is the set of the triangles in $\mathcal{T}_h$ that share the common edge e.
- $|\mathcal{T}_p|$ (resp. $|\mathcal{T}_e|$) is the number of triangles in $\mathcal{T}_p$ (resp. $\mathcal{T}_e$).
- Let $e \in \mathcal{E}_h^b$. Then T_e is the triangle in $\mathcal{T}_h$ such that $\mathcal{T}_e = \{T_e\}$.

We will consider both quasi-uniform and graded triangulations. For a quasi-uniform triangulation $\mathcal{T}_h$, we have

$$h_T \approx h \qquad \forall T \in \mathcal{T}_h. \tag{2.1}$$

Let $p_1, \ldots, p_L$ be the corners of Ω and ω_ℓ be the interior angle at p_ℓ for $1 \le \ell \le L$. For a graded triangulation $\mathcal{T}_h$, we have

$$h_T \approx h\Phi(c_T) \qquad \forall T \in \mathcal{T}_h, \tag{2.2}$$

where c_T is the center of T,

$$\Phi(x) = \prod_{\ell=1}^{L} |p_\ell - x|^{1-\alpha_\ell}, \tag{2.3}$$

and the grading parameters $\alpha_\ell > 0$ are determined as follows:

$$\begin{cases} \alpha_\ell = 1 & \text{if } \omega_\ell \le \dfrac{\pi}{2}, \\ \alpha_\ell < \left(\dfrac{\pi}{\omega_\ell}\right) - 1 & \text{if } \dfrac{\pi}{2} < \omega_\ell < \pi. \end{cases} \tag{2.4}$$

Note that (2.2) and (2.3) imply

$$h_T^{\alpha_\ell} \approx h \tag{2.5}$$

if $T \in \mathcal{T}_h$ touches the corner p_ℓ.

REMARK 2.1. *We can take* $\alpha = \min_{1 \le \ell \le L} \alpha_\ell$ *to be the index of elliptic regularity* (*cf. Appendix A*).

REMARK 2.2. *The construction of regular triangulations that satisfy* (2.2) *is discussed, for example, in [1, 10, 14].*

2.2. Jumps and Averages. The jumps and averages of the normal derivatives for functions in the piecewise Sobolev spaces

$$H^s(\Omega, \mathcal{T}_h) = \{v \in L_2(\Omega) : v_T = v|_T \in H^s(T) \quad \forall T \in \mathcal{T}_h\}$$

are defined as follows.

Let $e \in \mathcal{E}_h^i$ be the common edge of $T_\pm \in \mathcal{T}_h$ and n_e be the unit normal of e pointing from T_- to T_+. We define on e

$$\left\{\!\!\left\{ \frac{\partial^2 v}{\partial n^2} \right\}\!\!\right\} = \frac{1}{2}\left(\left.\frac{\partial^2 v_+}{\partial n_e^2}\right|_e + \left.\frac{\partial^2 v_-}{\partial n_e^2}\right|_e \right) \qquad \forall v \in H^s(\Omega, \mathcal{T}_h), s > \frac{5}{2}, \tag{2.6a}$$

$$\left[\!\!\left[\frac{\partial v}{\partial n} \right]\!\!\right] = \left.\frac{\partial v_+}{\partial n_e}\right|_e - \left.\frac{\partial v_-}{\partial n_e}\right|_e \qquad \forall v \in H^2(\Omega, \mathcal{T}_h), \tag{2.6b}$$

where $v_\pm = v\big|_{T_\pm}$. Similarly, we define on e

$$\left\{\!\!\left\{\frac{\partial v}{\partial n_e}\right\}\!\!\right\} = \frac{1}{2}\left(\frac{\partial v_+}{\partial n_e}\bigg|_e + \frac{\partial v_-}{\partial n_e}\bigg|_e\right) \qquad \forall v \in H^2(\Omega,\mathcal{T}_h), \tag{2.7a}$$

$$\left[\!\!\left[\frac{\partial^2 v}{\partial n_e^2}\right]\!\!\right] = \frac{\partial^2 v_+}{\partial n_e^2}\bigg|_e - \frac{\partial^2 v_-}{\partial n_e^2}\bigg|_e \qquad \forall v \in H^s(\Omega,\mathcal{T}_h), s > \frac{5}{2}. \tag{2.7b}$$

REMARK 2.3. *Note that the definitions for the average $\{\!\{\partial^2 v/\partial n^2\}\!\}$ and the jump $[\![\partial v/\partial n]\!]$ in (2.6), which appear in C^0 interior penalty methods, are independent of the choice of $T_\pm$ (or n_e). On the other hand, the definitions in (2.7) for $\{\!\{\partial v/\partial n_e\}\!\}$ and $[\![\partial^2 v/\partial n_e^2]\!]$, which appear only in the analysis, do depend on the choice of $T_\pm$ (or n_e). However their product is also independent of the choice of $T_\pm$ (or n_e).*

Let $e \in \mathcal{E}_h^b$ be a boundary edge and n_e be the unit normal of e pointing towards the outside of Ω. We define on e

$$\left\{\!\!\left\{\frac{\partial v}{\partial n_e}\right\}\!\!\right\} = \frac{\partial v}{\partial n_e}\bigg|_e \qquad \forall v \in H^2(\Omega,\mathcal{T}_h), \tag{2.8a}$$

$$\left[\!\!\left[\frac{\partial^2 v}{\partial n_e^2}\right]\!\!\right] = -\frac{\partial^2 v}{\partial n_e^2}\bigg|_e \qquad \forall v \in H^s(\Omega,\mathcal{T}_h), s > \frac{5}{2}. \tag{2.8b}$$

2.3. The Discrete Obstacle Problem. Let $V_h \subset H_0^1(\Omega)$ be the $\mathbb{P}_2$ Lagrange finite element space associated with $\mathcal{T}_h$ whose members vanish on $\partial\Omega$. We define the bilinear form $a_h(\cdot,\cdot)$ on $V_h \times V_h$ by

$$\begin{aligned} a_h(v,w) = &\sum_{T\in\mathcal{T}_h}\int_T (D^2 v : D^2 w)\,dx + \sum_{e\in\mathcal{E}_h^i}\int_e \{\!\{\partial^2 v/\partial n^2\}\!\}[\![\partial w/\partial n]\!]\,ds \\ &+ \sum_{e\in\mathcal{E}_h^i}\int_e \{\!\{\partial^2 w/\partial n^2\}\!\}[\![\partial v/\partial n]\!]\,ds \\ &+ \sigma\sum_{e\in\mathcal{E}_h^i}|e|^{-1}\int_e [\![\partial v/\partial n]\!][\![\partial w/\partial n]\!]\,ds, \end{aligned} \tag{2.9}$$

where $\sigma > 0$ is a penalty parameter. Note that $a_h(\cdot,\cdot)$ is a consistent bilinear form for the biharmonic equation with the boundary conditions of simply supported plates.

It follows from (2.6a) and scaling that

$$\sum_{e\in\mathcal{E}_h^i}|e|\|\{\!\{\partial^2 v/\partial n^2\}\!\}\|_{L_2(e)}^2 \lesssim \sum_{T\in\mathcal{T}_h}|v|_{H^2(T)}^2 \qquad \forall v \in V_h. \tag{2.10}$$

Therefore, for sufficiently large σ, we have (cf. [21])

$$a_h(v,v) \gtrsim \left(\sum_{T\in\mathcal{T}_h}|v|_{H^2(T)}^2 + \sum_{e\in\mathcal{E}_h^i}|e|^{-1}\|[\![\partial v/\partial n]\!]\|_{L_2(e)}^2\right) \qquad \forall v \in V_h. \tag{2.11}$$

The discrete bilinear form that approximates $\mathcal{A}(\cdot,\cdot)$ is then given by

$$\mathcal{A}_h(v,w) = \beta a_h(v,w) + \gamma(v,w), \tag{2.12}$$

and

$$\|v\|_h = \left[\beta\left(\sum_{T\in\mathcal{T}_h}|v|^2_{H^2(T)} + \sum_{e\in\mathcal{E}_h^i}|e|^{-1}\|[\![\partial v/\partial n]\!]\|^2_{L_2(e)}\right) + \gamma\|v\|^2_{L_2(\Omega)}\right]^{\frac{1}{2}} \tag{2.13}$$

is the mesh-dependent energy norm. It follows from (2.10)–(2.13) that

$$|\mathcal{A}_h(v,w)| \lesssim \|v\|_h\|w\|_h \qquad \forall v,w\in V_h, \tag{2.14}$$

$$\mathcal{A}_h(v,v) \gtrsim \|v\|_h^2 \qquad \forall v\in V_h, \tag{2.15}$$

provided that σ is sufficiently large, which we assume to be the case.

Note that

$$\|v\|^2_{H^1(\Omega)} \lesssim \sum_{T\in\mathcal{T}_h}|v|^2_{H^2(T)} + \sum_{e\in\mathcal{E}_h^i}|e|^{-1}\|[\![\partial v/\partial n]\!]\|^2_{L_2(e)} \tag{2.16}$$

for all $v\in H^2(\Omega,\mathcal{T}_h)\cap H_0^1(\Omega)$ by a Poincaré–Friedrichs inequality [28, Example 5.4], and hence

$$\|v\|_{H^1(\Omega)} \lesssim \|v\|_h \tag{2.17}$$

for all $v\in H^2(\Omega,\mathcal{T}_h)\cap H_0^1(\Omega)$ ($\supset V_h + H^2(\Omega)\cap H_0^1(\Omega)$).

We can now define the discrete obstacle problem for (1.7):

$$\text{Find}\quad \bar{y}_h = \operatorname*{argmin}_{y_h\in K_h}\left[\frac{1}{2}\mathcal{A}_h(y_h,y_h) - (f,y_h)\right], \tag{2.18}$$

where

$$K_h = \{v\in V_h : \psi_1(p)\le v(p)\le\psi_2(p)\quad\forall p\in\mathcal{V}_h\}. \tag{2.19}$$

Let Π_h be the nodal interpolation operator for the $\mathbb{P}_2$ Lagrange finite element space. Then Π_h maps $H^2(\Omega)\cap H_0^1(\Omega)$ into V_h and K into K_h. Therefore K_h is a nonempty closed convex subset of V_h. Moreover the bilinear form $\mathcal{A}_h(\cdot,\cdot)$ is symmetric positive definite by (2.15). Hence the discrete problem (2.18) has a unique solution $\bar{y}_h\in K_h$ characterized by the discrete variational inequality:

$$\mathcal{A}_h(\bar{y}_h, y_h-\bar{y}_h) \ge (f, y_h-\bar{y}_h) \qquad \forall y_h\in K_h. \tag{2.20}$$

Let Π_T be the nodal interpolation operator for the $\mathbb{P}_2$ Lagrange finite element on a triangle T. We have a standard local interpolation error estimate [20, 35]

$$\sum_{m=0}^{2} h_T^{m-2}|\zeta - \Pi_T\zeta|_{H^m(T)} \lesssim h_T^s|\zeta|_{H^{2+s}(T)} \tag{2.21}$$

for all $\zeta \in H^{2+s}(T)$, $T \in \mathcal{T}_h$ and $s \in [0,1]$.

The following lemma provides global interpolation error estimates for the solution $\bar{y}$ of (1.7)/(1.9).

LEMMA 2.1. *There exists a positive constant C independent of h such that*

$$\|\bar{y} - \Pi_h\bar{y}\|_h \leq Ch^\tau,$$

where $\tau = \alpha$ if $\mathcal{T}_h$ is quasi-uniform and $\tau = 1$ if $\mathcal{T}_h$ is graded according to (2.2)–(2.4).

Proof. Since the estimate for a quasi-uniform $\mathcal{T}_h$ is standard (cf. [21]), we will focus on a graded $\mathcal{T}_h$. Let $\mathcal{T}_h^I$ be the set of triangles in $\mathcal{T}_h$ that do not touch any corner of Ω and $\mathcal{T}_h^C = \mathcal{T}_h \setminus \mathcal{T}_h^I = \cup_{1\leq\ell\leq L}\mathcal{T}_{h,\ell}^C$, where $\mathcal{T}_{h,\ell}^C$ is the set of the triangles that touch the corner p_ℓ.

Since $\bar{y}_T \in H^3(T)$ for $T \in \mathcal{T}_h^I$ (cf. Appendix A), we have, by (2.21),

$$\begin{aligned}\sum_{T\in\mathcal{T}_h^I}\sum_{m=0}^{2} h_T^{2(m-2)}|\bar{y} - \Pi_h\bar{y}|_{H^m(T)}^2 \\ \lesssim \sum_{T\in\mathcal{T}_h^I}\big(\Phi^{-2}(c_T)h_T^2\big)\Phi^2(c_T)|\bar{y}|_{H^3(T)}^2\end{aligned} \tag{2.22}$$

where the function Φ is defined in (2.3).

It follows from (2.2), (2.3), (2.22), and (A.7) that

$$\sum_{T\in\mathcal{T}_h^I}\sum_{m=0}^{2} h_T^{2(m-2)}|\bar{y} - \Pi_h\bar{y}|_{H^m(T)}^2 \lesssim h^2. \tag{2.23}$$

Let $T \in \mathcal{T}_{h,\ell}^C$ be a triangle that touches a corner p_ℓ. Then $\bar{y} \in H^{2+\alpha_\ell}(T)$ (cf. Appendix A) and we have, by (2.21),

$$\sum_{m=0}^{2} h_T^{m-2}|\bar{y} - \Pi_h\bar{y}|_{H^m(T)} \lesssim h_T^{\alpha_\ell}|\bar{y}|_{H^{2+\alpha_\ell}(T)} \qquad \forall\, T \in \mathcal{T}_{h,\ell}^C. \tag{2.24}$$

It follows from (2.5) and (2.24) that

$$\sum_{T\in\mathcal{T}_h^C}\sum_{m=0}^{2} h_T^{2(m-2)}|\bar{y} - \Pi_h\bar{y}|_{H^m(T)}^2 \tag{2.25}$$

$$= \sum_{\ell=1}^{L} \sum_{T \in \mathcal{T}_{h,\ell}^C} \sum_{m=0}^{2} h_T^{2(m-2)} |\bar{y} - \Pi_h \bar{y}|_{H^m(T)}^2 \lesssim h^2.$$

Combining (2.23) and (2.25), we find

$$\sum_{T \in \mathcal{T}_h} \sum_{m=0}^{2} h_T^{2(m-2)} |\bar{y} - \Pi_h \bar{y}|_{H^m(T)}^2 \lesssim h^2. \tag{2.26}$$

By the trace theorem with scaling, (2.6b) and (2.26), we also have

$$\begin{aligned} &\sum_{e \in \mathcal{E}_h^i} |e|^{-1} \|[\![\partial(\bar{y} - \Pi_h \bar{y})/\partial n]\!]\|_{L_2(e)}^2 \qquad (2.27) \\ &\lesssim \sum_{T \in \mathcal{T}_h} \left(h_T^{-2} |\bar{y} - \Pi_h \bar{y}|_{H^1(T)}^2 + |\bar{y} - \Pi_h \bar{y}|_{H^2(T)}^2 \right) \lesssim h^2. \end{aligned}$$

The lemma for a graded $\mathcal{T}_h$ follows from (2.13), (2.16), (2.26), and (2.27). □

3. An Intermediate Obstacle Problem. As mentioned in Sect. 1, the difficulties due to the lack of $H^4_{\text{loc}}(\Omega)$ regularity can be bypassed if the convergence analysis does not rely on the complementarity form of the variational inequality (1.9). We can accomplish this by introducing the following intermediate obstacle problem:

$$\text{Find} \qquad \bar{y}_h^* = \operatorname*{argmin}_{y_h^* \in K_h^*} \left[\frac{1}{2} \mathcal{A}(y_h^*, y_h^*) - (f, y_h^*) \right], \tag{3.1}$$

where

$$K_h^* = \{ v \in H^2(\Omega) \cap H_0^1(\Omega) : \psi_1(p) \le v(p) \le \psi_2(p) \qquad \forall p \in \mathcal{V}_h \}. \tag{3.2}$$

By the standard theory (3.1) has a unique solution $\bar{y}_h^*$ characterized by the variational inequality

$$\mathcal{A}(\bar{y}_h^*, y_h^* - \bar{y}_h^*) \ge (f, y_h^* - \bar{y}_h^*) \qquad \forall y_h^* \in K_h^*. \tag{3.3}$$

Note that, on the one hand, $\bar{y}_h^* \in H^2(\Omega) \cap H_0^1(\Omega)$ minimizes the same functional as $\bar{y}$ but on the larger set $K_h^* \supset K$, and, on the other hand, $\bar{y}_h^*$ shares the same pointwise constraints as $\bar{y}_h$. Thus the intermediate obstacle problem connects the continuous obstacle problem (1.7) and the discrete obstacle problem (2.18). We will carry out the convergence analysis using (1.9), (2.20), and (3.3), but not their complementarity forms.

3.1. Relation Between $\bar{y}$ and $\bar{y}_h^*$. Using the fact that $H^2(\Omega)$ is compactly embedded in $C(\bar{\Omega})$, it was shown in [25] that there exist two nonnegative functions $\phi_1, \phi_2 \in C_0^\infty(\Omega)$ and a positive number h_0 such that for any $h \leq h_0$ we can find two positive numbers $\delta_{h,1}$ and $\delta_{h,2}$ with the following properties:

$$\hat{y}_h := \bar{y}_h^* + \delta_{h,1}\phi_1 - \delta_{h,2}\phi_2 \in K \quad \text{and} \quad \delta_{h,i} \lesssim h^2. \tag{3.4}$$

Note that we can treat $\bar{y}$ as an internal approximation of $\bar{y}_h^*$ since $K \subset K_h^*$. It then follows from (3.4) and a standard result [3] that

$$\|\bar{y}_h^* - \bar{y}\|_{H^2(\Omega)} \lesssim \left[\inf_{y\in K} \|\bar{y}_h^* - y\|_{H^2(\Omega)}\right]^{\frac{1}{2}} \lesssim \|\bar{y}_h^* - \hat{y}_h\|_{H^2(\Omega)}^{\frac{1}{2}} \lesssim h. \tag{3.5}$$

Remark 3.1. *Even though the results in [25] are obtained for clamped Kirchhoff plates on convex polygonal domains, these results are also valid for general boundary conditions and general polygonal domains because they are interior results that only require the following ingredients: (i) The set K_h^* is a closed convex subset of $H^2(\Omega)$. (ii) The constraints and the boundary conditions are separated. (iii) The obstacle functions ψ_1, ψ_2 and the solution $\bar{y}$ belong to $C^2(\Omega)$.*

3.2. Connection Between K_h and K_h^*. We can connect K_h and K_h^* by an enriching operator E_h that maps V_h into $H^2(\Omega) \cap H_0^1(\Omega)$. By construction E_h is a linear operator that preserves the nodal values at the vertices of $\mathcal{T}_h$, i.e.,

$$(E_h v)(p) = v(p) \qquad \forall p \in \mathcal{V}_h,\ v \in V_h, \tag{3.6}$$

which, in view of (2.19) and (3.2), implies

$$E_h K_h \subset K_h^*. \tag{3.7}$$

Moreover we have (cf. the notation in Sect. 2.1),

$$\begin{aligned} &\sum_{m=0}^{2} h_T^{2m} |v - E_h v|_{H^m(T)}^2 \\ &\lesssim h_T^4 \left(\sum_{T'\in\mathcal{T}_T} |v|_{H^2(T')}^2 + \sum_{e\in\mathcal{E}_{\mathcal{V}_T}^i} |e|^{-1} \|[\![\partial v/\partial n]\!]\|_{L_2(e)}^2 \right) \end{aligned} \tag{3.8}$$

for any $v \in V_h$ and $T \in \mathcal{T}_h$, and

$$\sum_{m=0}^{2} h_T^{m-2} |\zeta - E_h \Pi_h \zeta|_{H^m(T)} \lesssim h_T^s |\zeta|_{H^{2+s}(\mathcal{S}_T)} \tag{3.9}$$

for all $\zeta \in H^{2+s}(\mathcal{S}_T)$, $T \in \mathcal{T}_h$ and $s \in [0,1]$. The construction of E_h, which is similar to the constructions of the enriching operators in [16, 21] under different boundary conditions, is given in Appendix B, where we also derive the estimates (3.8) and (3.9).

The estimate (3.8) implies

$$\sum_{T\in\mathcal{T}_h} h_T^{2(m-2)}|v - E_h v|^2_{H^m(T)} \lesssim \|v\|_h^2 \qquad \forall\, v \in V_h, \tag{3.10}$$

and in particular,

$$|E_h v|_{H^2(\Omega)} \lesssim \|v\|_h \qquad \forall\, v \in V_h. \tag{3.11}$$

Combining (2.7a), (2.8a), (3.10) and the trace theorem with scaling, we also have

$$\sum_{e\in\mathcal{E}_h} |e|^{-1}\|\{\!\{\partial(v - E_h v)/\partial n_e\}\!\}\|^2_{L_2(e)} \lesssim \|v\|_h^2 \qquad \forall\, v \in V_h. \tag{3.12}$$

Finally the quasi-local estimate (3.9) implies the following result for the solution $\bar{y}$ of (1.7). We omit the proof due to its similarity with the proof of Lemma 2.1.

LEMMA 3.1. *There exists a positive constant C independent of h such that*

$$\|\bar{y} - E_h\Pi_h\bar{y}\|_{L_2(\Omega)} + h|\bar{y} - E_h\Pi_h\bar{y}|_{H^1(\Omega)} + h^2|\bar{y} - E_h\Pi_h\bar{y}|_{H^2(\Omega)} \le Ch^{2+\tau},$$

where $\tau = \alpha$ if $\mathcal{T}_h$ is quasi-uniform and $\tau = 1$ if $\mathcal{T}_h$ is graded according to (2.2)–(2.4).

4. Preliminary Estimates. In this section we derive some preliminary estimates that are useful for the convergence analysis in Sect. 5. We begin by stating the following integration by parts formula that holds for $v, w \in V_h$:

$$\begin{aligned}
&\sum_{T\in\mathcal{T}_h}\int_T D^2 v : D^2(w - E_h w)\,dx \\
&\quad= \sum_{T\in\mathcal{T}_h}\int_{\partial T}\left[\left(\frac{\partial^2 v}{\partial n^2}\right)\left(\frac{\partial(w - E_h w)}{\partial n}\right)\right. \\
&\qquad\left.+\left(\frac{\partial^2 v}{\partial n\partial t}\right)\left(\frac{\partial(w - E_h w)}{\partial t}\right)\right] ds \\
&\quad= -\sum_{e\in\mathcal{E}_h}\int_e \left[\!\left[\frac{\partial^2 v}{\partial n_e^2}\right]\!\right]\left\{\!\!\left\{\frac{\partial(w - E_h w)}{\partial n_e}\right\}\!\!\right\} ds \\
&\qquad-\sum_{e\in\mathcal{E}_h^i}\int_e \left\{\!\!\left\{\frac{\partial^2 v}{\partial n^2}\right\}\!\!\right\}\left[\!\left[\frac{\partial(w - E_h w)}{\partial n}\right]\!\right] ds.
\end{aligned} \tag{4.1}$$

Note that $v \in \mathbb{P}_2(T)$ and hence on any edge e of $T \in \mathcal{T}_h$ we have

$$\int_e \left(\frac{\partial^2 v_T}{\partial n \partial t}\right)\left(\frac{\partial(w_T - E_h w)}{\partial t}\right) ds = \left(\frac{\partial^2 v_T}{\partial n \partial t}\right)\int_e \frac{\partial(w_T - E_h w)}{\partial t}\, ds = 0$$

because of (3.6).

Next we derive a basic estimate for $\bar{y} - \bar{y}_h$, where $\bar{y}$ (resp. $\bar{y}_h$) is the solution of (1.7)/(1.9) [resp. (2.18)/(2.20)].

LEMMA 4.1. *There exists a positive constant C independent of h such that*

$$\|\bar{y} - \bar{y}_h\|_h^2 \le 2\|\bar{y} - \Pi_h\bar{y}\|_h^2 + C\big[\mathcal{A}_h(\Pi_h\bar{y}, \Pi_h\bar{y} - \bar{y}_h) - (f, \Pi_h\bar{y} - \bar{y}_h)\big]. \quad (4.2)$$

Proof. Since $\Pi_h\bar{y} \in K_h$, we deduce from (2.15) and (2.20) that

$$\begin{aligned}\|\bar{y} - \bar{y}_h\|_h^2 &\le 2\|\bar{y} - \Pi_h\bar{y}\|_h^2 + 2\|\Pi_h\bar{y} - \bar{y}_h\|_h^2\\ &\le 2\|\bar{y} - \Pi_h\bar{y}\|_h^2 + C\mathcal{A}_h(\Pi_h\bar{y} - \bar{y}_h, \Pi_h\bar{y} - \bar{y}_h)\\ &\le 2\|\bar{y} - \Pi_h\bar{y}\|_h^2 + C\big[\mathcal{A}_h(\Pi_h\bar{y}, \Pi_h\bar{y} - \bar{y}_h) - (f, \Pi_h\bar{y} - \bar{y}_h)\big].\end{aligned}$$

□

In view of Lemmas 2.1 and 4.1, we can complete the error analysis by bounding the second term on the right-hand side of (4.2). This will be carried out in Sect. 5 after we have developed several technical lemmas in the remaining part of this section.

LEMMA 4.2. *There exists a positive constant C independent of h such that*

$$\sum_{e\in\mathcal{E}_h} |e|\big\|[\![\partial^2(\Pi_h\bar{y})/\partial n_e^2]\!]\big\|_{L_2(e)}^2 \le Ch^{2\tau},$$

where $\tau = \alpha$ if $\mathcal{T}_h$ is quasi-uniform and $\tau = 1$ if $\mathcal{T}_h$ is graded according to (2.2)–(2.4).

Proof. We will split the estimate into two cases. Let $\mathcal{E}_h^R = \{e \in \mathcal{E}_h :$ e is not an edge of any triangle that touches a corner of Ω where the angle is strictly greater than $\pi/2\}$ and $\mathcal{E}_h^S = \mathcal{E}_h \setminus \mathcal{E}_h^R$. Note that the number of edges in $\mathcal{E}_h^S$ is bounded by a constant determined by the minimum angle of $\mathcal{T}_h$.

Since away from the corners of Ω where the angles are strictly greater than $\pi/2$ the function $\bar{y}$ belongs to H^3 and $\partial^2\bar{y}/\partial n^2 = \Delta\bar{y}$ vanishes on $\partial\Omega$ (cf. Appendix A), we have, by (2.7b), (2.21) and the trace theorem with scaling,

$$\begin{aligned}\sum_{e\in\mathcal{E}_h^R} |e|\big\|[\![\partial^2(\Pi_h\bar{y})/\partial n_e^2]\!]\big\|_{L_2(e)}^2 &= \sum_{e\in\mathcal{E}_h^R} |e|\big\|[\![\partial^2(\Pi_h\bar{y} - \bar{y})/\partial n_e^2]\!]\big\|_{L_2(e)}^2\\ &\lesssim \sum_{e\in\mathcal{E}_h^R}\sum_{T\in\mathcal{T}_e} (h_T^2\Phi^{-2}(c_T))\Phi^2(c_T)|\bar{y}|_{H^3(T)}^2,\end{aligned}$$

where the function Φ is defined in (2.3). It then follows from (2.1)–(2.3) and (A.7) that

$$\sum_{e\in\mathcal{E}_h^R} |e| \left\| [\![\partial^2(\Pi_h\bar{y})/\partial n_e^2]\!] \right\|_{L_2(e)}^2 \lesssim h^{2\tau}, \tag{4.3}$$

where $\tau = \alpha$ if $\mathcal{T}_h$ is quasi-uniform and $\tau = 1$ if $\mathcal{T}_h$ satisfies (2.2)–(2.4).

Let $e \in \mathcal{E}_h^S$ be an edge of a triangle that touches a corner p_ℓ of Ω where the angle $\omega_\ell \in (\pi/2, \pi)$. It follows from scaling that

$$\begin{aligned} |e| \left\| [\![\partial^2(\Pi_h\bar{y})/\partial n_e^2]\!] \right\|_{L_2(e)}^2 &\lesssim \sum_{T\in\mathcal{T}_e} |\Pi_h\bar{y}|_{H^2(T)}^2 \\ &\lesssim \sum_{T\in\mathcal{T}_e} \left(|\Pi_h\bar{y} - \bar{y}|_{H^2(T)}^2 + |\bar{y}|_{H^2(T)}^2 \right). \end{aligned}$$

Let $T \in \mathcal{T}_e$. Since $\bar{y} \in H^{2+\alpha_\ell}(T)$ (cf. Appendix A), we have

$$|\Pi_h\bar{y} - \bar{y}|_{H^2(T)} \lesssim |\bar{y}|_{H^{2+\alpha_\ell}(T)} h_T^{\alpha_\ell}.$$

Moreover we have

$$\begin{aligned} |\bar{y}|_{H^2(T)} &\approx \sum_{|\mu|=2} \|\partial^\mu \bar{y}\|_{L_2(T)} \\ &= \sum_{|\mu|=2} \|\Psi^{-1}\big(\Psi(\partial^\mu \bar{y})\big)\|_{L_2(T)} \lesssim h_T^{\alpha_\ell} \sum_{|\mu|=2} \|\Psi(\partial^\mu \bar{y})\|_{L_2(T)}, \end{aligned}$$

where Ψ is define in (A.9). Therefore it follows from (2.1), (2.5), Remark 2.1 and (A.8) that

$$\sum_{e\in\mathcal{E}_h^S} |e| \left\| [\![\partial^2(\Pi_h\bar{y})/\partial n_e^2]\!] \right\|_{L_2(e)}^2 \lesssim h^{2\tau}, \tag{4.4}$$

where $\tau = \alpha$ if $\mathcal{T}_h$ is quasi-uniform and $\tau = 1$ if $\mathcal{T}_h$ satisfies (2.2)–(2.4).

The lemma follows from (4.3) and (4.4). □

LEMMA 4.3. *There exists a positive constant C independent of h such that*

$$\begin{aligned} &\left| a_h(\Pi_h\bar{y}, \Pi_h\bar{y} - \bar{y}_h) - \int_\Omega D^2\bar{y} : D^2 E_h(\Pi_h\bar{y} - \bar{y}_h)\, dx \right| \\ &\quad \le Ch^\tau \|\Pi_h\bar{y} - \bar{y}_h\|_h, \end{aligned} \tag{4.5}$$

where $\tau = \alpha$ if $\mathcal{T}_h$ is quasi-uniform and $\tau = 1$ if $\mathcal{T}_h$ is graded according to (2.2)–(2.4).

Proof. Since both $[\![\partial E_h(\Pi_h\bar{y} - \bar{y}_h)/\partial n]\!]$ and $[\![\partial\bar{y}/\partial n]\!]$ equal 0, we have, from (2.9),

$$
\begin{aligned}
&a_h(\Pi_h\bar{y},\Pi_h\bar{y}-\bar{y}_h)\\
&\quad=\sum_{T\in\mathcal{T}_h}\int_T D^2\bar{y}:D^2E_h(\Pi_h\bar{y}-\bar{y}_h)\,dx\\
&\qquad+\sum_{T\in\mathcal{T}_h}\int_T D^2(\Pi_h\bar{y}-\bar{y}):D^2E_h(\Pi_h\bar{y}-\bar{y}_h)\,dx\\
&\qquad+\sum_{T\in\mathcal{T}_h}\int_T D^2(\Pi_h\bar{y}):D^2\big[(\Pi_h\bar{y}-\bar{y}_h)-E_h(\Pi_h\bar{y}-\bar{y}_h)\big]\,dx\\
&\qquad+\sum_{e\in\mathcal{E}_h^i}\int_e\left\{\!\!\left\{\frac{\partial^2(\Pi_h\bar{y})}{\partial n^2}\right\}\!\!\right\}\left[\!\!\left[\frac{\partial[(\Pi_h\bar{y}-\bar{y}_h)-E_h(\Pi_h\bar{y}-\bar{y}_h)]}{\partial n}\right]\!\!\right]ds\qquad(4.6)\\
&\qquad+\sum_{e\in\mathcal{E}_h^i}\int_e\left\{\!\!\left\{\frac{\partial^2(\Pi_h\bar{y}-\bar{y}_h)}{\partial n^2}\right\}\!\!\right\}\left[\!\!\left[\frac{\partial(\Pi_h\bar{y}-\bar{y})}{\partial n}\right]\!\!\right]ds\\
&\qquad+\sum_{e\in\mathcal{E}_h^i}\frac{\sigma}{|e|}\int_e\left[\!\!\left[\frac{\partial(\Pi_h\bar{y}-\bar{y})}{\partial n}\right]\!\!\right]\left[\!\!\left[\frac{\partial(\Pi_h\bar{y}-\bar{y}_h)}{\partial n}\right]\!\!\right]ds,
\end{aligned}
$$

and we can use (2.6), (2.13), Lemma 2.1, (3.11) and scaling to estimate the second, fifth, and sixth terms on the right-hand side of (4.6) as follows:

$$
\begin{aligned}
&\left|\sum_{T\in\mathcal{T}_h}\int_T D^2(\Pi_h\bar{y}-\bar{y}):D^2E_h(\Pi_h\bar{y}-\bar{y}_h)\,dx\right|\\
&\qquad\le\left(\sum_{T\in\mathcal{T}_h}|\Pi_h\bar{y}-\bar{y}|^2_{H^2(T)}\right)^{\frac12}|E_h(\Pi_h\bar{y}-\bar{y}_h)|_{H^2(\Omega)}\qquad(4.7)\\
&\qquad\lesssim h^\tau\|\Pi_h\bar{y}-\bar{y}_h\|_h,
\end{aligned}
$$

$$
\begin{aligned}
&\left|\sum_{e\in\mathcal{E}_h^i}\int_e\left\{\!\!\left\{\frac{\partial^2(\Pi_h\bar{y}-\bar{y}_h)}{\partial n^2}\right\}\!\!\right\}\left[\!\!\left[\frac{\partial(\Pi_h\bar{y}-\bar{y})}{\partial n}\right]\!\!\right]ds\right|\\
&\quad\le\left(\sum_{e\in\mathcal{E}_h^i}|e|\|\{\!\{\partial^2(\Pi_h\bar{y}-\bar{y}_h)/\partial n^2\}\!\}\|^2_{L_2(e)}\right)^{\frac12}\\
&\qquad\times\left(\sum_{e\in\mathcal{E}_h^i}\frac{1}{|e|}\|[\![\partial(\Pi_h\bar{y}-\bar{y})/\partial n]\!]\|^2_{L_2(e)}\right)^{\frac12}\qquad(4.8)\\
&\quad\lesssim\left(\sum_{T\in\mathcal{T}_h}|\Pi_h\bar{y}-\bar{y}_h|^2_{H^2(T)}\right)^{\frac12}\left(\sum_{e\in\mathcal{E}_h^i}\frac{1}{|e|}\|[\![\partial(\Pi_h\bar{y}-\bar{y})/\partial n]\!]\|^2_{L_2(e)}\right)^{\frac12}\\
&\quad\lesssim h^\tau\|\Pi_h\bar{y}-\bar{y}_h\|_h,
\end{aligned}
$$

$$\begin{aligned}
&\left|\sum_{e\in\mathcal{E}_h^i}\frac{\sigma}{|e|}\int_e\left[\!\!\left[\frac{\partial(\Pi_h\bar{y}-\bar{y})}{\partial n}\right]\!\!\right]\left[\!\!\left[\frac{\partial(\Pi_h\bar{y}-\bar{y}_h)}{\partial n}\right]\!\!\right]ds\right| \\
&\quad\lesssim\left(\sum_{e\in\mathcal{E}_h^i}\frac{1}{|e|}\|[\![\partial(\Pi_h\bar{y}-\bar{y})/\partial n]\!]\|^2_{L_2(e)}\right)^{\frac{1}{2}} \\
&\qquad\times\left(\sum_{e\in\mathcal{E}_h^i}\frac{1}{|e|}\|[\![\partial(\Pi_h\bar{y}-\bar{y}_h)/\partial n]\!]\|^2_{L_2(e)}\right)^{\frac{1}{2}} \\
&\quad\lesssim h^\tau\|\Pi_h\bar{y}-\bar{y}_h\|_h.
\end{aligned}\tag{4.9}$$

Now we use (3.12), the integration by parts formula (4.1) together with Lemma 4.2 to estimate the sum of the third and fourth terms on the right-hand side of (4.6) by

$$\begin{aligned}
&\sum_{T\in\mathcal{T}_h}\int_T D^2(\Pi_h\bar{y}):D^2\left[(\Pi_h\bar{y}-\bar{y}_h)-E_h(\Pi_h\bar{y}-\bar{y}_h)\right]dx \\
&\qquad+\sum_{e\in\mathcal{E}_h^i}\int_e\left\{\!\!\left\{\frac{\partial^2(\Pi_h\bar{y})}{\partial n^2}\right\}\!\!\right\}\left[\!\!\left[\frac{\partial[(\Pi_h\bar{y}-\bar{y}_h)-E_h(\Pi_h\bar{y}-\bar{y}_h)]}{\partial n}\right]\!\!\right]ds \\
&\quad=-\sum_{e\in\mathcal{E}_h}\int_e\left[\!\!\left[\frac{\partial^2(\Pi_h\bar{y})}{\partial n_e^2}\right]\!\!\right]\left\{\!\!\left\{\frac{\partial[(\Pi_h\bar{y}-\bar{y}_h)-E_h(\Pi_h\bar{y}-\bar{y}_h)]}{\partial n_e}\right\}\!\!\right\}ds \\
&\quad\le\left(\sum_{e\in\mathcal{E}_h}|e|\left\|\left[\!\!\left[\partial^2(\Pi_h\bar{y})/\partial n_e^2\right]\!\!\right]\right\|^2_{L_2(e)}\right)^{\frac{1}{2}} \\
&\qquad\times\left(\sum_{e\in\mathcal{E}_h}\frac{1}{|e|}\left\|\left\{\!\!\left\{\partial[(\Pi_h\bar{y}-\bar{y}_h)-E_h(\Pi_h\bar{y}-\bar{y}_h)]/\partial n_e\right\}\!\!\right\}\right\|^2_{L_2(e)}\right)^{\frac{1}{2}} \\
&\quad\lesssim h^\tau\|\Pi_h\bar{y}-\bar{y}_h\|_h.
\end{aligned}\tag{4.10}$$

The lemma follows from (4.6)–(4.10). □

LEMMA 4.4. *There exists a positive constant C independent of h such that*

$$\mathcal{A}(\bar{y},E_h\Pi_h\bar{y}-\bar{y})\le Ch^{1+\tau},$$

where $\tau=\alpha$ if $\mathcal{T}_h$ is quasi-uniform and $\tau=1$ if $\mathcal{T}_h$ is graded according to (2.2)–(2.4).

Proof. Since $\Delta\bar{y}\in H_0^1(\Omega)$ (cf. Appendix A), we have, by (1.5) and Lemma 3.1,

$$\begin{aligned}\int_\Omega D^2\bar{y} : D^2(E_h\Pi_h\bar{y}-\bar{y})\,dx &= \int_\Omega (\Delta\bar{y})\big(\Delta(E_h\Pi_h\bar{y}-\bar{y})\big)\,dx\\ &= -\int_\Omega \nabla(\Delta\bar{y})\cdot\nabla(E_h\Pi_h\bar{y}-\bar{y})\,dx \qquad (4.11)\\ &\lesssim |E_h\Pi_h\bar{y}-\bar{y}|_{H^1(\Omega)} \lesssim h^{1+\tau}.\end{aligned}$$

Moreover Lemma 3.1 also implies

$$(\bar{y}, E_h\Pi_h\bar{y}-\bar{y}) \lesssim h^{2+\tau}. \qquad (4.12)$$

The lemma follows from (1.8), (4.11), and (4.12). □

5. Convergence Analysis. In this section we complete the error analysis by finding a bound for the second term on the right-hand side of (4.2). We will show that

$$\mathcal{A}_h(\Pi_h\bar{y},\Pi_h\bar{y}-\bar{y}_h)-(f,\Pi_h\bar{y}-\bar{y}_h) \lesssim h^{2\tau}+h^\tau\|\Pi_h\bar{y}-\bar{y}_h\|_h, \qquad (5.1)$$

where $\tau=\alpha$ if $\mathcal{T}_h$ is quasi-uniform and $\tau=1$ if $\mathcal{T}_h$ satisfies (2.2)–(2.4). But first we use (5.1) to establish the main result of this paper.

THEOREM 5.1. *There exists a positive constant C independent of h such that*

$$\|\bar{y}-\bar{y}_h\|_h \le Ch^\tau, \qquad (5.2)$$

where $\tau=\alpha$ if $\mathcal{T}_h$ is quasi-uniform and $\tau=1$ if $\mathcal{T}_h$ is graded according to (2.2)–(2.4).

Proof. It follows from Lemma 2.1, (4.2), (5.1), and the inequality of arithmetic and geometric means that

$$\begin{aligned}\|\bar{y}-\bar{y}_h\|_h^2 &\le C\big(h^{2\tau}+h^\tau\|\Pi_h\bar{y}-\bar{y}_h\|_h\big)\\ &\le C\big(h^{2\tau}+h^\tau\|\bar{y}-\bar{y}_h\|_h\big) \le Ch^{2\tau}+\frac{1}{2}\|\bar{y}-\bar{y}_h\|_h^2,\end{aligned}$$

which implies (5.2). □

The following lemma reduces the derivation of (5.1) to an estimate at the continuous level.

LEMMA 5.1. *There exists a positive constant C independent of h such that*

$$\begin{aligned}&\mathcal{A}_h(\Pi_h\bar{y},\Pi_h\bar{y}-\bar{y}_h)-(f,\Pi_h\bar{y}-\bar{y}_h)\\ &\quad\le Ch^\tau\|\Pi_h\bar{y}-\bar{y}_h\|_h+\mathcal{A}\big(\bar{y},E_h(\Pi_h\bar{y}-\bar{y}_h)\big)-\big(f,E_h(\Pi_h\bar{y}-\bar{y}_h)\big),\end{aligned}$$

where $\tau=\alpha$ if $\mathcal{T}_h$ is quasi-uniform and $\tau=1$ if $\mathcal{T}_h$ is graded according to (2.2)–(2.4).

Proof. From (1.8) and (2.12) we have

$$\begin{aligned}
&\mathcal{A}_h(\Pi_h\bar{y},\Pi_h\bar{y}-\bar{y}_h)-(f,\Pi_h\bar{y}-\bar{y}_h)\\
&\quad=\beta\left[a_h(\Pi_h\bar{y},\Pi_h\bar{y}-\bar{y}_h)-\int_\Omega\big(D^2\bar{y}:D^2E_h(\Pi_h\bar{y}-\bar{y}_h)\big)\,dx\right]\\
&\qquad+\gamma\big[(\Pi_h\bar{y},\Pi_h\bar{y}-\bar{y}_h)-\big(\bar{y},E_h(\Pi_h\bar{y}-\bar{y}_h)\big)\big]\\
&\qquad-\big(f,(\Pi_h\bar{y}-\bar{y}_h)-E_h(\Pi_h\bar{y}-\bar{y}_h)\big)\\
&\qquad+\mathcal{A}\big(\bar{y},E_h(\Pi_h\bar{y}-\bar{y}_h)\big)-\big(f,E_h(\Pi_h\bar{y}-\bar{y}_h)\big),
\end{aligned}\tag{5.3}$$

and we can bound the second and third terms on the right-hand side of (5.3) as follows:

$$\begin{aligned}
&(\Pi_h\bar{y},\Pi_h\bar{y}-\bar{y}_h)-\big(\bar{y},E_h(\Pi_h\bar{y}-\bar{y}_h)\big)\\
&\quad=(\Pi_h\bar{y}-\bar{y},\Pi_h\bar{y}-\bar{y}_h)+\big(\bar{y},(\Pi_h\bar{y}-\bar{y}_h)-E_h(\Pi_h\bar{y}-\bar{y}_h)\big)\\
&\quad\lesssim h^2\|\Pi_h\bar{y}-\bar{y}_h\|_h
\end{aligned}\tag{5.4}$$

by (2.21) and (3.10); and

$$\big|\big(f,(\Pi_h\bar{y}-\bar{y}_h)-E_h(\Pi_h\bar{y}-\bar{y}_h)\big)\big|\lesssim h^2\|\Pi_h\bar{y}-\bar{y}_h\|_h\tag{5.5}$$

by (3.10).

The lemma follows from Lemma 4.3 and (5.3)–(5.5). □

In view of Lemma 5.1, it only remains to show that

$$\mathcal{A}\big(\bar{y},E_h(\Pi_h\bar{y}-\bar{y}_h)\big)-\big(f,E_h(\Pi_h\bar{y}-\bar{y}_h)\big)\lesssim h^{2\tau}+h^\tau\|\Pi_h\bar{y}-\bar{y}_h\|_h.\tag{5.6}$$

We will use the relation $A\mathrel{\dot{\leq}}B$ to streamline the derivation of (5.6), where

$$A\mathrel{\dot{\leq}}B\quad\text{means that}\quad A-B\lesssim h^{2\tau}+h^\tau\|\Pi_h\bar{y}-\bar{y}_h\|_h.$$

The estimate (5.6) can then be written as

$$\mathcal{A}\big(\bar{y},E_h(\Pi_h\bar{y}-\bar{y}_h)\big)\mathrel{\dot{\leq}}\big(f,E_h(\Pi_h\bar{y}-\bar{y}_h)\big).\tag{5.7}$$

It follows from (1.9), (3.3), (3.4), (3.7), and Lemma 4.4 that

$$\begin{aligned}
\mathcal{A}\big(\bar{y},E_h(\Pi_h\bar{y}-\bar{y}_h)\big)&=\mathcal{A}(\bar{y},E_h\Pi_h\bar{y}-\bar{y})+\mathcal{A}(\bar{y},\bar{y}-E_h\bar{y}_h)\\
&\mathrel{\dot{\leq}}\mathcal{A}(\bar{y},\bar{y}-E_h\bar{y}_h)\\
&=\mathcal{A}(\bar{y},\bar{y}-\hat{y}_h)+\mathcal{A}(\bar{y},\hat{y}_h-E_h\bar{y}_h)\\
&\leq(f,\bar{y}-\hat{y}_h)+\mathcal{A}(\bar{y},\bar{y}_h^*-E_h\bar{y}_h)+\mathcal{A}(\bar{y},\delta_{h,1}\phi_1-\delta_{h,2}\phi_2)\\
&\mathrel{\dot{\leq}}(f,\bar{y}-\hat{y}_h)+\mathcal{A}(\bar{y},\bar{y}_h^*-E_h\bar{y}_h)\\
&=(f,\bar{y}-\hat{y}_h)+\mathcal{A}(\bar{y}_h^*,\bar{y}_h^*-E_h\bar{y}_h)+\mathcal{A}(\bar{y}-\bar{y}_h^*,\bar{y}_h^*-E_h\bar{y}_h)\\
&\leq(f,\bar{y}-\hat{y}_h)+(f,\bar{y}_h^*-E_h\bar{y}_h)+\mathcal{A}(\bar{y}-\bar{y}_h^*,\bar{y}_h^*-E_h\bar{y}_h)\\
&\mathrel{\dot{\leq}}(f,\bar{y}-E_h\bar{y}_h)+\mathcal{A}(\bar{y}-\bar{y}_h^*,\bar{y}_h^*-E_h\bar{y}_h).
\end{aligned}\tag{5.8}$$

Moreover we have

$$\begin{aligned}(f,\bar{y}-E_h\bar{y}_h) &= \big(f,E_h(\Pi_h\bar{y}-\bar{y}_h)\big)+(f,\bar{y}-E_h\Pi_h\bar{y}) \\ &\overset{\cdot}{\leq} \big(f,E_h(\Pi_h\bar{y}-\bar{y}_h)\big)\end{aligned} \tag{5.9}$$

by Lemma 3.1, and

$$\begin{aligned}&\mathcal{A}(\bar{y}-\bar{y}_h^*,\bar{y}_h^*-E_h\bar{y}_h) \\ &\quad=\mathcal{A}(\bar{y}-\bar{y}_h^*,\bar{y}_h^*-\bar{y})+\mathcal{A}(\bar{y}-\bar{y}_h^*,\bar{y}-E_h\bar{y}_h) \\ &\quad\leq \mathcal{A}(\bar{y}-\bar{y}_h^*,\bar{y}-E_h\Pi_h\bar{y})+\mathcal{A}\big(\bar{y}-\bar{y}_h^*,E_h(\Pi_h\bar{y}-\bar{y}_h)\big) \\ &\quad\overset{\cdot}{\leq} 0\end{aligned} \tag{5.10}$$

by (3.5), (3.11), and Lemma 3.1. The relation (5.7) then follows from (5.8)–(5.10). Therefore we have established (5.6) and hence (5.1).

The following corollary is an immediate consequence of (2.17) and Theorem 5.1.

COROLLARY 5.1. *There exists a positive constant C independent of h such that*

$$|\bar{y}-\bar{y}_h|_{H^1(\Omega)}\leq Ch^{\tau},$$

where $\tau=\alpha$ if $\mathcal{T}_h$ is quasi-uniform and $\tau=1$ if $\mathcal{T}_h$ is graded according to (2.2)–(2.4).

Since the energy norm $\|\cdot\|_h$ is an H^2-like norm, we can also deduce an L_∞ norm error estimate from Theorem 5.1. The proof of the following theorem, which is based on Lemmas 2.1, 3.1, Theorem 5.1, standard inverse estimates and the Sobolev inequality, is identical to the proof of Theorem 4.1 in [23] and thus omitted.

THEOREM 5.2. *There exists a positive constant C independent of h such that*

$$\|\bar{y}-\bar{y}_h\|_{L_\infty(\Omega)}\leq Ch^{\tau},$$

where $\tau=\alpha$ if $\mathcal{T}_h$ is quasi-uniform and $\tau=1$ if $\mathcal{T}_h$ is graded according to (2.2)–(2.4).

REMARK 5.1. *Since the norms $\|\cdot\|_{L_\infty(\Omega)}$ and $|\cdot|_{H^1(\Omega)}$ are weaker than the energy norm $\|\cdot\|_h$, the order of convergence in these norms should be higher than the order of convergence in $\|\cdot\|_h$. This is confirmed by the numerical results in Sect. 6. Therefore the estimates for $\|\bar{y}-\bar{y}_h\|_{L_\infty(\Omega)}$ and $|\bar{y}-\bar{y}_h|_{H^1(\Omega)}$ in Corollary 5.1 and Theorem 5.2 are not sharp.*

For the optimal control problem defined by (1.1)–(1.3), we can take the approximation for the optimal control $\bar{u}$ to be the function $\bar{u}_h\in V_h$ defined by

$$\int_\Omega \nabla\bar{y}_h\cdot\nabla v\,dx=\int_\Omega \bar{u}_h v\,dx \qquad \forall\, v\in V_h. \tag{5.11}$$

THEOREM 5.3. *There exists a positive constant C independent of h such that*

$$\|\bar{u} - \bar{u}_h\|_{L_2(\Omega)} \leq Ch^{\tau}, \tag{5.12}$$

where $\tau = \alpha$ if $\mathcal{T}_h$ is quasi-uniform and $\tau = 1$ if $\mathcal{T}_h$ is graded according to (2.2)–(2.4).

Proof. Let $Q_h : L_2(\Omega) \longrightarrow V_h$ be the orthogonal projection. From (1.2) we have

$$\int_\Omega \nabla\bar{y} \cdot \nabla v \, dx = \int_\Omega \bar{u} v \, dx = \int_\Omega (Q_h\bar{u}) v \, dx \qquad \forall\, v \in V_h. \tag{5.13}$$

Let $v \in V_h$ be arbitrary. Using integration by parts, the Cauchy–Schwarz inequality, scaling, (2.7b), (2.13), and Theorem 5.1, we find

$$\begin{aligned}
&\int_\Omega \nabla(\bar{y} - \bar{y}_h) \cdot \nabla v \, dx \\
&\quad = -\sum_{e \in \mathcal{E}_h^i} \int_e [\![\partial(\bar{y} - \bar{y}_h)/\partial n]\!] v \, ds - \sum_{T \in \mathcal{T}_h} \int_T [\Delta(\bar{y} - \bar{y}_h)] v \, dx \\
&\quad \leq \left(\sum_{e \in \mathcal{E}_h^i} |e|^{-1} \|[\![\partial(\bar{y} - \bar{y}_h)/\partial n]\!]\|^2_{L_2(e)} \right)^{\frac{1}{2}} \left(\sum_{e \in \mathcal{E}_h^i} |e| \|v\|^2_{L_2(e)} \right)^{\frac{1}{2}} \\
&\qquad + \left(\sum_{T \in \mathcal{T}_h} |\bar{y} - \bar{y}_h|^2_{H^2(T)} \right)^{\frac{1}{2}} \|v\|_{L_2(\Omega)} \\
&\quad \lesssim \|\bar{y} - \bar{y}_h\|_h \|v\|_{L_2(\Omega)} \lesssim h^{\tau} \|v\|_{L_2(\Omega)}.
\end{aligned}$$

It then follows from (5.11), (5.13) and duality that

$$\begin{aligned}
&\|Q_h\bar{u} - \bar{u}_h\|_{L_2(\Omega)} \\
&\quad = \sup_{v \in V_h \setminus \{0\}} \left(\int_\Omega (Q_h\bar{u} - \bar{u}_h) v \, dx \right) / \|v\|_{L_2(\Omega)} \\
&\quad = \sup_{v \in V_h \setminus \{0\}} \left(\int_\Omega \nabla(\bar{y} - \bar{y}_h) \cdot \nabla v \, dx \right) / \|v\|_{L_2(\Omega)} \lesssim h^{\tau}.
\end{aligned} \tag{5.14}$$

Furthermore, we have, by a standard interpolation error estimate [61],

$$\|Q_h\bar{u} - \bar{u}\|_{L_2(\Omega)} \lesssim |\bar{u}|_{H^1(\Omega)} h. \tag{5.15}$$

The estimate (5.12) follows from (5.14) and (5.15). ∎

REMARK 5.2. *One can also take the piecewise constant function* $\bar{u}_h = -\Delta_h \bar{y}_h$ *to be an approximation of the optimal control* $\bar{u}$*, where* Δ_h *is the piecewise Laplacian with respect to* $\mathcal{T}_h$*. The estimate* (5.12) *then immediately follows from* (2.13) *and Theorem 5.1. But numerical results indicate that the approximation of* $\bar{u}$ *defined by* (5.11) *is a better choice.*

REMARK 5.3. *By tracing the constants in all the estimates* (*including* (3.4)) *one can show* (*using* (A.7), (A.8), *and* (A.10)) *that the constant* C *in Theorem 5.1, Corollary 5.1, Theorem 5.2, and Theorem 5.3 is of the form*

$$\mathfrak{C}\left(\|f\|_{L_2(\Omega)} + \sum_{i=1}^{2} \|\psi_i\|_{W^2_\infty(K)} + |\Delta\bar{y}|_{H^1(\Omega)} + \sum_{|\mu|=3} \|\Phi(\partial^\mu \bar{y})\|_{L_2(\Omega)} + \sum_{|\mu|=2} \|\Psi(\partial^\mu \bar{y})\|_{L_2(\Omega)} + \|\bar{y}\|_{W^2_\infty(K)}\right),$$

where Φ (*resp.* Ψ) *is defined in* (2.3) (*resp.* (A.9)), $K \subset\subset \Omega$ *is a compact neighborhood of the contact set where* $(\bar{y}-\psi_1)(\bar{y}-\psi_2) = 0$*, and the positive constant* $\mathfrak{C}$ *depends only on* Ω *and the shape regularity of* $\mathcal{T}_h$*.*

6. Numerical Results. In this section we present several numerical examples for the obstacle problem (1.7) with $\psi_1(x) = -\infty$. The computational domain for the first four examples is the square $(-0.5, 0.5) \times (-0.5, 0.5)$. The discrete problems are defined on uniform triangulations $\mathcal{T}_j$ with mesh parameter $h_j = 2^{-j}$ (= the length of the horizontal and vertical edges) for $1 \le j \le 8$, and the penalty parameter σ is chosen to be 5 which ensures the coercivity of the discrete bilinear form on uniform meshes. The solutions of the discrete problems are denoted by $\bar{y}_j$ $(1 \le j \le 8)$, which are obtained by a primal–dual active set algorithm [4, 47].

Example 1. In this example we validate our numerical scheme by solving (1.7)/(1.9) with a known solution. We begin with the obstacle problem on the disc $\{x : |x| < 2\}$ with $\gamma = 0$, $\beta = 1$, $f = 0$ and $\psi_2(x) = \frac{1}{2}|x|^2 - 1$. This problem can be solved analytically because of rotational symmetry and the exact solution is given by

$$y_\dagger(x) = \begin{cases} C_1|x|^2 \ln|x| + C_2|x|^2 + C_3 \ln|x| + C_4 & |x| > r_0 \\ \frac{1}{2}|x|^2 - 1 & |x| \leqslant r_0 \end{cases}, \tag{6.1}$$

where $r_0 = 0.31078820\ldots$, $C_1 = -0.26855864\ldots$, $C_2 = 0.45470930\ldots$, $C_3 = -0.02593989\ldots$, and $C_4 = -1.05625438\ldots$.

Let $\bar{y}$ be the restriction of $y_\dagger$ to $\Omega = (-0.5, 0.5)^2$. Then we have

$$\bar{y} = \operatorname*{argmin}_{y \in \tilde{K}} \left[\frac{1}{2}\int_\Omega (D^2 y : D^2 y)dx - \int_{\partial\Omega} \left(\frac{\partial^2 y_\dagger}{\partial n^2}\right)\left(\frac{\partial y}{\partial n}\right) ds\right], \tag{6.2}$$

where n is the unit outer normal on $\partial\Omega$ and

$$\tilde{K} = \{v \in H^2(\Omega) : v - y_\dagger \in H^1_0(\Omega) \quad \text{and} \quad v \le \psi_2 \quad \text{in } \Omega\},$$

i.e., $\bar{y}$ is the solution of an obstacle problem for a simply supported plate with nonhomogeneous boundary conditions.

As in the case of clamped plates [23], our results for simply supported plates with homogeneous boundary conditions (Theorems 5.1 and 5.2) can be extended to the nonhomogeneous case. Let $\tilde{V}_h$ be the $\mathbb{P}_2$ Lagrange finite element space associated with the triangulation $\mathcal{T}_h$. The discrete problem for (6.2) is to find

$$\bar{y}_h = \operatorname*{argmin}_{y_h \in \tilde{K}_h} \left[\frac{1}{2} a_h(y_h, y_h) - \sum_{e \in \mathcal{E}_h^b} \int_e \left(\frac{\partial^2 y_\dagger}{\partial n^2} \right) \left(\frac{\partial y_h}{\partial n} \right) ds \right], \tag{6.3}$$

where

$$\tilde{K}_h = \{ v \in \tilde{V}_h : \ v - \Pi_h y_\dagger \in H_0^1(\Omega) \quad \text{and} \quad v(p) \le \psi_2(p) \quad \forall\, p \in \mathcal{V}_h \}.$$

Let $\bar{y}_j$ be the solution of (6.3) for the jth level triangulation and Π_j be the Lagrange nodal interpolation operator for the jth level finite element space V_j. We evaluate the error $e_j = \Pi_j \bar{y} - \bar{y}_j$ in the energy norm $\| \cdot \|_{h_j}$ and in the ℓ_∞ norm $\| \cdot \|_\infty$ defined by

$$\|e_j\|_\infty = \max_{p \in \mathcal{N}_j} |e_j(p)|,$$

where $\mathcal{N}_j$ is the set of the vertices and midpoints of $\mathcal{T}_j$. We also compute the order of convergence in these norms by the formulas

$$\ln(\|e_{j-1}\|_{h_{j-1}} / \|e_j\|_{h_j}) / \ln 2 \quad \text{and} \quad \ln(\|e_{j-1}\|_\infty / \|e_j\|_\infty) / \ln 2.$$

The numerical results are presented in Table 1. The order of convergence in the energy norm is observed to be 1.5, which is better than the order of 1 predicted by Theorem 5.1. This is likely due to the fact that $\bar{y}$ is actually a C^∞ function on Ω away from the circle with radius r_0 and therefore superconvergence occurs since we use uniform triangulations. We also observe that the order of convergence in the ℓ_∞ norm is close to 2, better than the order of 1 predicted by Theorem 5.2.

We plot the discrete coincidence sets I_7 and I_8 in Fig. 1, where

$$I_j = \{ p \in \mathcal{N}_j : \ \bar{y}_j(p) \ge \psi_2(p) - \|e_j\|_\infty \}.$$

The black circle represents the exact free boundary $F = \{x \in \Omega : |x| = r_0\}$ (cf. (6.1)). It is evident that the discrete coincidence sets (resp. free boundaries) are converging to the exact coincidence set (resp. free boundary).

The second set of examples are optimal control problems with state constraints that come from [6, 55]. The value of γ is taken to be 1. Since the exact solutions are not known, we take $\tilde{e}_{\bar{y},j} = \bar{y}_{j-1} - \bar{y}_j$ and evaluate $\|\tilde{e}_{\bar{y},j}\|_{h_j}$ (the error of the state in the energy norm), $|\tilde{e}_{\bar{y},j}|_{H^1}$ (the error of

TABLE 1
Energy and l_∞ errors for Example 1

j	$\|e_j\|_{h_j}/\|\bar{y}_8\|_{h_8}$	Order	$\|e_j\|_\infty$	Order
1	2.1840×10^{-1}		7.0940×10^{-3}	
2	6.8348×10^{-2}	1.68	5.9691×10^{-4}	3.57
3	3.1394×10^{-2}	1.12	5.7224×10^{-4}	0.06
4	9.8571×10^{-3}	1.67	1.1579×10^{-4}	2.31
5	3.8462×10^{-3}	1.36	3.5461×10^{-5}	1.71
6	1.4533×10^{-3}	1.40	1.0669×10^{-5}	1.73
7	5.4157×10^{-4}	1.42	3.3085×10^{-6}	1.69
8	1.9884×10^{-4}	1.45	8.9654×10^{-7}	1.88

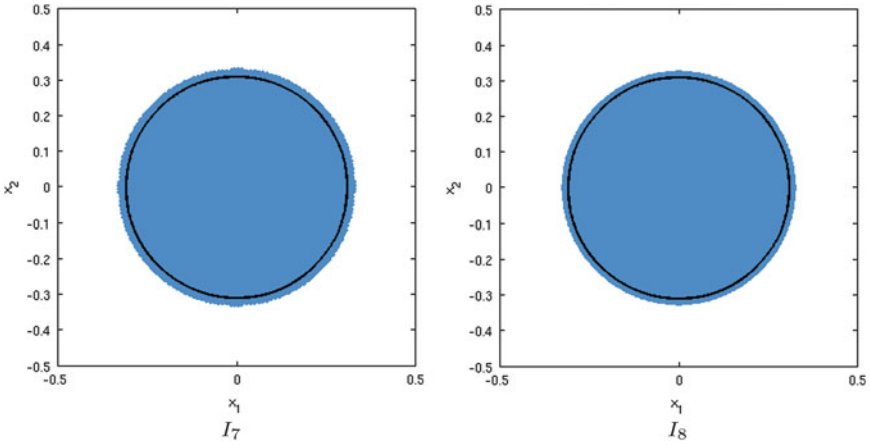


FIG. 1. Discrete coincidence sets I_7 (*left*) and I_8 (*right*) for Example 1

the state in the $H^1(\Omega)$ seminorm), and $\|\tilde{e}_{\bar{y},j}\|_\infty$ (the error of the state in the l_∞ norm) defined by

$$\|\tilde{e}_{\bar{y},j}\|_\infty = \max_{p\in\mathcal{N}_j} |\tilde{e}_{\bar{y},j}(p)|.$$

The approximations of the optimal control in these examples are given by the piecewise quadratic functions $\bar{u}_j \in V_j$ defined by (5.11). We take $\tilde{e}_{\bar{u},j} = \bar{u}_{j-1} - \bar{u}_j$ and evaluate $\|\tilde{e}_{\bar{u},j}\|_{L_2}$ (the error of the control in the $L_2(\Omega)$ norm). The orders of convergence in these examples are generated by the formulas

$$\ln(\|\tilde{e}_{\bar{y},j-1}\|/\|\tilde{e}_{\bar{y},j}\|)/\ln(2) \quad \text{and} \quad \ln(\|\tilde{e}_{\bar{u},j-1}\|/\|\tilde{e}_{\bar{u},j}\|)/\ln(2).$$

Example 2. In this example we take $y_d(x) = 10(\sin(2\pi(x_1+0.5)) + (x_2+0.5))$, $\psi_2(x) = 0.01$ and $\beta = 0.1$. The errors for the approximations of the state and the control are reported in Tables 2 and 3. The discrete state $\bar{y}_8$ and control $\bar{u}_8$ are depicted in Fig. 2.

TABLE 2
Energy and ℓ_∞ state errors for Example 2

j	$\|\tilde{e}_{\bar{y},j}\|_{h_j}/\|\bar{y}_8\|_{h_8}$	Order	$\|\tilde{e}_{\bar{y},j}\|_\infty$	Order
1	3.3661×10^{0}		1.1842×10^{-1}	
2	1.9062×10^{0}	0.82	3.9252×10^{-2}	1.59
3	7.4142×10^{-1}	1.36	6.5358×10^{-3}	2.59
4	4.4582×10^{-1}	0.73	2.0856×10^{-3}	1.66
5	2.2066×10^{-1}	1.01	6.2389×10^{-4}	1.74
6	1.0916×10^{-1}	1.02	1.8209×10^{-4}	1.78
7	5.4174×10^{-2}	1.01	4.5582×10^{-5}	2.00
8	2.7011×10^{-2}	1.00	1.1677×10^{-5}	1.96

TABLE 3
H^1 state errors and L_2 control errors for Example 2

j	$\|\tilde{e}_{\bar{y},j}\|_{H^1}/\|\bar{y}_8\|_{H^1}$	Order	$\|\tilde{e}_{\bar{u},j}\|_{L_2}/\|\bar{u}_8\|_{L_2}$	Order
1	4.9436×10^{0}		3.6418×10^{0}	
2	1.9541×10^{0}	1.34	2.1388×10^{0}	0.77
3	3.6305×10^{-1}	2.43	7.9054×10^{-1}	1.44
4	1.1593×10^{-1}	1.65	3.3568×10^{-1}	1.24
5	3.4745×10^{-2}	1.74	1.2506×10^{-1}	1.42
6	9.7768×10^{-3}	1.83	4.0060×10^{-2}	1.64
7	2.5550×10^{-3}	1.94	1.3141×10^{-2}	1.61
8	6.4538×10^{-4}	1.99	4.4595×10^{-3}	1.56

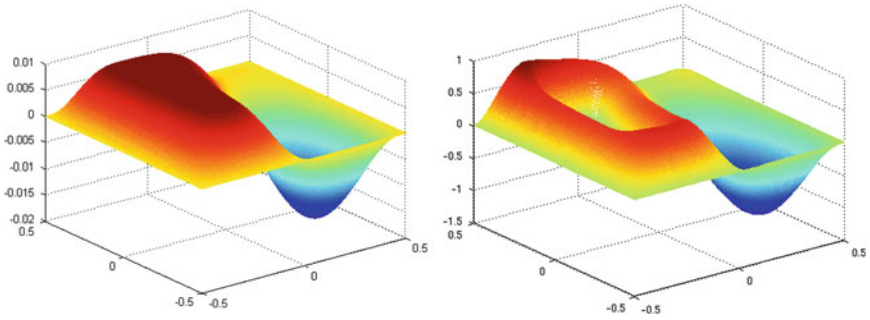


FIG. 2. Discrete state $\bar{y}_8$ (*left*) and control $\bar{u}_8$ (*right*) for Example 2

Example 3. In this example we take $y_d(x) = \sin(2\pi(x_1 + 0.5)(x_2 + 0.5))$, $\psi_2(x) = 0.1$ and $\beta = 10^{-3}$. The errors for the approximations of the state and the control are given in Tables 4 and 5. Figure 3 contains the plots for the discrete state $\bar{y}_8$ and the discrete control $\bar{u}_8$.

TABLE 4
Energy and ℓ_∞ state errors for Example 3

j	$\|\tilde{e}_{\bar{y},j}\|_{h_j}/\|\bar{y}_8\|_{h_8}$	Order	$\|\tilde{e}_{\bar{y},j}\|_\infty$	Order
1	2.6968×10^{0}		6.3179×10^{-1}	
2	1.2439×10^{0}	1.12	1.2247×10^{-1}	2.37
3	6.7643×10^{-1}	0.88	3.7137×10^{-2}	1.72
4	3.4552×10^{-1}	0.97	7.2368×10^{-3}	2.36
5	1.7485×10^{-1}	0.98	2.5667×10^{-3}	1.50
6	8.6434×10^{-2}	1.01	7.3986×10^{-4}	1.79
7	4.2673×10^{-2}	1.02	1.9661×10^{-4}	1.91
8	2.1230×10^{-2}	1.01	4.9541×10^{-5}	1.99

TABLE 5
H^1 state errors and L_2 control errors for Example 3

j	$\|\tilde{e}_{\bar{y},j}\|_{H^1}/\|\bar{y}_8\|_{H^1}$	Order	$\|\tilde{e}_{\bar{u},j}\|_{L_2}/\|\bar{u}_8\|_{L_2}$	Order
1	3.2873×10^{0}		2.6748×10^{0}	
2	1.1542×10^{0}	1.51	1.5626×10^{0}	0.78
3	3.3936×10^{-1}	1.77	9.4478×10^{-1}	0.73
4	9.2061×10^{-2}	1.88	3.5294×10^{-1}	1.42
5	2.6639×10^{-2}	1.79	1.2171×10^{-1}	1.54
6	7.2818×10^{-3}	1.87	4.1983×10^{-2}	1.54
7	1.8560×10^{-3}	1.97	1.3128×10^{-2}	1.68
8	4.6711×10^{-4}	1.99	4.4419×10^{-3}	1.56

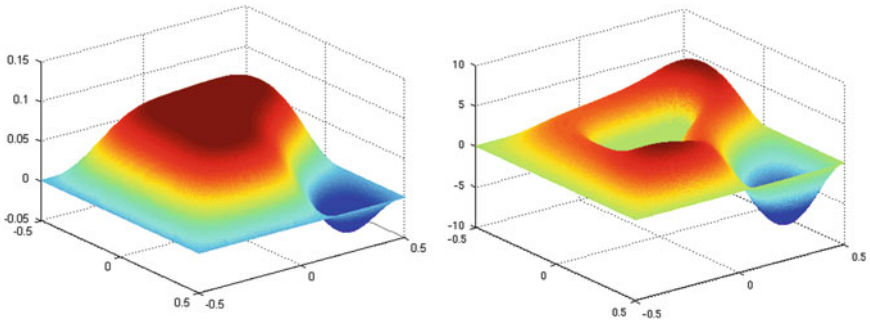


FIG. 3. Discrete state $\bar{y}_8$ (*left*) and control $\bar{u}_8$ (*right*) for Example 3

Example 4. In this example we take $y_d(x) = \sin(4\pi(x_1+0.5)(x_2+0.5))+1.5$, $\psi_2(x) = 1$ and $\beta = 10^{-4}$. The errors for the approximations in the state and the control are presented in Tables 6 and 7. The plots of the discrete state $\bar{y}_8$ and the discrete control $\bar{u}_8$ are given in Fig. 4.

TABLE 6
Energy and ℓ_∞ state errors for Example 4

j	$\|\tilde{e}_{\bar{y},j}\|_{h_j}/\|\bar{y}_8\|_{h_8}$	Order	$\|\tilde{e}_{\bar{y},j}\|_\infty$	Order
1	8.6145×10^{-1}		1.9915×10^{0}	
2	7.0373×10^{-1}	0.29	1.3112×10^{0}	0.60
3	3.6102×10^{-1}	0.96	3.7238×10^{-1}	1.82
4	2.3689×10^{-1}	0.61	8.4619×10^{-2}	2.14
5	1.1894×10^{-1}	0.99	1.6099×10^{-2}	2.39
6	5.9093×10^{-2}	1.01	5.6989×10^{-3}	1.50
7	2.9179×10^{-2}	1.02	1.5619×10^{-3}	1.87
8	1.4505×10^{-2}	1.01	3.4243×10^{-4}	2.19

TABLE 7
H^1 state errors and L_2 control errors for Example 4

j	$\|\tilde{e}_{\bar{y},j}\|_{H^1}/\|\bar{y}_8\|_{H^1}$	Order	$\|\tilde{e}_{\bar{u},j}\|_{L_2}/\|\bar{u}_8\|_{L_2}$	Order
1	1.3273×10^{0}		1.2796×10^{0}	
2	8.1485×10^{-1}	0.70	1.2466×10^{0}	0.04
3	2.9527×10^{-1}	1.46	8.4385×10^{-1}	0.56
4	1.0283×10^{-1}	1.52	4.0479×10^{-1}	1.06
5	3.3447×10^{-2}	1.62	1.6078×10^{-1}	1.33
6	8.4522×10^{-3}	1.98	5.3193×10^{-2}	1.60
7	2.2334×10^{-3}	1.92	1.7536×10^{-2}	1.60
8	5.5791×10^{-4}	2.00	5.7008×10^{-3}	1.62

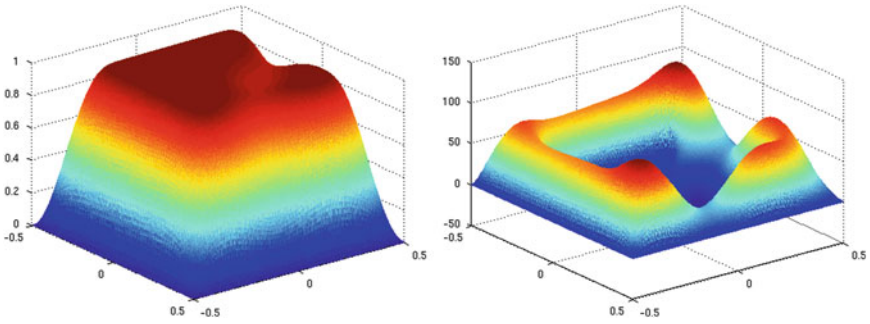


FIG. 4. Discrete state $\bar{y}_8$ (*left*) and control $\bar{u}_8$ (*right*) for Example 4

The numerical results in Tables 2–7 confirm the error estimate for $\|\bar{y}-\bar{y}_h\|_h$ in Theorem 5.1, since the index of elliptic regularity $\alpha = 1$ for a rectangular domain. On the other hand, the order of convergence for $\bar{y}_h$ is 2 for both the ℓ_∞ norm and the $H^1(\Omega)$ seminorm, which is better than the first order convergence predicted by Theorem 5.2 and Corollary 5.1; and the order of convergence for $\bar{u}_h$ is around 1.5, which is also better than the

first order convergence predicted by Theorem 5.3. The plots of the state and control in Figs. 2–4 also agree with the ones reported in [6, 55].

Example 5. In this example we take Ω to be the pentagonal domain obtained from the square $(-0.5, 0.5)^2$ by deleting the triangle with vertices $(0.5, 0)$, $(0.5, 0.5)$ and $(0, 0.5)$. We use the same data as Example 3, i.e., $\psi_2(x) = 0.1$, $y_d(x) = \sin(2\pi(x_1 + 0.5)(x_2 + 0.5))$, $\beta = 10^{-3}$ and $\gamma = 1$. The mesh parameter for the jth level uniform triangulation $\mathcal{T}_j$ is $h_j = 2^{-(j+1)}$. The errors for the approximate state $\bar{y}_j$ and approximate control $\bar{u}_j$ are presented in Tables 8 and 9. Since the index of elliptic regularity α for the pentagonal domain can be taken to be any number less than $1/3$ (cf. Remark 2.1), the results in Tables 8 and 9 agree with Theorems 5.1 and 5.3. However, for this example the magnitude of the l_∞ error of the state seems to be $O(h^{2\alpha})$ and the magnitude of the $H^1(\Omega)$ error of the state seems to be $O(h)$.

We also plot the discrete state $\bar{y}_8$ and control $\bar{u}_8$ in Fig. 5. The singular nature of $\bar{y}$ near the corners at $(0.5, 0)$ and $(0, 0.5)$ can be observed in the plot of $\bar{u}_6$.

TABLE 8
Energy and ℓ_∞ state errors for Example 5

j	$\|\tilde{e}_{\bar{y},j}\|_{h_j}/\|\bar{y}_8\|_{h_8}$	Order	$\|\tilde{e}_{\bar{y},j}\|_\infty$	Order
1	1.2749×10^{0}		1.2541×10^{-1}	
2	7.3054×10^{-1}	0.80	3.7113×10^{-2}	1.76
3	3.6072×10^{-1}	1.02	4.6868×10^{-3}	2.99
4	1.9576×10^{-1}	0.88	1.3685×10^{-3}	1.78
5	1.1763×10^{-1}	0.73	3.4423×10^{-4}	1.99
6	7.8971×10^{-2}	0.57	1.4986×10^{-4}	1.20
7	5.7723×10^{-2}	0.45	7.7115×10^{-5}	0.96
8	4.4159×10^{-2}	0.39	4.6283×10^{-5}	0.74

TABLE 9
H^1 state errors and L_2 control errors for Example 5

j	$\|\tilde{e}_{\bar{y},j}\|_{H^1}/\|\bar{y}_8\|_{H^1}$	Order	$\|\tilde{e}_{\bar{u},j}\|_{L_2}/\|\bar{u}_8\|_{L_2}$	Order
1	1.1561×10^{0}		1.7291×10^{0}	
2	3.4840×10^{-1}	1.73	1.0841×10^{0}	0.67
3	8.2785×10^{-2}	2.07	3.9368×10^{-1}	1.46
4	2.2172×10^{-2}	1.90	1.3651×10^{-1}	1.53
5	6.5259×10^{-3}	1.76	5.3314×10^{-2}	1.36
6	2.1309×10^{-3}	1.61	2.5368×10^{-2}	1.07
7	8.7597×10^{-4}	1.28	1.7158×10^{-2}	0.56
8	4.5704×10^{-4}	0.94	1.3122×10^{-2}	0.39

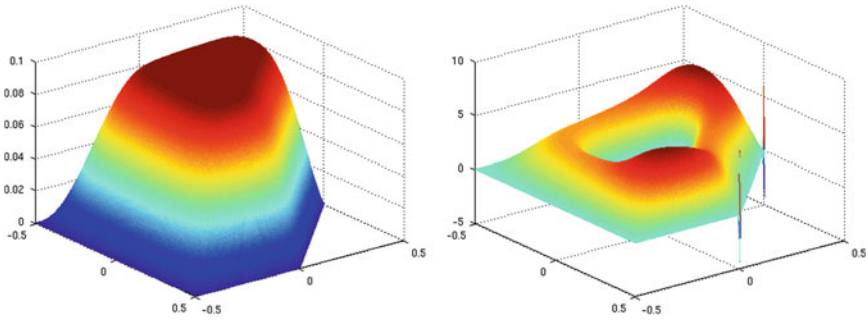


FIG. 5. Discrete state $\bar{y}_8$ (*left*) and control $\bar{u}_8$ (*right*) for Example 5

TABLE 10
Energy and ℓ_∞ state errors for Example 6

j	$\|\tilde{e}_{\bar{y},j}\|_{h_j}/\|\bar{y}_7\|_{h_7}$	Order	$\|\tilde{e}_{\bar{y},j}\|_\infty$	Order
1	5.1533×10^{-1}		2.1625×10^{-2}	
2	3.4325×10^{-1}	0.59	8.5234×10^{-3}	1.34
3	1.9843×10^{-1}	0.79	3.9322×10^{-3}	1.12
4	1.0957×10^{-1}	0.86	1.5313×10^{-3}	1.36
5	6.0125×10^{-2}	0.87	6.2314×10^{-4}	1.30
6	3.2836×10^{-2}	0.87	2.2180×10^{-4}	1.49
7	1.7795×10^{-2}	0.88	7.5681×10^{-5}	1.55

Example 6. In this example we solve the same problem in Example 5 on graded meshes obtained from a uniform triangulation $\mathcal{T}_0$ of the pentagonal domain by the refinement process in [10] (cf. Fig. 6), and we take the penalty parameter σ to be 20.

The errors of the approximate state $\bar{y}_j$ and approximate control $\bar{u}_j$ are reported in Tables 10 and 11. It is observed that the order of convergence for the state in the energy norm and for the control in the $L_2(\Omega)$ norm is about 1, which agrees with Theorems 5.1 and 5.3. On the other hand, the order of convergence for the state in the ℓ_∞ norm and the $H^1(\Omega)$ seminorm is about 1.5, which is better than the order of convergence predicted by Theorem 5.2 and Corollary 5.1.

The discrete state $\bar{y}_7$ and control $\bar{u}_3$ are depicted in Fig. 7. By comparing Figs. 5 and 7 we see that the graphs of the optimal states computed by a uniform mesh and a graded mesh are very similar. But the graph of the optimal control computed by graded meshes exhibited a more pronounced singular behavior near the corners $(0, 0.5)$ and $(0.5, 0)$ since the triangles at these corners are much smaller than the corresponding ones in a uniform mesh.

TABLE 11
H^1 state errors and L_2 control errors for Example 6

j	$\lvert\tilde{e}_{\bar{y},j}\rvert_{H^1}/\lvert\bar{y}_8\rvert_{H^1}$	Order	$\lVert\tilde{e}_{\bar{u},j}\rVert_{L_2}/\lVert\bar{u}_8\rVert_{L_2}$	Order
1	2.2461×10^{-1}		4.7166×10^{-1}	
2	1.0985×10^{-1}	1.03	2.9528×10^{-1}	0.68
3	4.5006×10^{-2}	1.29	1.5324×10^{-1}	0.95
4	1.5012×10^{-2}	1.58	7.7635×10^{-2}	0.98
5	5.7283×10^{-3}	1.39	4.0842×10^{-2}	0.93
6	1.9491×10^{-3}	1.56	2.1646×10^{-2}	0.92
7	6.6083×10^{-4}	1.56	1.1353×10^{-2}	0.93

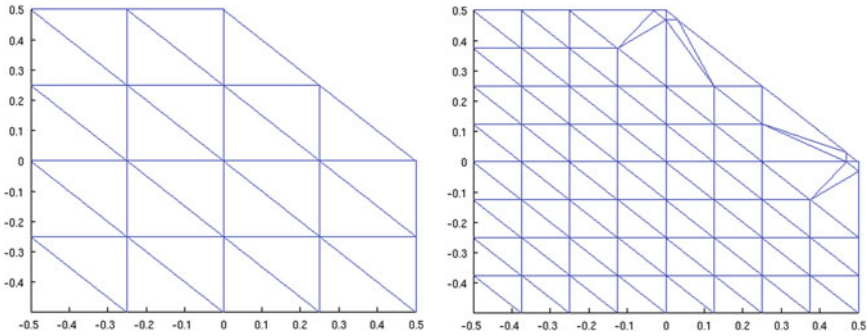


FIG. 6. Triangulation $\mathcal{T}_0$ (*left*) and $\mathcal{T}_1$ (*right*) for the pentagonal domain

7. Concluding Remarks. In this paper we have only considered the optimal control problem (1.1)–(1.3) on convex polygonal domains. It is possible to treat this problem on general polygonal domains, in which case the space $H^2(\Omega) \cap H_0^1(\Omega)$ will be replaced by the space $\{v \in H_0^1(\Omega) : \Delta v \in L_2(\Omega)\}$ that has been thoroughly analyzed in [45, 46] and the discretization will involve singular functions.

The three-dimensional version of (1.1)–(1.3) can also be solved as fourth order variational inequalities by finite element methods. For smooth domains, a straightforward extension of the approach in [15, 23–25] and this paper will lead to $O(h^{\frac{1}{2}})$ errors for the state in the energy norm and the control in the $L_2(\Omega)$ norm, similar to the error estimates in [37, 56]. Again we expect the convergence of the state in the $H^1(\Omega)$ norm and the $L_\infty(\Omega)$ norm to be of higher order.

These and other topics, such as the solution of optimal control problems with both state and control constraints as fourth order variational inequalities are subjects of ongoing investigations.

Acknowledgment. The authors would like to thank an anonymous referee for suggesting the piecewise quadratic approximation of the optimal control defined by (5.11). The work of S.C. Brenner, L.-Y. Sung, and Y.

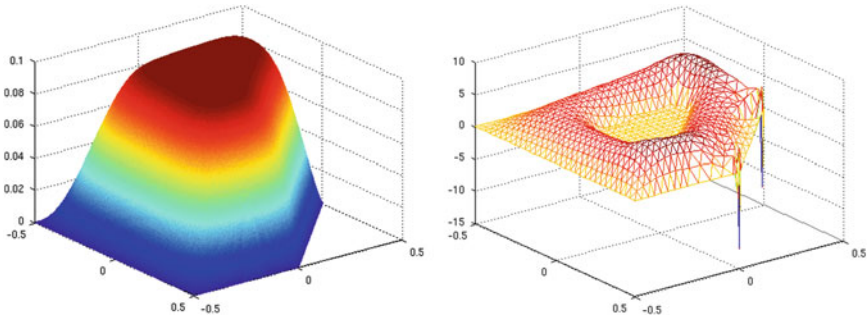


FIG. 7. Discrete state $\bar{y}_7$ (*left*) and control $\bar{u}_3$ (*right*) for Example 6

Zhang was supported in part by the National Science Foundation under Grant No. DMS-10-16332.

APPENDIX

A. Elliptic Regularity for Simply Supported Plates. In this appendix we summarize elliptic regularity results for the biharmonic equation on convex polygonal domains with the boundary conditions of simply supported plates and also discuss related results for the solution $\bar{y}$ of the obstacle problem (1.7). We will focus on the H^3 regularity (or lack thereof) for the solution since $\bar{y} \in H^3_{\text{loc}}(\Omega)$.

Let Ω be a convex polygonal domain with corners $p_1, \ldots, p_L$ and ω_ℓ be the interior angle of Ω at p_ℓ. Let $g \in L_2(\Omega)$ and $z \in H^2(\Omega) \cap H_0^1(\Omega)$ satisfy

$$\int_\Omega D^2 z : D^2 v \, dx = \int_\Omega g v \, dx \qquad \forall\, v \in H^2(\Omega) \cap H_0^1(\Omega). \tag{A.1}$$

It follows from (A.1) that $w = -\Delta z \in L_2(\Omega)$ has the following properties: (i) w is an H^2 function away from the corners of Ω, (ii) w vanishes on $\partial\Omega \setminus \{p_1, \ldots, p_L\}$. These two conditions then imply that

$$w = -\Delta z \text{ belongs to } H^2(\Omega) \cap H_0^1(\Omega) \tag{A.2}$$

and that z also satisfies

$$\int_\Omega \nabla z \cdot \nabla v \, dx = \int_\Omega w v \, dx \qquad \forall\, v \in H_0^1(\Omega).$$

Thus we can deduce the elliptic regularity of z from the elliptic regularity theory for the Laplace operator [36, 45, 58].

First of all,

$$z \text{ is an } H^4 \text{ function away from the corners of } \Omega, \tag{A.3}$$

which also follows directly from (A.1). Secondly we have

$$z \in H^3(\mathcal{N}_\ell) \qquad \text{if } \omega_\ell \leq \pi/2, \tag{A.4}$$

where $\mathcal{N}_\ell \subset \Omega$ is a neighborhood of p_ℓ. Finally, at a corner p_ℓ where $\omega_\ell > \pi/2$, we have

$$z - \kappa_\ell \varphi_\ell \in H^3(\mathcal{N}_\ell), \tag{A.5}$$

where $\mathcal{N}_\ell \subset \Omega$ is a neighborhood of p_ℓ, κ_ℓ is a constant (generalized stress intensity factor), and the singular function φ_ℓ is defined by

$$\varphi_\ell = r_\ell^{\pi/\omega_\ell} \sin\big((\pi/\omega_\ell)\theta_\ell\big). \tag{A.6}$$

Here (r_ℓ, θ_ℓ) are the polar coordinates at p_ℓ such that the two edges of Ω emanating from p_ℓ are given by $\theta_\ell = 0$ and $\theta_\ell = \omega_\ell$. Note that φ_ℓ is a harmonic function and $\varphi_\ell \in H^{1+(\pi/\omega_\ell)-\epsilon}(\mathcal{N}_\ell)$ for any $\epsilon > 0$.

Now we turn to the solution $\bar{y}$ of (1.7)/(1.9). Since the constraints in (1.3) are not active near $\partial\Omega$ because of (1.4b), we have

$$\int_\Omega \big[\beta\big(D^2(\rho_1\bar{y}) : D^2 w\big)\big]\, dx = \int_\Omega \beta\,\big[D^2(\rho_1\bar{y}) : D^2\big((1-\rho_2)w\big)\big]\, dx + \int_\Omega \rho_2(f - \gamma\bar{y})w\, dx$$

for all $w \in H^2(\Omega) \cap H_0^1(\Omega)$, where $\rho_1 = \rho_2 = 1$ near $\partial\Omega$, $\rho_1 = 1$ on the support of ρ_2, and the support of ρ_1 is disjoint from the active set where $\bar{y}(x) = \psi_1(x)$ or $\psi_2(x)$. Note that standard interior elliptic regularity [62, Sect. 20] implies

$$\int_\Omega \big[D^2(\rho_1\bar{y}) : D^2\big((1-\rho_2)w\big)\big]\, dx = \int_\Omega (1-\rho_2)[\Delta^2(\rho_1\bar{y})]w\, dx,$$

where $(1-\rho_2)\Delta^2(\rho_1\bar{y}) \in L_2(\Omega)$.

Therefore $z = \rho_1\bar{y}$ satisfies (A.1) with $g = \rho_2(f - \gamma\bar{y})/\beta + (1-\rho_2)\Delta^2(\rho_1\bar{y}) \in L_2(\Omega)$. Combining (A.2)–(A.6) and the fact that $\bar{y} \in H^3_{\rm loc}(\Omega)$, we can draw the following conclusions about $\bar{y}$.

- The function $\Delta\bar{y}$ belongs to $H_0^1(\Omega)$. Therefore $\bar{u} = -\Delta\bar{y}$ belongs to $H_0^1(\Omega)$ for the optimal control problem (1.1)–(1.3).
- Let α_ℓ be chosen according to (2.4). Then $\bar{y} \in H^{2+\alpha_\ell}(\mathcal{N}_\ell)$, where $\mathcal{N}_\ell\,(\subset \Omega)$ is a neighborhood of p_ℓ. Globally we have $\bar{y} \in H^{2+\alpha}(\Omega)$ where $\alpha = \min_{1\leq\ell\leq L} \alpha_\ell$.
- We can write $\bar{y} = \bar{y}_S + \bar{y}_R$, where $\bar{y}_R \in H^3(\Omega) \cap H_0^1(\Omega)$, $\Delta\bar{y}_R \in H_0^1(\Omega)$ and $\bar{y}_S$ have the following properties.
  - $\bar{y}_S$ is an H^3 function away from the corners of Ω where the angles are $> \pi/2$.

  - $\bar{y}_S$ is a multiple of φ_ℓ in a neighborhood $\mathcal{N}_\ell$ of a corner p_ℓ where $\omega_\ell > \pi/2$.
  - $\Delta\bar{y}_S$ belongs to $H_0^1(\Omega)$.
- Since $r_\ell^{1-\alpha_\ell}(\partial^\mu\varphi_\ell) \in L_2(\mathcal{N}_\ell)$ for $|\mu| = 3$, we have $\Phi(\partial^\mu\bar{y}_S) \in L_2(\Omega)$ for $|\mu| = 3$ and hence

$$\Phi(\partial^\mu\bar{y}) \in L_2(\Omega) \quad \text{for } |\mu| = 3, \tag{A.7}$$

  where the function Φ is defined by (2.3).
- Since $r_\ell^{-\alpha_\ell}(\partial^\mu\varphi_\ell) \in L_2(\mathcal{N}_\ell)$ for $|\mu| = 2$, we have $\Psi(\partial^\mu\bar{y}_S) \in L_2(\Omega)$ for $|\mu| = 2$ and hence

$$\Psi(\partial^\mu\bar{y}) \in L_2(\Omega) \quad \text{for } |\mu| = 2, \tag{A.8}$$

  where the function Ψ is defined by

$$\Psi(x) = \prod_{\ell=1}^{L} |p_\ell - x|^{-\alpha_\ell}. \tag{A.9}$$

Finally we note that (cf. [36, Theorem AA.3 and Theorem AA.7])

$$|\bar{y}|_{H^{2+\alpha}(\Omega)} \le C_\Omega \sum_{|\mu|=3} \|\Phi(\partial^\mu\bar{y})\|_{L_2(\Omega)}. \tag{A.10}$$

B. An Enriching Operator. In this appendix we construct the enriching operator introduced in Sect. 3.2. Such operators have played an important role in the design and analysis of fast solvers for nonconforming finite element methods [11, 12, 22, 26].

Let $\tilde{V}_h \subset H^1(\Omega)$ be the $\mathbb{P}_2$ Lagrange finite element space associated with $\mathcal{T}_h$ and $\tilde{W}_h \subset H^2(\Omega)$ be the $\mathbb{P}_6$ Argyris finite element space [2] associated with $\mathcal{T}_h$. The degrees of freedom of $w \in \tilde{W}_h$ (cf. Fig. 8) consist of the values of the derivatives of w up to second order at the vertices of $\mathcal{T}_h$, the values of w at the midpoints of the edges of $\mathcal{T}_h$ and at the centers of the triangles of $\mathcal{T}_h$, and the values of the normal derivative of w at two nodes on each edge in $\mathcal{E}_h$.

The enriching operator $E_h : \tilde{V}_h \longrightarrow \tilde{W}_h$ is defined by averaging as follows (cf. Sect. 2.1 for the notation).

(i) Let N be a degree of freedom associated with an interior node p. We define

$$N(E_h v) = \frac{1}{|\mathcal{T}_p|} \sum_{T\in\mathcal{T}_p} N(v_T).$$

(ii) Let N be a degree of freedom involving the normal derivative associated with a boundary node interior to an edge $e \in \mathcal{E}_h^b$. We define

$$N(E_h v) = N(v_{T_e}).$$

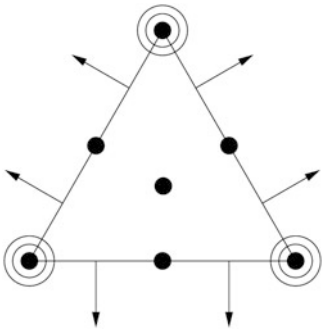


FIG. 8. Degrees of freedom for the $\mathbb{P}_6$ Argyris finite element

(iii) Let p be a boundary node which is not a corner of Ω such that p is the common endpoint of two edges $e_1, e_2 \in \mathcal{E}_h^b$. For any degree of freedom N associated with p, we define

$$N(E_h v) = \frac{1}{2}\big[N(v_{T_{e_1}}) + N(v_{T_{e_2}})\big].$$

(iv) Let p be a corner of Ω. Then p is the common endpoint of $e_1, e_2 \in \mathcal{E}_h^b$. Let t_j (resp. n_j) be a unit tangent (resp. normal) of e_j. We define

$$\begin{aligned}
(E_h v)(p) &= v(p), \\
(\partial(E_h v)/\partial t_j)(p) &= (\partial v_{T_{e_j}}/\partial t_j)(p) && \text{for } j = 1, 2, \\
(\partial^2(E_h v)/\partial t_j^2)(p) &= (\partial^2 v_{T_{e_j}}/\partial t_j^2)(p) && \text{for } j = 1, 2, \\
(\partial^2(E_h v)/\partial t_1 \partial n_1)(p) &= (\partial^2 v_{T_{e_1}}/\partial t_1 \partial n_1)(p).
\end{aligned}$$

REMARK B.1. *We can also replace the last equation in (iv) by*

$$(\partial^2(E_h v)/\partial t_2 \partial n_2)(p) = (\partial^2 v_{T_{e_2}}/\partial t_2 \partial n_2)(p).$$

Since v is continuous at the vertices, the relation (3.6) follows immediately from (i), (iii), and (iv). It is also easy to check that

$$E_h v \in W_h = \tilde{W}_h \cap H_0^1(\Omega) \subset H^2(\Omega) \cap H_0^1(\Omega) \quad \text{if} \quad v \in V_h = \tilde{V}_h \cap H_0^1(\Omega).$$

We now turn to the derivations of (3.8) and (3.9). Let $T \in \mathcal{T}_h$ be arbitrary. Since $v = E_h v$ at the vertices and the center of T, we have, by scaling,

$$\|v - E_h v\|^2_{L_2(T)} \lesssim h_T^4 \left(\sum_{p\in\mathcal{V}_T} |\nabla(v - E_h v)(p)|^2 + \sum_{p\in\mathcal{N}_T} \left| \frac{\partial(v - E_h v)}{\partial n}(p) \right|^2 + \sum_{p\in\mathcal{V}_T} h_T^2 |D^2(v - E_h v)(p)|^2 \right) \tag{B.1}$$

for all $v \in \tilde{V}_h$, where $\mathcal{N}_T$ is the set of the six nodes on ∂T associated with the degrees of freedom of the $\mathbb{P}_6$ Argyris finite element that involve the normal derivative (cf. Fig. 8).

Let $p \in \mathcal{V}_T$ be interior to Ω. Since the tangential derivative of $v - E_h v$ is continuous across element boundaries, we have, by the definition of E_h and a standard inverse estimate,

$$\begin{aligned} |\nabla(v - E_h v)(p)|^2 &= \left| \frac{1}{|\mathcal{T}_p|} \sum_{T'\in\mathcal{T}_p} (\nabla v_T(p) - \nabla v_{T'}(p)) \right|^2 \\ &\lesssim \sum_{e\in\mathcal{E}_p^i} |e|^{-1} \|[[\partial v/\partial n]]\|^2_{L_2(e)} \end{aligned} \tag{B.2}$$

where $\mathcal{E}_p^i$ is the set of the edges in $\mathcal{E}_h^i$ sharing p as a common endpoint. Similarly, we have

$$\begin{aligned} |D^2(v - E_h v)(p)|^2 &= \left| \frac{1}{|\mathcal{T}_p|} \sum_{T'\in\mathcal{T}_p} D^2(v_T - v_{T'})(p) \right|^2 \\ &\lesssim \sum_{T'\in\mathcal{T}_p} h_{T'}^{-2} |v|^2_{H^2(T')}. \end{aligned} \tag{B.3}$$

The estimates (B.2) and (B.3) are also valid for $p \in \partial\Omega$ by similar arguments.

Now we consider $p \in \mathcal{N}_T$. If p is a boundary node, then $\left|(\partial(v - E_h v)/\partial n)(p)\right| = 0$ by the definition of E_h. Otherwise we have, by a standard inverse estimate,

$$|\partial(v - E_h v)/\partial n(p)|^2 \lesssim |e|^{-1} \|[[\partial v/\partial n]]\|^2_{L_2(e)} \tag{B.4}$$

for some $e \in \mathcal{E}_h^i$.

Combining (B.1)–(B.4), we obtain the estimate (3.8) for $m = 0$, which then implies the estimates for $m = 1$ and 2 through standard inverse estimates.

For the operator $E_h \circ \Pi_h$, first we observe that it is a bounded linear operator from $H^{2+s}(\mathcal{S}_T)$ into $H^2(T)$ because of (2.21) and (3.8). Furthermore, by construction, $E_h \Pi_h \zeta = \zeta$ on T if $\zeta \in \mathbb{P}_2(\mathcal{S}_T)$. Hence the estimate (3.9) follows from the Bramble–Hilbert lemma (cf. [9, 38]).

REFERENCES

[1] T. Apel, A.-M. Sändig, and J. Whiteman, *Graded mesh refinement and error estimates for finite element solutions of elliptic boundary value problems in non-smooth domains*, Math. Methods Appl. Sci., 19 (1996), pp. 63–85.

[2] J. Argyris, I. Fried, and D. Scharpf, *The TUBA family of plate elements for the matrix displacement method*, Aero. J. Roy. Aero. Soc., 72 (1968), pp. 701–709.

[3] J. Aubin, *Approximation of variational inequations*, in Functional Analysis and Optimization, Academic Press, New York, 1966, pp. 7–14.

[4] M. Bergounioux, K. Ito, and K. Kunisch, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim., 37 (1999), pp. 1176–1194 (electronic).

[5] M. Bergounioux and K. Kunisch, *Augmented Lagrangian techniques for elliptic state constrained optimal control problems*, SIAM J. Control Optim., 35 (1997), pp. 1524–1543.

[6] ———, *Primal-dual strategy for state-constrained optimal control problems*, Comput. Optim. Appl., 22 (2002), pp. 193–224.

[7] ———, *On the structure of Lagrange multipliers for state-constrained optimal control problems*, Systems Control Lett., 48 (2003), pp. 169–176.

[8] H. Blum and R. Rannacher, *On the boundary value problem of the biharmonic operator on domains with angular corners*, Math. Methods Appl. Sci., 2 (1980), pp. 556–581.

[9] J. Bramble and S. Hilbert, *Estimation of linear functionals on Sobolev spaces with applications to Fourier transforms and spline interpolation*, SIAM J. Numer. Anal., 7 (1970), pp. 113–124.

[10] J. Brannick, H. Li, and L. Zikatanov, *Uniform convergence of the multigrid V-cycle on graded meshes for corner singularities*, Numer. Linear Algebra Appl., 15 (2008), pp. 291–306.

[11] S.C. Brenner, *A two-level additive Schwarz preconditioner for nonconforming plate elements*, Numer. Math., 72 (1996), pp. 419–447.

[12] ———, *Convergence of nonconforming multigrid methods without full elliptic regularity*, Math. Comp., 68 (1999), pp. 25–53.

[13] ———, *C^0 Interior Penalty Methods*, in Frontiers in Numerical Analysis-Durham 2010, J. Blowey and M. Jensen, eds., vol. 85 of Lecture Notes in Computational Science and Engineers, Springer-Verlag, Berlin-Heidelberg, 2012, pp. 79–147.

[14] S.C. Brenner and C. Carstensen, *Finite Element Methods*, in Encyclopedia of Computational Mechanics, E. Stein, R. de Borst, and T. Hughes, eds., Wiley, Weinheim, 2004, pp. 73–118.

[15] S.C. Brenner, C. Davis, and L.-Y. Sung, *A generalized finite element method for the displacement obstacle problem of clamped Kirchhoff plates*, arXiv:1212.3026 [math.NA].

[16] S.C. Brenner, S. Gu, T. Gudi, and L.-Y. Sung, *A quadratic C^0 interior penalty method for linear fourth order boundary value problems with boundary conditions of the Cahn-Hilliard type*, SIAM J. Numer. Anal., 50 (2012), pp. 2088–2110.

[17] S.C. Brenner, T. Gudi, and L.-Y. Sung, *An a posteriori error estimator for a quadratic C^0 interior penalty method for the biharmonic problem*, IMA J. Numer. Anal., 30 (2010), pp. 777–798.

[18] S.C. Brenner and M. Neilan, *A C^0 interior penalty method for a fourth order elliptic singular perturbation problem*, SIAM J. Numer. Anal., 49 (2011), pp. 869–892.

[19] S.C. Brenner, M. Neilan, and L.-Y. Sung, *Isoparametric C^0 interior penalty methods for plate bending problems on smooth domains*, Calcolo, 49 (2012), pp. 35–67.

[20] S.C. Brenner and L.R. Scott, *The Mathematical Theory of Finite Element Methods* (*Third Edition*), Springer-Verlag, New York, 2008.
[21] S.C. Brenner and L.-Y. Sung, *C^0 interior penalty methods for fourth order elliptic boundary value problems on polygonal domains*, J. Sci. Comput., 22/23 (2005), pp. 83–118.
[22] ———, *Multigrid algorithms for C^0 interior penalty methods*, SIAM J. Numer. Anal., 44 (2006), pp. 199–223.
[23] S.C. Brenner, L.-Y. Sung, H. Zhang, and Y. Zhang, *A quadratic C^0 interior penalty method for the displacement obstacle problem of clamped Kirchhoff plates*, SIAM J. Numer. Anal., 50 (2012), pp. 3329–3350.
[24] ———, *A Morley finite element method for the displacement obstacle problem of clamped Kirchhoff plates*, J. Comp. Appl. Math., DOI:10.1016/j.cam.2013.02.028, published online May 29, 2013.
[25] S.C. Brenner, L.-Y. Sung, and Y. Zhang, *Finite element methods for the displacement obstacle problem of clamped plates*, Math. Comp., 81 (2012), pp. 1247–1262.
[26] S.C. Brenner and K. Wang, *Two-level additive Schwarz preconditioners for C^0 interior penalty methods*, Numer. Math., 102 (2005), pp. 231–255.
[27] ———, *An iterative substructuring algorithm for a C^0 interior penalty method*, Elect. Trans. Numer. Anal., 39 (2012), pp. 313–332.
[28] S.C. Brenner, K. Wang, and J. Zhao, *Poincaré-Friedrichs inequalities for piecewise H^2 functions*, Numer. Funct. Anal. Optim., 25 (2004), pp. 463–478.
[29] H. Brézis and G. Stampacchia, *Sur la régularité de la solution d'inéquations elliptiques*, Bull. Soc. Math. France, 96 (1968), pp. 153–180.
[30] F. Brezzi, W. Hager, and P.-A. Raviart, *Error estimates for the finite element solution of variational inequalities*, Numer. Math., 28 (1977), pp. 431–443.
[31] ———, *Error estimates for the finite element solution of variational inequalities. II. Mixed methods*, Numer. Math., 31 (1978/79), pp. 1–16.
[32] L. Caffarelli and A. Friedman, *The obstacle problem for the biharmonic operator*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 6 (1979), pp. 151–184.
[33] E. Casas, *Control of an elliptic problem with pointwise state constraints*, SIAM J. Control Optim., 24 (1986), pp. 1309–1318.
[34] S. Cherednichenko, K. Krumbiegel, and A. Rösch, *Error estimates for the Lavrentiev regularization of elliptic optimal control problems*, Inverse Problems, 24 (2008), p. 055003.
[35] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[36] M. Dauge, *Elliptic Boundary Value Problems on Corner Domains,* Lecture Notes in Mathematics 1341, Springer-Verlag, Berlin-Heidelberg, 1988.
[37] K. Deckelnick and M. Hinze, *Convergence of a finite element approximation to a state-constrained elliptic control problem*, SIAM J. Numer. Anal., 45 (2007), pp. 1937–1953 (electronic).
[38] T. Dupont and R. Scott, *Polynomial approximation of functions in Sobolev spaces*, Math. Comp., 34 (1980), pp. 441–463.
[39] G. Engel, K. Garikipati, T. Hughes, M. Larson, L. Mazzei, and R. Taylor, *Continuous/discontinuous finite element approximations of fourth order elliptic problems in structural and continuum mechanics with applications to thin beams and plates, and strain gradient elasticity*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 3669–3750.
[40] R. Falk, *Error estimates for the approximation of a class of variational inequalities*, Math. Comp., 28 (1974), pp. 963–971.
[41] J. Frehse, *Zum Differenzierbarkeitsproblem bei Variationsungleichungen höherer Ordnung*, Abh. Math. Sem. Univ. Hamburg, 36 (1971), pp. 140–149.
[42] ———, *On the regularity of the solution of the biharmonic variational inequality*, Manuscripta Math., 9 (1973), pp. 91–103.
[43] A. Friedman, *Variational Principles and Free-Boundary Problems*, Robert E. Krieger Publishing Co. Inc., Malabar, FL, second ed., 1988.

[44] W. Gong and N. Yan, *A mixed finite element scheme for optimal control problems with pointwise state constraints*, J. Sci. Comput., 46 (2011), pp. 182–203.
[45] P. Grisvard, *Elliptic Problems in Non Smooth Domains*, Pitman, Boston, 1985.
[46] ———, *Singularities in Boundary Value Problems*, Masson, Paris, 1992.
[47] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2002), pp. 865–888 (2003).
[48] M. Hintermüller and K. Kunisch, *PDE-constrained optimization subject to pointwise constraints on the control, the state, and its derivative*, SIAM J. Optim., 20 (2009), pp. 1133–1156.
[49] M. Hintermüller and W. Ring, *A level set approach for the solution of a state-constrained optimal control problem*, Numer. Math., 98 (2004), pp. 135–166.
[50] M. Hinze and C. Meyer, *Variational discretization of Lavrentiev-regularized state constrained elliptic optimal control problems*, Comput. Optim. Appl., 46 (2010), pp. 487–510.
[51] M. Hinze and A. Schiela, *Discretization of interior point methods for state constrained elliptic optimal control problems: optimal error estimates and parameter adjustment*, Comput. Optim. Appl., 48 (2011), pp. 581–600.
[52] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Society for Industrial and Applied Mathematics, Philadelphia, 2000.
[53] H. Lewy and G. Stampacchia, *On the regularity of the solution of a variational inequality*, Comm. Pure Appl. Math., 22 (1969), pp. 153–188.
[54] J.-L. Lions and G. Stampacchia, *Variational inequalities*, Comm. Pure Appl. Math., 20 (1967), pp. 493–519.
[55] W. Liu, W. Gong, and N. Yan, *A new finite element approximation of a state-constrained optimal control problem*, J. Comput. Math., 27 (2009), pp. 97–114.
[56] C. Meyer, *Error estimates for the finite-element approximation of an elliptic control problem with pointwise state and control constraints*, Control Cybernet., 37 (2008), pp. 51–83.
[57] C. Meyer, A. Rösch, and F. Tröltzsch, *Optimal control of PDEs with regularized pointwise state constraints*, Comput. Optim. Appl., 33 (2006).
[58] S. Nazarov and B. Plamenevsky, *Elliptic Problems in Domains with Piecewise Smooth Boundaries*, de Gruyter, Berlin-New York, 1994.
[59] J.-F. Rodrigues, *Obstacle Problems in Mathematical Physics*, North-Holland Publishing Co., Amsterdam, 1987.
[60] A. Schiela and A. Günther, *An interior point algorithm with inexact step computation in function space for state constrained optimal control*, Numer. Math., 119 (2011), pp. 373–407.
[61] L. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.
[62] J. Wloka, *Partial Differential Equations*, Cambridge University Press, Cambridge, 1987.

A LOCAL TIMESTEPPING RUNGE–KUTTA DISCONTINUOUS GALERKIN METHOD FOR HURRICANE STORM SURGE MODELING

CLINT DAWSON*

Abstract. In this paper, we describe a local timestepping (LTS) approach within the Runge–Kutta discontinuous Galerkin (RKDG) method, and the application of this method to the modeling of hurricane storm surge. Modeling storm surge requires the numerical solution of the shallow water equations with wind and atmospheric pressure forcing, over complex domains which include wet and dry regions. The RKDG method is well suited for these applications; however, well-resolved simulations of storm surge can require highly graded meshes, which can lead to severe global CFL constraints. The LTS approach allows for elements to use timesteps which approximately satisfy only local CFL conditions. We describe a fully parallel implementation of the LTS method within an RKDG shallow water simulator, developed by the author and collaborators over a period of several years. We demonstrate that, for a specific hurricane, namely Hurricane Ike, the LTS method can reduce parallel run-times by nearly a factor of two with no degradation in accuracy.

Key words. Local timestepping, Multirate methods, Shallow water equations, Runge–Kutta discontinuous Galerkin methods, Hurricane storm surge

AMS(MOS) subject classifications.

1. Introduction.

In this paper, we continue our investigation of local timestepping (LTS) methods for Runge–Kutta discontinuous Galerkin (RKDG) discretizations of conservation laws. We are specifically interested in these methods for the numerical solution of shallow water systems in the coastal ocean. The application of RKDG methods to the shallow water equations has been extensively described in a series of papers by the author and collaborators; see [4, 8, 15–17, 23]. In particular, the method has been shown to be robust for a wide variety of flows in coastal seas, from tidal flows to more extreme flows such as hurricane-driven storm surge. Extensive verification and validation work has been performed, and the result is a robust shallow water simulator based on the RKDG formulation.

One of the remaining research issues with the RKDG method is the efficiency question. It has been debated in the literature that DG methods in general are inefficient, due to the fact that compared to standard continuous Galerkin methods, they have far more degrees of freedom, at least on the same computational mesh. Another issue with RKDG methods are the timestep constraints due to the CFL condition. For coastal ocean applications, the meshes are often highly graded in the continental shelf, near-shore, and coastal inland regions. Mesh elements may vary in

*The Institute for Computational Science and Engineering, The University of Texas at Austin, Austin, TX 78712, USA, clint@ices.utexas.edu

X. Feng et al. (eds.), *Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations*, The IMA Volumes in Mathematics and its Applications 157, DOI 10.1007/978-3-319-01818-8_5,

size from several square kilometers to $\mathcal{O}(100)$ square meters. Furthermore, the eigenvalues of the system may also vary by orders of magnitude over the domain. Thus, enforcing a global CFL timestep constraint based on satisfying a highly localized CFL condition can lead to inefficiencies in the timestepping in RKDG methods, and in explicit methods in general.

LTS is one way to deal with this problem. In LTS, the timestep size varies on each element and is determined by a local CFL condition. Such methods have been previously derived and applied to conservation laws by a number of authors [6, 7, 14, 19, 20]. This procedure is also similar to multirate methods and adaptive mesh refinement (AMR) methods. The AMR method used in the GeoClaw software [2, 3] uses forward Euler timestepping with timesteps dictated by local CFL constraints on each refinement patch. The fluxes at the interfaces between levels are conserved in the same way described here, and as described in [19]. The multirate methods described in [5] for conservation laws are shown to preserve second order accuracy and the TVD property.

In previous work [9, 21], we developed and applied an LTS method within the framework of a second order RKDG method and applied the method to the solution of the shallow water equations. In [21], the accuracy of the method was examined and comparisons with RKDG solutions with no LTS were given for some relatively small-scale model problems. In [9], we extended the LTS method to large-scale coastal flow applications. These applications require large domains, highly unstructured meshes, and can involve simulation of phenomena over several days. Thus, efficient simulation requires the use of parallel computing. The extension of LTS methodologies to distributed memory parallel computers is nontrivial, as we discuss in more detail below.

The rest of this paper is arranged as follows. In the next section, we outline the shallow water equations, the RKDG method and discuss the implementation of the LTS method in a parallel computing environment. In Sect. 3, we provide results for LTS in a highly challenging application, the modeling of hurricane storm surge, focusing in particular on Hurricane Ike (2008).

2. The Shallow Water Equations. The shallow water equations (SWE) are based on the three-dimensional Reynold's averaged Navier–Stokes equations for a Newtonian fluid. Averaging these equations over the vertical depth of the water H and applying kinematic and no-flow boundary conditions at the top and the bottom gives rise to the conservative form of the SWE:

$$\frac{\partial H}{\partial t} + S_p \frac{\partial (uH)}{\partial x} + \frac{\partial (vH)}{\partial y} = 0, \tag{2.1}$$

$$\frac{\partial (uH)}{\partial t} + S_p \frac{\partial \left(u^2 H + \frac{1}{2} g H^2\right)}{\partial x} + \frac{\partial (uvH)}{\partial y} = g S_p H \frac{\partial \eta}{\partial x} + (\tau_x^{\xi} - \tau_x^{\eta}) + F_x, \tag{2.2}$$

$$\frac{\partial(vH)}{\partial t} + \frac{\partial\left(v^2H + \frac{1}{2}gH^2\right)}{\partial y} + S_p\frac{\partial(uvH)}{\partial x} = gH\frac{\partial\eta}{\partial y} + (\tau_y^\xi - \tau_y^\eta) + F_y, \tag{2.3}$$

where u and v are depth-average velocities, ξ is the water elevation relative to the geoid, $\eta = H - \xi$ is the bathymetry relative to the geoid, g is gravitational acceleration, $\{\tau_{x,y}^\xi, \tau_{x,y}^\eta\}$ are the surface (wind) and bed (bottom friction) stresses, respectively, and $F_{x,y}$ accounts for other external forces, such as Coriolis force and tidal potential. The function $S_p(y)$ is a spherical correction factor which transforms the SWE in spherical coordinates ϕ, λ to Cartesian coordinates x, y using an orthogonal cylindrical projection; see [8]. To arrive at these equations, a number of assumptions have been made; (1) the vertical acceleration of a fluid particle is small in comparison with the acceleration of gravity, (2) shear stresses due to the vertical velocity are small, and (3) the horizontal shear terms, $\{\partial^2u/\partial x^2, \partial^2u/\partial y^2, \partial^2v/\partial x^2, \partial^2v/\partial y^2\}$ are small compared to vertical shears, $\{\partial^2u/\partial z^2, \partial^2v/\partial z^2\}$.

For closure, the bed stress terms must be parameterized via the depth-averaged velocities. The bed stress is often approximated by linear or quadratic functions of the velocities; however, we have used a hybrid form proposed by Westerink et al. [22] which varies the bottom-friction coefficient with the water column depth:

$$\tau_x^\eta = uH\left(C_f\frac{\sqrt{u^2+v^2}}{H}\right), \quad \tau_y^\eta = vH\left(C_f\frac{\sqrt{u^2+v^2}}{H}\right), \tag{2.4}$$

where,

$$C_f = C_{\text{fmin}}\left(1 + \left(\frac{H_{\text{break}}}{H}\right)^{f_\theta}\right)^{f_\gamma/f_\theta}. \tag{2.5}$$

This formulation applies a depth-dependent, Manning-type friction law below the break depth (H_{break}) and a standard Chezy friction law when the depth is greater than the break depth. For the applications below, C_{fmin} is allowed to vary, since the bed surfaces change.

The wind surface stress is computed by a standard quadratic drag law. Define

$$\tau_x^\xi = C_d\rho_{\text{air}}|\mathbf{W}|W_x, \tag{2.6}$$

$$\tau_y^\xi = C_d\rho_{\text{air}}|\mathbf{W}|W_y. \tag{2.7}$$

Here $\mathbf{W} = (W_x, W_y)$ is the wind speed sampled at a 10 m height over a 15-min time period and ρ_{air} is the air density. The drag coefficient is defined by Garratt's drag formula [10]:

$$C_d = (0.75 + 0.06|\mathbf{W}|) * 10^{-3}. \tag{2.8}$$

We also remark that the wind surface stress is capped so that its magnitude is never greater than 0.002.

3. Numerical Methods.

3.1. The Discontinuous Galerkin Finite Element Method. Rewrite the SWE as a hyperbolic system with a source term:

$$\frac{\partial w}{\partial t} + \nabla \cdot \mathbf{F}(w) = s. \tag{3.1}$$

To formulate the semi-discrete DG method for (3.1), the physical domain, Ω, is first partitioned into non-overlapping finite elements, K_i for $i = 1, 2, \ldots N$. If $P^k(K_i)$ is defined as the space of polynomials of degree $\leq k$ over element i, the DG method can be formulated as seeking a piecewise smooth approximation $w_h \mid_{K_i} \in P^k(K_i)$, obtained by multiplying (3.1) by a test function v_h, and integrating by parts:

$$\int_{K_i} \frac{\partial w_h}{\partial t}\, v_h \, dx - \int_{K_i} \mathbf{F}(w_h) \cdot \nabla v_h \, dx + \int_{\partial K_i} \hat{\mathbf{F}} \cdot \mathbf{n}_i v_h \, ds = \int_{K_i} s v_h \, dx. \tag{3.2}$$

Here $\hat{\mathbf{F}}(w_{h,-}, w_{h,+})$ is an approximation to the normal flux at the element boundaries given discontinuous left and right states $w_{h,-}$ and $w_{h,+}$, respectively, and $\mathbf{n}_i$ is the unit outward normal to ∂K_i. Many different numerical fluxes $\hat{\mathbf{F}}$ have been proposed in the literature. For the results presented here, the local Lax-Friedrichs flux is used.

Expanding the solution w_h in terms of its degrees of freedom, (3.1) can be written as system of ODEs:

$$\mathbf{M} \frac{d\tilde{\mathbf{w}}}{dt} = \mathbf{b}. \tag{3.3}$$

3.2. Runge–Kutta Time Discretization. For time integration, the system of equations

$$\frac{d\tilde{\mathbf{w}}}{dt} = L_h(\tilde{\mathbf{w}}) \equiv \mathbf{M}^{-1}\, \mathbf{b}, \tag{3.4}$$

is discretized in time using an explicit, strong stability preserving (SSP) Runge–Kutta scheme. For linear basis functions in space, we use a second order SSP Runge–Kutta scheme. Given a timestep Δt, and $t^n = n\Delta t$, $n = 0, 1, \ldots$, the method is defined as

$$\begin{aligned} \tilde{\mathbf{w}}^0 &= \tilde{\mathbf{w}}(t^n), \\ \tilde{\mathbf{w}}^l &= \tilde{\mathbf{w}}^{l-1} + \Delta t L_h(\tilde{\mathbf{w}}^{l-1}), \quad \text{for } l = 1, 2 \\ \tilde{\mathbf{w}}(t^{n+1}) &= \tfrac{1}{2}(\tilde{\mathbf{w}}^0 + \tilde{\mathbf{w}}^2). \end{aligned} \tag{3.5}$$

Other aspects of the DG implementation, such as slope limiting and wetting and drying, which are more specific to the shallow water application, are described in [21] and the references therein. Therefore we will not repeat them here except to say that in the numerical results below, we use a vertex-based slope limiter (the Bell–Dawson–Shubin limiter) as described in [1, 18], and the wetting and drying algorithm described in [4].

3.3. Local Timestepping (LTS). From here on, we will restrict our attention to piecewise linear approximations and second-order SSP Runge–Kutta timestepping. The LTS method that we employ is described in [9, 21], it is based on a simple modification of the second order SSP Runge–Kutta method described above, to allow for different timesteps in different regions, and to conserve mass.

To describe the method, consider the 1D (one space dimension plus time) scenario shown in Fig. 1. Here we have highlighted three elements, denoted by K_j, $j = i-1, i, i+1$, where $K_j = [x_{j-1/2}, x_{j+1/2}]$. Element K_{i-1} has the smallest timestep, which we label ΔT_1. Element K_i has timestep $\Delta T_2 = 2\Delta T_1$ and element K_{i+1} has the largest timestep $\Delta T_3 = 2\Delta T_2$. We focus on computing the solution in elements K_{i-1} and K_i, given a global solution at time t^n. Let $w_h^0 = w_h(t^n)$ and $w_h^{0,l} = w_h(t^n + l\Delta T_1)$. On K_{i-1}, we take two Euler steps:

$$\begin{aligned}\int_{K_{i-1}} w_h^{k,l} v_h dx &= \int_{K_{i-1}} w_h^{k-1,l} v_h dx \\ &-\Delta T_1 \left[\hat{F}\left(w_{h,-}^{k-1,l}, w_{h,+}^{0} \right) \Big|_{x_{i-1/2}} - \hat{F}\left(w_{h,-}^{k-1,l}, w_{h,+}^{k-1,l} \right) \Big|_{x_{i-3/2}} \right] \\ &+\Delta T_1 \int_{K_{i-1}} \left[F\left(w_h^{k-1,l} \right) v_h'(x) + s v_h \right] dx, \end{aligned} \tag{3.6}$$

for $k = 1, 2$. Then define

$$w_h^{0,l+1} = \frac{1}{2}\left[w_h^{0,l} + w_h^{2,l} \right].$$

We repeat this step for $l = 0, 1$ until we arrive at a solution $w_h^{0,2} = w_h(t^n + 2\Delta T_1)$ on element K_{i-1}. Note that we have kept the right state at $x_{i-1/2}$ fixed at the old time level during the calculation.

Now moving to element K_i, we again take two Euler steps:

$$\begin{aligned}\int_{K_i} w_h^{k} v_h dx &= \int_{K_i} w_h^{k-1} v_h dx \\ &-\Delta T_2 \left\{ \hat{F}\left(w_{h,-}^{k-1}, w_{h,+}^{0} \right) \Big|_{x_{i+1/2}} \right. \\ &\left. -\frac{1}{2}\left[\hat{F}\left(w_{h,-}^{k-1,0}, w_{h,+}^{0} \right) \Big|_{x_{i-1/2}} + \hat{F}\left(w_{h,-}^{k-1,1}, w_{h,+}^{0} \right) \Big|_{x_{i-1/2}} \right] \right\} \\ &+\Delta T_2 \int_{K_{i-1}} \left[F\left(w_h^{k-1} \right) v_h'(x) + s v_h \right] dx. \end{aligned} \tag{3.7}$$

for $k = 1, 2$. Then on K_i:

$$w_h(t^n + \Delta T_2) = \frac{1}{2}\left[w_h^0 + w_h^2 \right].$$

Note that the sum of all the fluxes at the interface $x_{i-1/2}$ between K_{i-1} and K_i is zero, insuring that we have global conservation of mass at the discrete time levels t^n, $t^{n+1}, \ldots$.

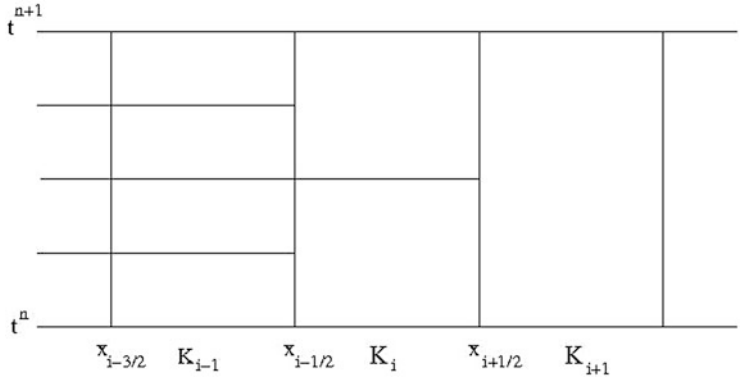


FIG. 1. Example of LTS on three elements with three different local timesteps in one space dimension

Next we go back to K_{i-1} and compute the solution up to time t^{n+1}, using the same procedure outlined above. Then we compute the solution on K_i up to time t^{n+1}, and finally the solution on K_{i+1} is computed.

3.4. Implementation and Parallelization. The LTS method described above can be generalized under the assumption that neighboring elements are assumed to have timesteps which differ by some integer M. We assume that each element K is placed into a timestepping group or level, where level 1 will denote the elements with the smallest timestep, level 2 the next smallest, and so forth. The total number of levels is denoted by $\bar{N}$.

We first calculate a local CFL timestep. On each element K, we compute the local timestep

$$\Delta t_K = \alpha \frac{\bar{h}_K}{\lambda_K} \tag{3.8}$$

where α is a CFL parameter which is $\mathcal{O}(1)$, typically $\alpha = 1/\sqrt{2}$, $\bar{h}_K$ is the minimum distance between the centroid of the element and the midpoint of the edges of K, and λ_K is an estimate of the maximum eigenvalue of the Jacobian associated with the normal flux $\hat{\mathbf{F}}$. As above, let ΔT_l, $l = 1, \ldots, \bar{N}$ denote timesteps associated with each timestepping level, where $M\Delta T_l = \Delta T_{l+1}$. We assume that

$$\Delta T_1 \leq \min_K \Delta t_K.$$

Then element K is placed into timestep group l if

$$\Delta T_l \leq \Delta t_K < \Delta T_{l+1}. \tag{3.9}$$

If $\Delta t_K \geq \Delta T_{\bar{N}}$, then element K is placed into level $\bar{N}$. The element timesteps are then reset, thus if element K is in group l, then $\Delta t_K \leftarrow \Delta T_l$.

For large-scale applications of interest, solutions cannot be computed in serial due to memory and CPU limitations, therefore parallel computing is necessary. We have implemented the LTS method in parallel with our shallow water model. The parallelization approach is based on domain decomposition, where the domain is first decomposed using the METIS software library [11, 12]. METIS divides the domain into overlapping subdomains with "ghost" regions based on a graph-partition of the nodes that make up the finite element mesh. In our implementation, the ghost region consists of elements which are shared by neighboring processors. MPI is used to pass solution information defined on the ghost elements to the neighboring processor. METIS attempts to divide the domain to balance the work-load among processors, to preserve locality of the elements and nodes within the subdomain and to minimize the "surface-to-volume" ratio; that is, to keep the ratio of ghost nodes to resident nodes low in order to reduce the communications overhead. For improved load balancing, METIS allows the user to weight nodes in the finite element mesh using an estimate of the "work" related to the node; for example, by estimating the maximum amount of work performed in elements which are attached to the node.

For a fixed global timestep, the parallelization of the DG method is quite straightforward. Each element has the same amount of work and takes the same timestep, and parallelization is achieved by each subdomain communicating with neighboring subdomains at the end of each Runge–Kutta timestep. The communication remains constant throughout the simulation. For LTS, the situation is much more complicated.

First, there is the question of load balancing. To march from time t^n to time t^{n+1}, the amount of work per element depends on the local timestep. We have attempted to address this in METIS by weighting each node by a factor which depends on the local timesteps associated with elements attached to the node. This factor is determined by the number of sub-cycling steps required for the element with the smallest timestep to go from time t^n to time t^{n+1}.

Second, there is the question of the timing of interprocessor communication. For example, consider the 1-D example in Fig. 1. There are $\bar{N} = 3$ timestepping levels with $M = 2$. Element K_{i-1} is on level 1, with the smallest timestep, element K_i is on level 2, and K_{i+1} is on level 3. Now assume element K_{i-1} is on processor 0 (PE0), and K_i and K_{i+1} are on the neighboring processor 1 (PE1), with elements K_{i-1} and K_i in the ghost region. Both processors PE0 and PE1 compute the solution on these two elements, but element K_{i-1} is "owned" by PE0 while element K_i is "owned" by PE1. For the solution to be computed correctly in the ghost region, information in element K_{i-1} must be passed from PE0 to PE1 at each level 1 timestep, and the information in element K_i must be passed from PE1 to PE0 at all level 2 timesteps.

In general, each timestepping level must communicate information with neighboring processors which share elements on the same level, if these elements are within the ghost region. Therefore, we have implemented a message-passing construct which is level-dependent. This may reduce parallel efficiency in the sense of strong scalability, since not all subdomains may have the same number of elements on each level, in fact some subdomains may have no elements on a given timestepping level. Or, subdomains may have elements within a level but none in the ghost region, while other subdomains may have many elements within a certain level in the ghost region, and thus require message-passing. One could try to address this problem by attempting to evenly divide the elements on each level among the processors; however, this approach would most likely destroy locality and result in a large number of isolated elements on each processor.

In summary, determining an optimal parallel strategy for LTS is complicated by several factors; however, as we will see in the results section below, LTS can still lead to an efficient and accurate approach in parallel as it does in serial.

4. Applications to the SWE. The eigenvalues of the normal flux for the SWE (with $S_p = 1$) are

$$\lambda_{1,2} = un_x + vn_y \pm \sqrt{gH}, \quad \lambda_3 = un_x + vn_y. \tag{4.1}$$

In shallow water simulations, one typically initializes the simulation by assuming a "cold-start"; i.e., water elevations are initially constant and water velocity is zero. Thus the largest eigenvalue initially is $\sqrt{gH}$, and the local timesteps are computed by

$$\Delta t_K = \alpha \frac{h_K}{\sqrt{gH_K}} \tag{4.2}$$

where H_K is the average water depth over the element. As the simulation progresses, the local timesteps may need to be adjusted based on the water velocity. In many cases $\sqrt{gH} >> |un_x + vn_y|$ and the local timesteps can be fixed during the computation. For more challenging applications, for example, modeling hurricane storm surges, this is not the case. Therefore, at certain intervals during the computation, we may recompute the local timesteps by

$$\Delta t_K = \alpha \frac{h_K}{\lambda_K} \tag{4.3}$$

where $\lambda_K = |\mathbf{u}_K| + \sqrt{gH_K}$. Here $|\mathbf{u}_K|$ is the magnitude of the cell average of velocity over the element K. The elements are then redistributed among the levels on each processor. That is, the number of levels $\bar{N}$ and the ratio M is left fixed, but elements are allowed to move between levels, depending on Δt_K.

4.1. Hurricane Ike Storm Surge Hindcast. In this section, we describe the application of LTS to a severe storm surge event, namely Hurricane Ike, which struck the upper Texas and Louisiana coasts in 2008.

The track of Ike is seen in Fig. 2. The storm progressed through the Western North Atlantic, through the Caribbean Sea making landfall in Cuba, and moved across the Gulf of Mexico, finally making a second landfall at Galveston, TX in the early morning of September 13, 2008. By this time, Ike had high category 2 winds but had an unusually large wind field and produced a category 4 storm surge in an area east of Houston, TX. In [8], we compared results computed using the RKDG method with no LTS to data taken from another model, namely the Advanced Circulation or ADCIRC code, which was used to study Hurricane Ike in [13]. Here we focus on computing results from this same study using no LTS and LTS (with a slight difference that we ignore eddy viscosity in the SWE, which is required to keep the ADCIRC model stable). The purpose of this exercise is to investigate the performance of the parallel DG code with LTS in this complex scenario and compare to results generated using the RKDG method described in [8] which did not use LTS.

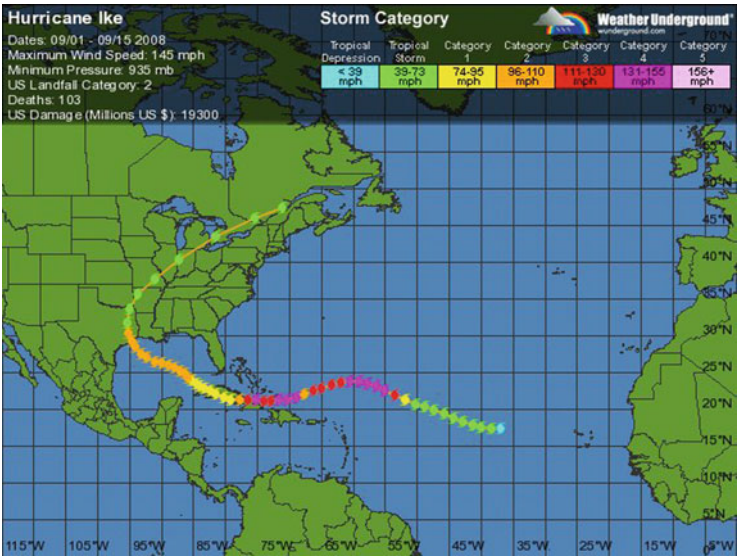


FIG. 2. Track of Hurricane Ike, taken from http://www.wunderground.com

The domain used in these simulations is the Western North Atlantic Ocean, Gulf of Mexico and Caribbean Sea, see Fig. 3. Here we also include most sections of the coast which are less than 50 feet above sea level, since these regions could be wetted in a storm event. The contours in the figure represent bathymetry measured in meters. In Fig. 4, we zoom in on the Galveston Bay region, the narrow channel in the figure is the Houston Ship

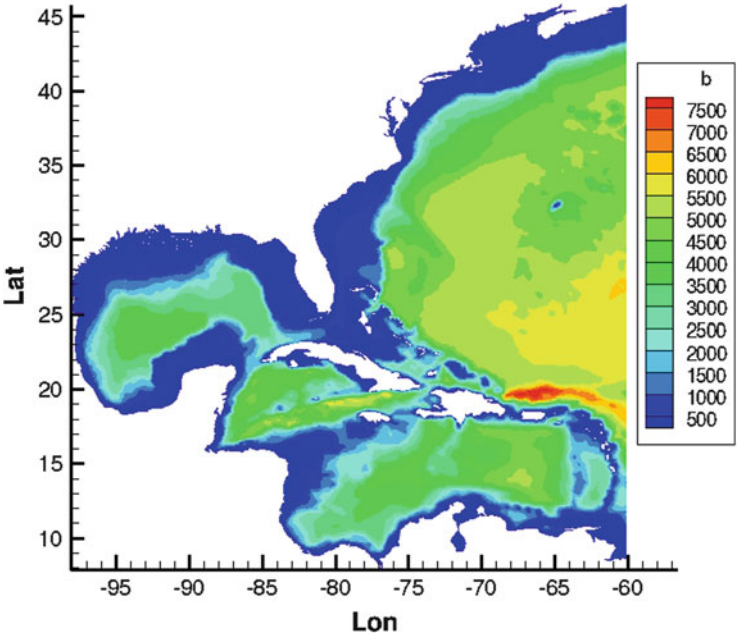


FIG. 3. Western North Atlantic/Texas domain with bathymetry (m)

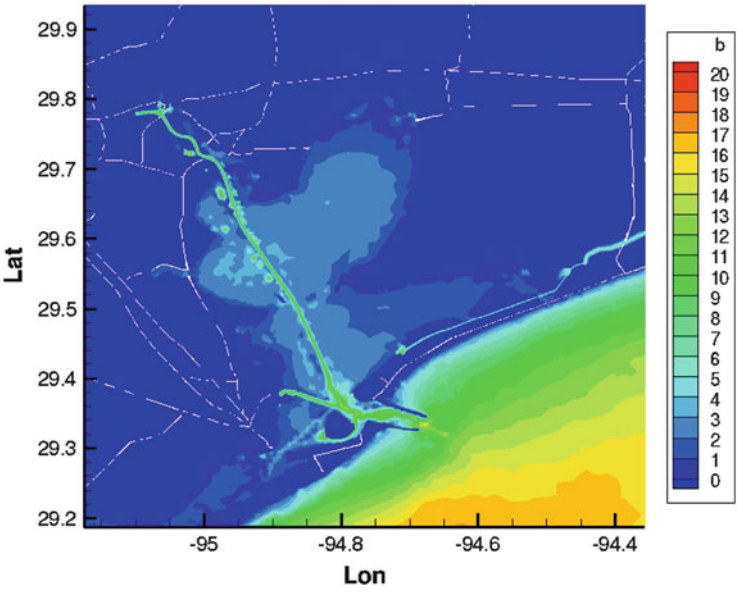


FIG. 4. Galveston bay with bathymetry (m)

Channel, which connects the Port of Houston to the Gulf of Mexico. The land regions shown in the figure are also included in the computational domain.

We present results of simulations of a 10-day period during the storm, beginning at 12:00 p.m. on September 5, 2008 and progressing through 12:00 p.m. on September 15, 2008. Hurricane winds were generated after the event by a data assimilated parametric wind model and were obtained from Ocean Weather, Inc.; see [13] for more details. The finite element mesh for these simulations consisted of 6,633,623 elements and 3,331,560 nodes, with most elements located in the Louisiana–Texas inland regions and continental shelf. The mesh is highly graded, with element areas on the order of several square kilometers in the deeper oceanic basins, transitioning to element areas on the order of 2,000 m^2 in the coastal regions of Texas and Louisiana. For simulations with no LTS, a global timestep of 0.5 s was used throughout the simulation. This was close to the minimum timestep computed using the CFL criteria (3.8) with velocity of zero.

An initial check of the local CFL constraints on each element revealed that only a small fraction of elements required the minimum CFL timestep of 0.5 s. The vast majority of elements had a local CFL timestep of 2 s or greater. The initial local CFL calculation revealed that 46,327 elements were in level 1 with a timestep of 0.5 s, 953,602 elements were in level 2 with a timestep of 1 s, and the remaining 5,633,694 elements could take a timestep of 2 s or larger. After testing several LTS scenarios, the most efficient was to use only three timestepping levels ($\bar{N} = 3$) and with $\bar{M} = 2$. The local CFL constraints were recomputed every 500 timesteps, and elements were redistributed among levels if needed; however, during the course of the simulation, only about 100,000 elements changed their timestep, mostly migrating from level 3 to level 2.

We ran simulations on the Ranger parallel computer at the Texas Advanced Computing Center*with 4,096 processing cores.

To compare the results of the LTS approach described above with no LTS, we look at two types of results, contours of maximum water elevation and hydrographs. The maximum water surface elevation is computed as

$$\eta_{\max}(x, y) = \max_{0 \le t \le T} \eta(x, y, t).$$

This quantity is of interest since it indicates where storm surge had the most impact over the course of the simulation. In Figs. 5 and 6, we compare the maximum water levels for no LTS and LTS over the impact area (the upper Texas coast extending to southeastern Louisiana). We also computed the difference between the two solutions in Fig. 7. Overall the agreement between the two solutions is remarkably close. There are a few small differences in the solutions in some isolated elements, primarily in regions which

*The Ranger system is comprised of 3,936 16-way SMP compute nodes providing 15,744 AMD Opteron processors for a total of 62,976 compute cores, 123 TB of total memory and 1.7 PB of raw global disk space. It has a theoretical peak performance of 579 TFLOPS. All Ranger nodes are interconnected using InfiniBand technology in a full-CLOS topology providing a 1GB/sec point-to-point bandwidth.

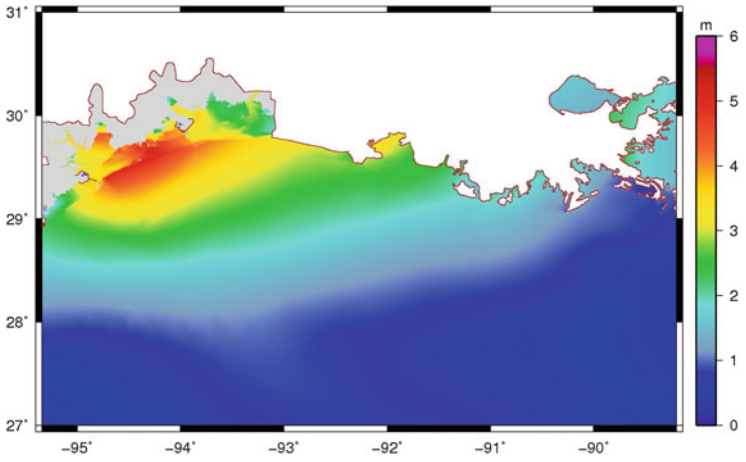


FIG. 5. Maximum water surface elevation in meters for Hurricane Ike hindcast, no LTS

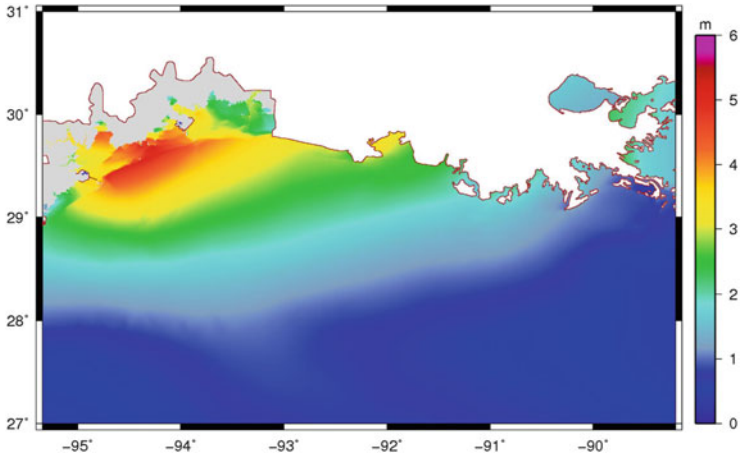


FIG. 6. Maximum water surface elevation in meters for Hurricane Ike hindcast with LTS

experience wetting and drying. These differences are most likely due to sensitivities in the wetting and drying algorithm used in the code.

We also compare hydrographs of solutions at three locations along the upper Texas coast, where actual instruments were deployed just before the storm, as described in [13]. These measurement locations are labeled as X, Y, and Z in Fig. 8 and are in the region of maximum storm surge. The LTS and no LTS solutions are plotted together in Fig. 9, where we observe that the solutions are virtually identical.

Finally, we remark on the CPU time of the simulations. The overall wall clock time for the 10-day simulation for no LTS was 1,058 min, and

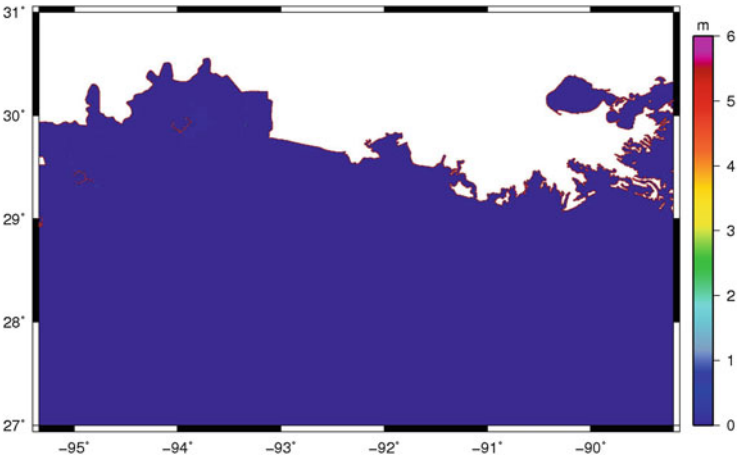


Fig. 7. Difference between LTS and no LTS

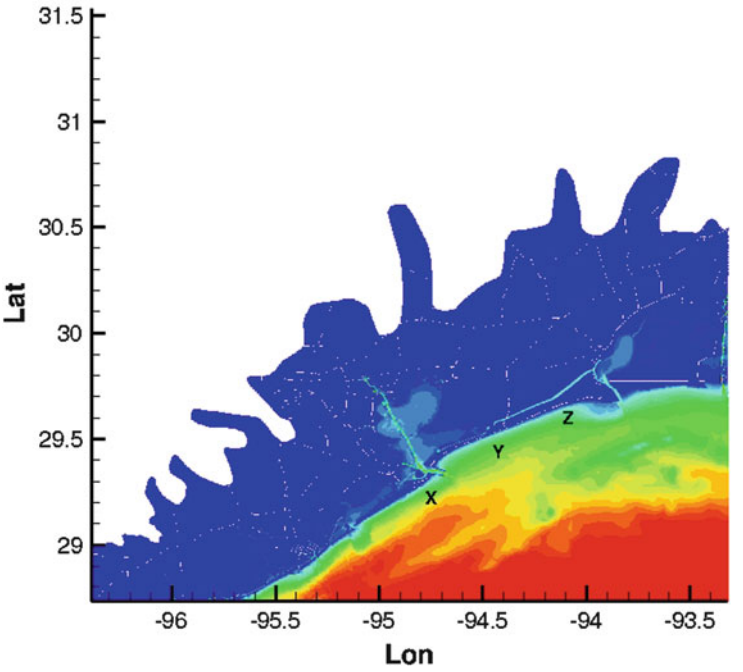


Fig. 8. Measurement locations X, Y, and Z

622 min with LTS, representing a 41 % decrease in wall clock time. While one might expect an even more dramatic decrease in wall clock time is possible, one must take into account the sequential nature of the local timestepping approach and the effect that this has on parallel efficiency.

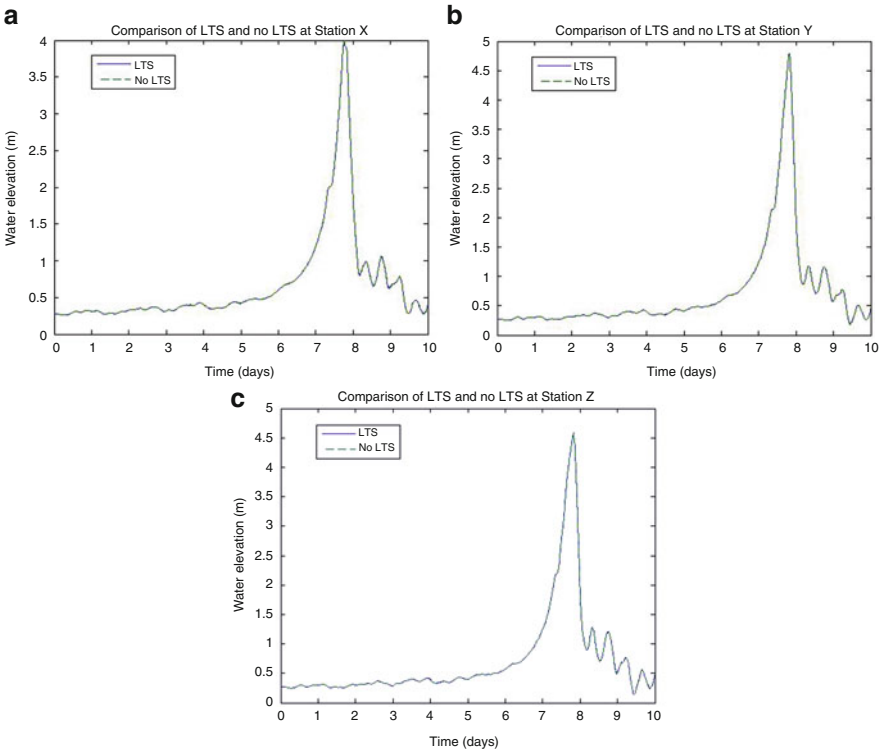


FIG. 9. Comparison of hydrographs for Hurricane Ike at measurement locations X, Y, and Z (**a**) Station X. (**b**) Station Y. (**c**) Station Z

5. Conclusion. In this paper, the LTS approach first described in [21] has been applied to a large-scale application in coastal ocean modeling, namely hurricane storm surge. The LTS method has proven to be robust and accurate even in these extreme events.

6. Acknowledgment. The author acknowledges the support of National Science Foundation grant DMS-0915118.

REFERENCES

[1] J. B. BELL, C. N. DAWSON, AND G. R. SHUBIN, *An unsplit, higher-order Godunov method for scalar conservation laws*, Journal of Computational Physics, 74 (1988), pp. 1–24.

[2] M. BERGER, D. L. GEORGE, R. J. LEVEQUE, AND K. T. MANDLI, *The GeoClaw software for depth-averaged flows with adaptive refinement*, Advances in Water Resources, (2011).

[3] M. BERGER AND R. J. LEVEQUE, *Adaptive mesh refinement using wave-propagation algorithms for hyperbolic systems*, SIAM Journal on Numerical Analysis, 35 (1998), pp. 2298–2316.

[4] S. Bunya, E. Kubatko, J. Westerink, and C. Dawson, *A wetting and drying treatment for the runge-kutta discontinuous galerkin solution to the shallow water equations.*, Computer Methods in Applied Mechanics and Engineering, 198 (2009), pp. 1548 – 1562.

[5] E. M. Constantinescu and A. Sandu, *Multirate timestepping methods for hyperbolic conservation laws*, Journal of Scientific Computing, 33 (2007), pp. 239–278.

[6] C. Dawson, *High resolution upwind-mixed finite element methods for advection-diffusion equations with variable time-stepping.*, Numer. Methods Partial Differential Equations, 11 (1995), pp. 525–538.

[7] C. Dawson and R. Kirby, *High resolution schemes for conservation laws with locally varying time steps*, SIAM J. Sci. Comput., 22 (2000), pp. 2256–2281.

[8] C. Dawson, E. Kubatko, J. Westerink, C. Trahan, C. Mirabito, C. Michoski, and N. Panda, *Discontinuous Galerkin methods for modeling hurricane storm surge*, Advances in Water Resources, doi:10.1016/j.advwatres.2010.11.004 (2010).

[9] C. Dawson, C. Trahan, E. Kubatko, and J. J. Westerink, *A parallel local timestepping Runge-Kutta Discontinuous Galerkin method with appliations to coastal ocean modeling*, submitted, (2012).

[10] J. Garratt, *Review of drag coefficients over oceans and continents*, Monthly Weather Review, 105 (1977), pp. 915–929.

[11] G. Karypis and V. Kumar, *METIS: A software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices*, University of Minnesota, Department of Computer Science/Army HPC Research Center, Minneapolis, MN, (1998).

[12] ———, *A fast and high quality scheme for partitioning irregular graphs*, SIAM Journal on Scientific Computing, 20 (1999), pp. 359–392.

[13] A. Kennedy, U. Gravois, B. Zachry, J. Westerink, M. Hope, J. Dietrich, M. Powell, A. Cox, J. R.A. Luettich, and R. Dean, *Origin of the hurricane ike forerunner surge*, Geophysical Research Letters, 38 (2011).

[14] R. Kirby, *On the convergence of high resolution methods with multiple time scales for hyperbolic conservation laws.*, Mathematics of Computation, 72 (2002), pp. 1239–1250.

[15] E. Kubatko, S. Bunya, C. Dawson, and J. Westerink, *Dynamic p-adaptive Runge-Kutta discontinuous Galerkin methods for the shallow water equations*, Computer Methods in Applied Mechanics and Engineering, 198 (2009), pp. 1766–1774.

[16] E. J. Kubatko, S. Bunya, C. Dawson, and J. J. Westerink, *A performance comparison of continuous and discontinuous finite element shallow water models*, Journal of Scientific Computing, 40 (2009), pp. 315–339.

[17] E. J. Kubatko, J. J. Westerink, and C. Dawson, *hp discontinuous Galerkin methods for advection dominated problems in shallow water flow*, Comput. Methods Appl. Mech. Engrg., 196 (2006), pp. 437–451.

[18] C. Michoski, C. Mirabito, C. Dawson, D. Wirasaet, E. J. Kubatko, and J. J. Westerink, *Adaptive hierarchi transformations over dynamic p-enriched schemes applied to generalized DG systems*, Journal of Computational Physics, 230 (2011), pp. 8028–8056.

[19] S. Osher and R. Sanders, *Numerical approximations to nonlinear conservation laws with locally varying time and space grids.*, Mathematics of Computation, 41 (1983), pp. 321 – 336.

[20] B. F. Sanders, *Integration of a shallow water model with a local time-step.*, Journal of Hydraulic Research, 46 (2008), pp. 466 – 475.

[21] C. J. Trahan and C. Dawson, *Local time-stepping in Runge-Kutta discontinuous Galerkin finite element methods applied to the shallow water equations*, Comput. Methods Appl. Mech. Engrg., 217–220 (2012), pp. 139–152.

[22] J. Westerink, R. Luettich, J. Feyen, J. Atkinson, C. Dawson, H. Roberts, M. Powell, J. Dunion, E. Kubatko, and H. Pourtaheri, *A basin to channel scale unstructured grid hurricane storm surge model applied to sourthern louisiana.*, American Meteorological Society, 136 (2008), pp. 833 – 864.

[23] D. Wirasaet, S. Tanaka, E. J. Kubatko, J. J. Westerink, and C. Dawson, *A performance comparison of nodal discontinuous Galerkin methods on triangles and quadrilaterals*, Internal Journal of Numerical Methods in Fluids, 64 (2010), pp. 1336–1362.

AN OVERVIEW OF THE DISCONTINUOUS PETROV GALERKIN METHOD

LESZEK F. DEMKOWICZ* AND JAY GOPALAKRISHNAN†

Abstract. We discuss our current understanding of the discontinuous Petrov Galerkin (DPG) Method with Optimal Test Functions and provide a literature review on the subject.

Key words. Discontinuous Petrov Galerkin, Optimal testing

AMS(MOS) subject classifications. 65N30, 35L15.

1. Introduction. The adventure with the discontinuous Petrov Galerkin (DPG) method started in Spring 2009. Analyzing spectral methods for the simplest 1D convection problem, we realized that the choice of test function $v = u$ leading to the standard DG method was far from an optimal one in terms of implied stability properties [23]. The main breakthrough came with a realization that the use of ultraweak variational formulation and discontinuous test functions allowed for the computation of (approximate) optimal test functions [25]. After reporting the exciting results in a Mafelap plenary talk, in June 2009, we had learned that we owned neither the concept of the ultraweak formulation nor even the name—the DPG method. Both were introduced several years earlier by the Italian colleagues [4, 5, 12, 13].*

But the concept of computing the optimal test functions on the fly was new, and we pursued a numerical implementation of *hp*-adaptivity quickly in [27] demonstrating the superior stability properties of the new method.

We devoted a considerable amount of our time and resources to the DPG research in the next three years. As it usually happens, our understanding did not grow in a systematic "monotone" mode and, hence, attempting to follow the DPG work in a chronological order would rather be confusing. Instead we present a review of the main concepts behind the DPG methodology as we understand them today: minimization of residuals in dual norms in Sect. 2, use of discontinuous test functions in Sect. 3, ultraweak variational formulations in Sect. 4, selection of optimal test norm for singular perturbation problems in Sect. 5, and the important interpretation of the DPG method as a localization of the PG method with

*Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX 78712, USA, leszek@ices.utexas.edu

†Department of Mathematics, Portland State University, Portland, OR 97207, USA, gjay@pdx.edu

*To our credit, the ultraweak formulation was used at that point very formally, without a proper Functional Analysis setting which we established later in [24].

X. Feng et al. (eds.), *Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations*, The IMA Volumes in Mathematics and its Applications 157, DOI 10.1007/978-3-319-01818-8_6,

global optimal test functions in Sect. 6. We conclude with an outline of our current work in Sect. 7.

The work on the DPG methodology has barely begun and we hope that more colleagues will get interested in the subject and join us in this endeavor.

2. DPG Is a Minimum Residual Method. DPG methods, like least squares methods, belong to the class of *minimum residual methods*. We start with a (linear) variational problem,

$$\begin{cases} u \in U \\ b(u,v) = l(v) \quad \forall v \in V \, . \end{cases} \tag{2.1}$$

Here U is a *trial space* and V is a *test space*. We shall assume that both U and V are Hilbert spaces, b is a sesquilinear (bilinear in the real case) and continuous form on $U \times V$, and l is an antilinear (linear) continuous form on V, i.e. an element of the dual space V'. It is well known that every such form $b(u,v)$ generates two linear operators, $B : U \to V', \quad B' : V \to U'$,

$$b(u,v) = \langle Bu, v \rangle_{V' \times V}, \quad \overline{b(u,v)} = \langle B'v, u \rangle_{U' \times U} \quad u \in U, v \in V \, . \tag{2.2}$$

As every Hilbert spaces is reflexive, i.e. it is isomorphic and isometric with its bidual, the two maps are actually conjugates of each other. The abstract variational problem (2.1) is equivalent to the operator equation:

$$Bu = l \, . \tag{2.3}$$

One might argue that the nature of variational problems lies in the fact that the corresponding operator takes values in a dual space.

Banach Closed Range Theorem[†] states that the following four conditions are equivalent to each other:

$$\begin{gathered} B \text{ has closed range,} \\ B' \text{ has closed range,} \\ B|_{\mathcal{N}(B)^\perp} \text{ is bounded below,} \\ B'|_{\mathcal{N}(B')^\perp} \text{ is bounded below.} \end{gathered}$$

Thus, at the expense of replacing U with the orthogonal complement of $\mathcal{N}(B)$, and V with the orthogonal complement of $\mathcal{N}(B')$, we can assume that both B and B' are bounded from below,

$$\|Bu\|_{V'} \geq \gamma \|u\|_U \quad \forall u \in U, \qquad \|B'v\|_{U'} \geq \gamma \|v\|_V \quad \forall v \in V \, . \tag{2.4}$$

Notice that the constant $\gamma = \|B^{-1}\| = \|(B')^{-1}\|$ is the same for both operators.

[†] See [44], p. 205, and [39], Thm. 5.18.2.

Let $U_h \subset U$ be now a finite-dimensional approximate trial space. The minimum residual method seeks a solution $u_h \in U_h$ that minimizes the corresponding residual:

$$u_h = \arg\min_{w_h \in U_h} J(w_h), \quad J(w_h) := \frac{1}{2}\|Bw_h - l\|^2_{V'} . \tag{2.5}$$

Of course, we have squared the norm of the residual and placed the half in front of it for elegance only. The use of the dual norm is a must, the operator takes values in the dual space V'. Problem (2.5) is equivalent to the minimization of the quadratic functional:

$$\frac{1}{2}(Bu_h, Bu_h)_{V'} - \mathrm{Re}(Bu_h, l)_{V'} \tag{2.6}$$

where $(\cdot,\cdot)_{V'}$ denotes the inner product in the dual space V'. Indeed, boundedess from below of B implies that the sesquilinear form:

$$(Bu_h, Bw_h)_{V'}$$

is U-coercive. Consequently, the minimum residual method can be classified as the classical Ritz method that experiences no preasymptotic behavior and delivers the best approximation error in the energy norm:

$$\|u\|_E := \|Bu\|_{V'} = \sup_{v\neq 0} \frac{|\langle Bu, v\rangle|}{\|v\|_V} = \sup_{v\neq 0} \frac{|b(u,v)|}{\|v\|_V} . \tag{2.7}$$

The dual norm, induced by the norm in test space,[‡]

$$\|l\|_{V'} = \sup_{v\neq 0} \frac{|l(v)|}{\|v\|_V} = \sup_{\|v\|_V \leq 1} |l(v)| = \sup_{\|v\|_V = 1} |l(v)| , \tag{2.8}$$

is not available analytically, unless we are dealing with the L^2-norm (possibly with a weight). We cannot thus compute directly with the dual norm. Coming to the rescue is the Riesz operator for the test space:

$$R_V \ : \ V \ni v \to (v,\cdot) \in V' , \tag{2.9}$$

which is an isometric isomorphism. At the expense of introducing the inverse of the Riesz operator, we can now trade the dual norm for the test norm and reformulate the minimum residual method (2.5) in a new form,

$$u_h = \arg\min_{w_h \in U_h} w_h \in U_h \tag{2.10}$$

Computing the Gâteaux derivative of the quadratic functional,

$$\langle \delta J(u_h); \delta u_h\rangle = \mathrm{Re}\,(R_V^{-1}(Bu_h - l), R_V^{-1} B\delta u_h)_V , \tag{2.11}$$

[‡]For Hilbert space, the supremum is attained and can be replaced with maximum.

we arrive at the linear problem equivalent[§] to minimization problem (2.10):

$$\begin{cases} u_h \in U_h \\ \left(R_V^{-1}\left(Bu_h - l\right), R_V^{-1} B\delta u_h\right)_V = 0 \quad \forall \delta u_h \in U_h \, . \end{cases} \tag{2.12}$$

There are two ways to proceed now.

Petrov–Galerkin Method with Optimal Test Functions. We introduce the *trial-to-test operator*:

$$T \, : U_h \to V, \quad T := R_V^{-1} B \, , \tag{2.13}$$

with the corresponding range $V_h := U_h$ identified as the *optimal test function space*. The linear problem (2.12) reduces to:

$$\left(R_V^{-1}\left(Bu_h - l\right), v_h\right)_V = 0 \quad \forall v_h \in V_h := TU_h \, . \tag{2.14}$$

Recalling the definition of Riesz operator, we can rewrite it in the variational form:

$$\begin{cases} u_h \in U_h \\ b(u_h, v_h) = l(v_h) \quad \forall v_h \in V_h \end{cases} \tag{2.15}$$

The minimum residual method is thus *equivalent* to a Petrov–Galerkin method with the optimal test functions. Computation of the optimal test functions involves inverting the Riesz operator,

$$\begin{cases} v_{\delta u_h} \in V \\ (v_{\delta u_h}, \delta v)_V = b(\delta u_h, \delta v) \quad \forall \delta v \in V \, , \end{cases} \tag{2.16}$$

and, unfortunately, it requires the solution of another boundary-value problem. Consequently, we have not got any practical method yet.

Being a minimum residual method,[¶] the PG method (2.15) delivers a hermitian, positive-definite stiffness matrix. Indeed, utilizing (2.16), we get:

$$b(u_h, v_{\delta u_h}) = (v_{u_h}, v_{\delta u_h})_V = \overline{(v_{\delta u_h}, v_{u_h})_V} = \overline{b(\delta u_h, v_{u_h})} \, . \tag{2.17}$$

The energy norm of the Galerkin error equals the residual and can be computed without knowing the exact solution,

$$\|u_h - u\|_E = \|B(u_h - u)\|_{V'} = \|Bu_h - l\|_{V'} = \|R_V^{-1}(Bu_h - l)\|_V \, . \tag{2.18}$$

[§] Functional $I(\delta u_h) := (R_V^{-1}(Bu_h - l), R_V^{-1} B\delta u_h)_V$ is antilinear. Real part of an antilinear functional vanishes if and only if the whole functional vanishes. This follows from the fact that, for any antilinear functional $I(v)$, $\mathrm{Im}\, I(v) = \mathrm{Re}\, I(iv)$.

[¶] One might say, a generalized least squares method.

We shall call $\psi := R_V^{-1}(Bu_h - l)$ the *error representation function*. Computing ψ involves solving the same variational problem as for the optimal test functions but with the residual on the right-hand side,

$$\begin{cases} \psi \in V \\ (\psi, \delta v)_V = b(u_h, \delta v) - l(\delta v) \quad \forall \delta v \in V \ . \end{cases} \tag{2.19}$$

Thus, the method comes with a "built-in" a-posteriori error estimation or, more precisely, a-posteriori *error evaluation*. Of course, all of these nice properties will be available if we come up with a practical method of inverting the Riesz operator.

A Mixed Formulation. Another way to proceed was proposed by Dahmen et al. in [20, 21]. Instead of identifying the second argument in (2.12) as the optimal test function, we identify the first argument as the error representation function,

$$(\psi, R_V^{-1} B \delta u_h)_V = 0 \quad \forall \delta u_h \ . \tag{2.20}$$

Taking out the Riesz operator and combining it with the definition of ψ, we obtain a saddle-point problem:

$$\begin{cases} \psi \in V,\ u_h \in U_h & & \\ (\psi, \delta v)_V - b(u_h, \delta v) & = -l(\delta v) & \forall \delta v \in V \\ b(\delta u_h, \psi) & = 0 & \forall \delta u_h \in U_h \ . \end{cases} \tag{2.21}$$

In this unconventional saddle-point problem, the approximate solution u_h comes from a finite-dimensional trial space and plays the role of the Lagrange multiplier for the error representation function.

Remark 2.1. *The PG scheme with optimal test functions was proposed in [23, 25]. It is perhaps interesting that we had arrived at the concept of optimal test functions from a completely different angle. Babuška's theorem [1] assures that the* discrete stability and approximability imply convergence. *More precisely, if $M := \|b\|$ is the continuity constant for the form $b(u, v)$ and the form satisfies the discrete inf-sup condition with constant γ_h,*

$$\sup_{v_h \in V_h} \frac{|b(u_h, v_h)|}{\|v_h\|_V} \geq \gamma_h \|u_h\|_U \ , \tag{2.22}$$

then the Galerkin error satisfies the estimate,

$$\|u_h - u\|_U \leq \frac{M}{\gamma_h} \inf_{w_h \in U_h} \|w_h - u\|_U \ . \tag{2.23}$$

The idea of optimal testing relies on employing test functions that realize *the supremum (maximum) in the discrete inf-sup condition (2.22). For*

a class of simple convection problems discussed in [23], such optimal test functions can be determined analytically. With such optimal test functions, the Petrov–Galerkin method inherits automatically the stability from the continuous level, i.e.

$$\gamma_h \geq \gamma \,, \tag{2.24}$$

where γ is the infinite-dimensional inf-sup constant. This holds for any possible trial space U. If we use the energy norm (2.7) in place of original norm $\|\cdot\|_U$, both the corresponding continuity and inf-sup constants are equal one, $M = \gamma = 1$. Babuška's estimate (2.23) implies then,

$$\|u_h - u\|_E \leq \frac{M}{\gamma_h}\|w_h - u\|_E \leq \frac{M}{\gamma}\|w_h - u\|_E = \|w_h - u\|_E \quad \forall w_h \in U_h \,. \tag{2.25}$$

Thus we have arrived at the minimum residual method. The moral of the story is that the minimum residual method is the most stable Petrov–Galerkin method *we can come up with.*

REMARK 2.2. *The most well-known minimum residual approach is the* Least Squares Method*, see the monograph of Bochev and Gunzburger [3]. Least squares are based on a strong operator setting with operator values in the L^2-space, and minimization of the L^2-residual. For the L^2 test space, Riesz operator reduces to an identity and there is no need for determining optimal test functions. The concept of minimizing the residual in dual norms is also not new, see, e.g., [6]. The novelty of the DPG method lies in the idea of computing the optimal test functions on the fly, made possible by use of broken test spaces and ultraweak variational formulation discussed in the next two sections, see also also Remark 3.2.*

3. Broken Test Spaces. As we have learned in Sect. 2, the Petrov Galerkin scheme with Optimal Test Functions requires inversion of the Riesz operator. With test norms involving standard exact sequence energy spaces, i.e. the use of H^1, $H(\mathrm{curl})$ and $H(\mathrm{div})$ inner products, inversion of the Riesz operator is equivalent to the solution of a separate boundary value problem. This would make the PG scheme unfeasible. Critical for the practicality of the method is the use of *broken* energy spaces typical for discontinuous Galerkin (DG) methods. Given a mesh $\mathcal{T}_h$ consisting of elements K, the corresponding broken energy spaces are defined as follows,

$$\begin{aligned} H^1(\Omega_h) &:= \{u \in L^2(\Omega) : u|_K \in H^1(K) \quad \forall K \in \mathcal{T}_h\}, \\ H(\mathrm{curl}, \Omega_h) &:= \{E \in (L^2(\Omega)^n : E|_K \in H(\mathrm{curl}, K) \quad \forall K \in \mathcal{T}_h\}, \\ H(\mathrm{div}, \Omega_h) &:= \{v \in (L^2(\Omega)^n : v|_K \in H(\mathrm{div}, K) \quad \forall K \in \mathcal{T}_h\}. \end{aligned} \tag{3.1}$$

The corresponding (standard) inner products are defined elementwise,

$$\begin{aligned}(u,\delta u)_{H^1(\Omega_h)} &:= \textstyle\sum_K (u|_K,\delta u|_K)_{H^1(K)},\\ (E,\delta E)_{H(\mathrm{curl},\Omega_h)} &:= \textstyle\sum_K (E|_K,\delta E|_K)_{H(\mathrm{curl},K)},\\ (v,\delta v)_{H(\mathrm{div},\Omega_h)} &:= \textstyle\sum_K (v|_K,\delta v|_K)_{H(\mathrm{div},K)}.\end{aligned} \tag{3.2}$$

While the definition of the broken energy spaces is unique, we use frequently other (equivalent) inner products defined on them. Inner products (3.2) are examples of *localizable inner products*, i.e. each element contribution defines an inner product on the corresponding element energy space. Not every standard energy inner product is localizable. For instance, the standard $H^1_0(\Omega)$ product:

$$(u,\delta u)_{H^1_0(\Omega)} = \int_\Omega \nabla u\,\overline{\nabla \delta u} = \sum_K \int_K \nabla u\,\overline{\nabla \delta u}\,, \tag{3.3}$$

is an inner product on $H^1_0(\Omega)$ but it is not longer definite on the corresponding broken energy space. Indeed, for an element K that is not adjacent to the boundary of Ω, $\int_K |u|^2 = 0$ implies only that u is a constant on K but not necessarily zero.

The main point in using the broken energy test spaces and localizable test norms is that the corresponding inversion of the Riesz operator *localizes*, i.e. it is done elementwise. Problem (2.16) decouples into independent element problems[‖]

$$\begin{cases} v_{\delta u_h} \in V(K) \\ (v_{\delta u_h},\delta v)_{V(K)} = b_K(\delta u_h,\delta v) \quad \forall \delta v \in V(K)\,. \end{cases} \tag{3.4}$$

Here b_K is the element contribution to the global sesquilinear form, and the left-hand side of the equation is the element contribution to the global test inner product. Function $\delta u_h = (\delta u_h)|_K$ denotes the restriction of a trial basis function to element K (element trial shape function), and $V(K)$ stands for the element test space.

Problem (3.4) is still infinite dimensional and it is equivalent to a boundary-value problem with Neumann (natural) boundary conditions. Except for simple problems (like advection with constant velocity vector [23]), we can only solve it approximately. As the inner product is hermitian and positive definite, the standard Bubnov–Galerkin method is a natural choice. We introduce an approximate element test space $\tilde V(K) \subset V(K)$ and seek *approximate optimal test functions*,

$$\begin{cases} \tilde v_{\delta u_h} \in \tilde V(K) \\ (\tilde v_{\delta u_h},\widetilde{\delta v})_{V(K)} = b_K(\delta u_h,\widetilde{\delta v}) \quad \forall \widetilde{\delta v} \in \tilde V(K)\,. \end{cases} \tag{3.5}$$

[‖] Note that the local problems are well defined by the assumption that the test norm is localizable.

This leads to the *approximate trial-to-test operator*:

$$\tilde{T} \,:\, U_h \ni \delta u_h \to \tilde{v}_{\delta u_h} \in \tilde{V}(K)\,, \tag{3.6}$$

and the corresponding *approximate optimal test space*:

$$\tilde{V}_h := \tilde{T} U_h\,. \tag{3.7}$$

The ultimate, *practical DPG method*, is obtained by replacing in (2.15) the optimal test functions with their approximate counterparts,

$$\left\{ \begin{array}{l} u_h \in U_h \\ b(u_h, \tilde{v}_h) = l(v_h) \quad \forall \tilde{v}_h \in \tilde{V}_h\,. \end{array} \right. \tag{3.8}$$

In practice, the approximate test space $\tilde{V}_h$ is obtained by raising locally the polynomial order of approximation. Roughly speaking, if the trial space involves polynomials of order p, we use polynomials of order $p + \Delta p$ for approximating the optimal test functions. Typically, $\Delta p = 2$. We might say that we are using the p-method for approximating element problem (3.5).

Example: Poisson Problem. We shall use the simplest example of the Poisson equation with Dirichlet boundary condition to illustrate the main points made so far. We seek $u \in H^1(\Omega)$ that satisfies the boundary-value problem:

$$\left\{ \begin{array}{ll} -\Delta u = f & \text{in } \Omega \\ u = u_0 & \text{on } \Gamma := \partial\Omega \end{array} \right. \tag{3.9}$$

where $f \in L^2(\Omega), u_0 \in H^{1/2}(\Omega)$ are given data.

Let $\mathcal{T}_h$ be a FE mesh. Take an element K, multiply both sides of Eq. $(3.9)_1$ with a test function v, integrate over element K, and integrate the left-hand side by parts, to obtain:

$$\int_K \nabla u \, \nabla v - \int_{\partial K} \frac{\partial u}{\partial n}\, v = \int_K f\, v\,. \tag{3.10}$$

Summing up over all elements K, we get:

$$\sum_K \int_K \nabla u \, \nabla v - \sum_K \int_{\partial K} \frac{\partial u}{\partial n}\, v = \sum_K \int_K f\, v\,. \tag{3.11}$$

The second term on the left-hand side represents jump terms and, for regular solution u, can be rewritten by summing up over all edges (faces) e in the mesh,

$$\sum_K \int_{\partial K} \frac{\partial u}{\partial n}\, v = \sum_e \frac{\partial u}{\partial n_e}[v]\,. \tag{3.12}$$

Here n_e is a predefined unit normal for edge e and $[v]$ represents the jump term:

$$[v](x) := \begin{cases} v & \text{if } e \subset \Gamma \\ \lim\limits_{\epsilon \to 0}(v(x+\epsilon n_e) - v(x-\epsilon n_e)) & \text{otherwise}\,. \end{cases} \tag{3.13}$$

We have several choices now.

1. Assume that test functions are globally conforming and vanish on Γ, $v \in H_0^1(\Omega)$. All boundary terms vanish and we arrive at the classical variational formulation:

$$\begin{cases} u \in H^1(\Omega), u = u_0 \text{ on } \Gamma \\ \displaystyle\int_\Omega \nabla u \, \nabla v = \int_\Omega f \, v \quad \forall v \in H_0^1(\Omega)\,. \end{cases} \tag{3.14}$$

2. We may assume that the test functions are globally conforming but do not necessarily vanish on Γ (we "do test" on Γ). The jump terms vanish but we are left with the normal derivative term on the domain boundary Γ:

$$\int_\Omega \nabla u \, \nabla v - \int_\Gamma \frac{\partial u}{\partial n} \, v = \int_\Omega f \, v\,. \tag{3.15}$$

   For $u \in H^1(\Omega)$, the normal derivative (the *flux*) is not well defined and we identify it as a new, separate unknown (placing the hat symbol over the normal derivative). The right energy setting is as follows:

$$\begin{cases} u \in H^1(\Omega), u = u_0 \text{ on } \Gamma, \ \widehat{\frac{\partial u}{\partial n}} \in H^{-1/2}(\Gamma) \\ \displaystyle\int_\Omega \nabla u \, \nabla v - \int_\Gamma \widehat{\frac{\partial u}{\partial n}} \, v = \int_\Omega f \, v \quad \forall v \in H^1(\Omega)\,. \end{cases} \tag{3.16}$$

3. We test with *discontinuous test functions*, i.e. $v \in H^1(\Omega_h)$. The jump terms remain. We introduce a new unknown, the *flux*

$$\widehat{\frac{\partial u}{\partial n}} \in H^{-1/2}(\Gamma_h)\,, \tag{3.17}$$

   that lives not only on Γ but the whole *mesh skeleton:*

$$\Gamma_h = \bigcup_K \partial K \tag{3.18}$$

   and, for regular solutions, coincides with the normal derivative of u. Space $H^{-1/2}(\Gamma_h)$, introduced in [24], is identified as the space of traces of functions from $H(\mathrm{div}, \Omega)$ to skeleton Γ_h, equipped with

the minimum energy extension norm. The ultimate variational formulation looks as follows:

$$\left\{ \begin{array}{l} u \in H^1(\Omega), u = u_0 \text{ on } \Gamma, \ \widehat{\frac{\partial u}{\partial n}} \in H^{-1/2}(\Gamma_h) \\ \displaystyle\int_\Omega \nabla u \, \nabla v - \sum_K \int_{\partial K} \widehat{\frac{\partial u}{\partial n}} \, v = \int_\Omega f \, v \quad \forall v \in H^1(\Omega_h) \, . \end{array} \right. \tag{3.19}$$

One can show that the last two problems are well posed [28].

With the discontinuous test functions, we can pursue now the idea of optimal testing. The price paid for the localization is the introduction of new unknown: the flux. The unknown solution is a group variable consisting of the original unknown u, and the flux $\widehat{\frac{\partial u}{\partial n}}$. A typical discretization uses standard H^1-conforming elements for u, and traces of Raviart–Thomas $H(\mathrm{div}, \Omega)$-conforming elements on mesh skeleton Γ_h for the flux. For instance, for 2D quadrilateral elements, the standard choice would be the isoparametric $\mathcal{P}^p \otimes \mathcal{P}^p$ element for u, and the discontinuous $\mathcal{P}^{p-1}$ element (with the Piola pull back map) for the flux. The lowest order elements use bilinear vertex shape functions for u and piecewise constant functions for the flux.

Let u be an H^1-conforming trial basis function. For each element K in the support of u, we solve for the optimal test function v_u,

$$\left\{ \begin{array}{l} v_u \in H^1(K) \\ \displaystyle\int_K \nabla v_u \, \nabla \delta v + v_u \, \delta v = \int_K \nabla u \, \nabla \delta v \quad \forall \delta v \in H^1(K) \, . \end{array} \right. \tag{3.20}$$

If u is discretized with $\mathcal{P}^p \otimes \mathcal{P}^p$ element, the corresponding optimal test function is approximated with $\mathcal{P}^{p+\Delta p} \otimes \mathcal{P}^{p+\Delta p}$ element (in practice, $\Delta p = 2$).

The support of v_u coincides with the support of u. If u is a vertex shape function, the corresponding optimal test function v_u will span over all elements sharing the vertex, if u represents an edge basis function, the support of v_u will include all elements sharing the edge, etc. In particular, if u is an element bubble, so is the corresponding test function v_u. Note though that, contrary to the continuous basis function u, the corresponding optimal test function (exact or approximate) v_u is *discontinuous*.

Similarly, let $g = g_e$ be a discontinuous flux function that lives on an edge (face in 3D) e. For each element K *adjacent* to the edge e, we solve for the optimal test function v_g,

$$\left\{ \begin{array}{l} v_g \in H^1(K) \\ \displaystyle\int_K \nabla v_g \, \nabla \delta v + v_g \, \delta v = - \int_{\partial K} g \, \mathrm{sgn}_K \, \delta v \quad \forall \delta v \in H^1(K) \, , \end{array} \right. \tag{3.21}$$

with

$$\text{sgn}_K = \begin{cases} 1 & \text{if } n_K = n_e \\ -1 & \text{if } n_K = -n_e \end{cases} \tag{3.22}$$

where n_e is the predefined edge normal and n_K is the outward normal for element K.

The support of v_g will span over all elements sharing the edge. We use the same $\mathcal{P}^{p+\Delta p} \otimes \mathcal{P}^{p+\Delta p}$ element to compute the approximate optimal test function.

REMARK 3.1. *It is important to notice that the DPG method does not destroy the classical logical flow of finite elements. In a classical PG FE method, we enter an element with sesquilinear (stiffness) and antilinear (load) forms, and two sets of approximating functions: trial and test shape functions. We integrate then for the element stiffness matrix and load vector that are returned to a global direct or iterative solver. In the DPG method, we enter an element with trial shape functions and* test inner product*. We compute then the approximate optimal test (shape) functions and proceed with the computation of corresponding element matrices. The "on the fly" computation of approximate optimal test functions takes place in the element routine and it does not affect the rest of the code.*

REMARK 3.2. *With the use of the* enriched *approximate test space, the logic is actually much simpler. We enter element K with the following forms: sesquilinear stiffness form $b_K(u,v)$, antilinear load $l_K(v)$ form, and sesquilinear $(v,\delta v)_K$ test inner product form. We have two sets of shape functions: trial shape functions e_i and enriched space basis functions $\hat{e}_j$. We compute the following matrices:*

$$\begin{aligned} K_{ki} &:= b_K(e_i, \hat{e}_k) \quad \text{"extended" element stiffness matrix,} \\ l_k &:= l_K(\hat{e}_k) \quad \text{"extended" element load vector,} \\ G_{kl} &:= (\hat{e}_k, \hat{e}_l)_V \quad \text{Gram test matrix.} \end{aligned} \tag{3.23}$$

We invert (factorize) the Gram matrix and compute the ultimate element stiffness matrix and load vector using the simple formulas:

$$G^{-1}_{kl} K_{ki} \bar{K}_{lj}, \quad G^{-1}_{kl} l_k \bar{K}_{lj} \,. \tag{3.24}$$

For the L^2 test inner product, we obtain the standard least squares method. The DPG method can thus be viewed as a preconditioned least squares method.

Convergence of the DPG method with exact optimal test functions is analyzed in [28]. We will discuss convergence analysis in context of general ultraweak variational formulations in more detail in Sect. 4. The effect of using the approximate optimal test functions was studied in [33].

The development of the DPG method was not motivated with the solution of simple elliptic problems like the Poisson equation for which

the standard Bubnov Galerkin method works just fine, and we do not necessarily advocate the use of DPG method for such problems. The price paid for the localization is high. If we neglect the cost of all local degrees-of-freedom (static condensation), in the standard H^1-conforming FE method we solve for *traces* of u on the mesh skeleton Γ_h. In the DPG method, we solve for *both* traces and fluxes, so the number of unknowns essentially doubles.

On the positive side, remember that the DPG method comes with a "built-in" a-posteriori error evaluator. Once the solution has been determined, a calc copy of the element routine is used to evaluate the (approximate) error representation function, and the corresponding element contribution to the global residual. With the use of adaptivity, the additional cost of solving for fluxes becomes less significant. The idea of using broken spaces and approximate inverse Riesz operators was used a long time ago in context of implicit a-posteriori error estimation, see, e.g., [38].

4. Ultraweak Variational Formulations. Whereas the localization requires only the test functions to be discontinuous, it is also desirable to work with a variational setting in which the trial functions are discontinuous as well. We shall first present such a formulation and only then discuss its advantages. As in the theory of Schwartz's distributions, the idea behind the *ultraweak variational formulation* is to move *all derivatives* to test functions.

We return to our model Poisson problem and rewrite it in terms of a system of first order equations:

$$\left\{ \begin{array}{rl} \sigma - \nabla u & = 0 \\ \mathrm{div}\sigma & = -f \, . \end{array} \right. \tag{4.1}$$

We will discuss first a *global formulation.* We multiply the two equations with *globally conforming* test functions $\tau \in H(\mathrm{div}, \Omega), v \in H^1(\Omega)$, integrate over domain Ω, and integrate by parts to obtain:

$$\left\{ \begin{array}{rl} \displaystyle\int_\Omega \sigma \cdot \tau + \int_\Omega u \, \mathrm{div}\tau - \int_\Gamma u \, v & = 0 \\ \displaystyle - \int_\Omega \sigma \, \nabla v + \int_\Gamma \sigma n \, v & = - \displaystyle\int_\Omega f \, v \, . \end{array} \right. \tag{4.2}$$

With no derivatives left on solution u, σ, the natural energy space for both components of the solution is the L^2 space: $\sigma \in (L^2(\Omega)^2, u \in L^2(\Omega)$. With such functional setting for the solution, the *trace* u and *flux* $t := \sigma n$ are not well defined and we declare them to be independent, additional unknowns $\hat{u} \in H^{1/2}(\Gamma)$ and $\hat{t} = \widehat{\sigma n} \in H^{-1/2}(\Gamma)$. Additionally, with Dirichlet boundary condition imposed on the whole boundary, the trace $\hat{u}$ is known from the boundary condition. We can substitute u_0 for $\hat{u}$ and move it to the right-hand side. This leads to the final formulation in the form:

$$\begin{cases} (u,\sigma,\hat{t}\,) \in L^2(\Omega) \times (L^2(\Omega))^2 \times H^{-1/2}(\Gamma) \\ (u,\mathrm{div}\tau) + (\sigma,\tau - \nabla v) + \langle \hat{t}, v\rangle = -(f,v) + \langle u_0, v\rangle \end{cases} \tag{4.3}$$

for all $v \in H^1(\Omega), \tau \in H(\mathrm{div},\Omega)$, where $(\cdot,\cdot)$ stands for the $L^2(\Omega)$-product and $\langle\cdot,\cdot\rangle$ for the duality pairing in $H^{-1/2}(\Gamma) \times H^{1/2}(\Gamma)$.

The bilinear form:

$$b((u,\sigma,\hat{t}),(v,\tau)) := (u,\mathrm{div}\tau) + (\sigma,\tau - \nabla v) + \langle \hat{t}, v\rangle \tag{4.4}$$

generates two operators: the ultraweak operator B defining the problem and its conjugate, *strong version* of the same operator (the problem is self-adjoint):

$$H^1(\Omega) \times H(\mathrm{div},\Omega) \ni (v,\tau) \to (\tau - \nabla v, \mathrm{div}\tau, v|_\Gamma) \in (L^2(\Omega))^2 \times L^2(\Omega) \times H^{\frac{1}{2}}(\Gamma)\,. \tag{4.5}$$

The ultraweak variational formulation can now be repeated in the DPG setting. Given a mesh $\mathcal{T}_h$ consisting of elements K, we repeat step (4.2) over every element K,

$$\begin{cases} \displaystyle\int_K \sigma\cdot\tau + \int_K u\,\mathrm{div}\tau - \int_\Gamma u\,v &= 0 & \tau \in H(\mathrm{div},K) \\ \displaystyle -\int_K \sigma\,\nabla v + \int_\Gamma \sigma n\, v &= -\int_K f\,v & v \in H^1(K)\,. \end{cases} \tag{4.6}$$

Next we sum up over all elements to obtain:

$$\sum_K \left\{ \int_K u\,\mathrm{div}\tau + \int_K \sigma\,(\tau - \nabla v) - \int_{\partial K} u\,v + \int_{\partial K} \underbrace{\sigma n}_{=:t}\, v \right\} = -\sum_K \int_K f v\,. \tag{4.7}$$

As in the global ultraweak formulation, we introduce additional unknowns: trace $\hat{u} \in H^{1/2}(\Gamma_h)$ and flux $\hat{t} \in H^{-1/2}(\Gamma_h)$. Both are defined not only on domain boundary Γ but *the whole mesh skeleton* Γ_h. Spaces $H^{1/2}(\Gamma_h)$ and $H^{-1/2}(\Gamma_h)$ are defined as traces of functions from $H^1(\Omega)$ and $H(\mathrm{div},\Omega)$ (see [24, 40] for a detailed discussion) and equipped with minimum energy extension norms:

$$\begin{aligned} \|\hat{u}\|_{H^{1/2}(\Gamma_h)} &:= \inf\left\{ \|u\|_{H^1(\Omega)} : u \in H^1(\Omega),\ u|_{\Gamma_h} = \hat{u} \right\}, \\ \|\hat{t}\|_{H^{-1/2}(\Gamma_h)} &:= \inf\left\{ \|\sigma\|_{H(\mathrm{div},\Omega)} : \sigma \in H(\mathrm{div},\Omega),\ (\sigma n)|_{\Gamma_h} = \hat{t} \right\}. \end{aligned} \tag{4.8}$$

The unknown trace $\hat{u}$ is represented as a sum of a lift of the known boundary data $\tilde{u}_0$ to the mesh skeleton Γ_h, and unknown component $\hat{u}$ that vanishes on Γ (watch for the overloaded symbol),

$$\hat{u} := \tilde{u}_0 + \hat{u}\,. \tag{4.9}$$

More precisely, the unknown component comes from the space of traces of $H_0^1(\Omega)$, denoted $\tilde{H}^{1/2}(\Gamma_h^0)$, where $\Gamma_h^0 := \Gamma_h - \Gamma$ stands for the *interior mesh skeleton*.

The ultimate *DPG ultraweak variational formulation* reads as follows:

$$\begin{cases} (u,\sigma,\hat{u},\hat{t}) \in L^2(\Omega) \times (L^2(\Omega))^2 \times \tilde{H}^{1/2}(\Gamma_h^0) \times H^{-1/2}(\Gamma_h) \\ \underbrace{(u, \mathrm{div}_h \tau) + (\sigma, \tau - \nabla_h v) - \langle \hat{u}, [\tau n] \rangle + \langle \hat{t}, [v] \rangle}_{=: b((u,\sigma,\hat{u},\hat{t}),(v,\tau))} = -(f,v) + \langle \tilde{u}_0, v \rangle \\ \forall v \in H^1(\Omega_h), \tau \in H(\mathrm{div}, \Omega_h) \, . \end{cases} \tag{4.10}$$

Above, the test functions come from the broken test spaces $H^1(\Omega_h), \tau \in H(\mathrm{div}, \Omega_h)$, the duality pairings extend over Γ_h^0 and Γ_h, respectively, and the grad and div operators in the first two terms are *understood elementwise* as indicated by index h. Similarly to the global ultraweak formulation, the conjugate operator generated by the bilinear form b corresponds to a strong version of the operator applied elementwise and accompanied by interface conditions across interelement boundaries expressing continuity of v and τn,

$$\begin{aligned} H^1(\Omega_h) \times H(\mathrm{div}, \Omega_h) \ni (v,\tau) \to (\mathrm{div}_h, \tau - \nabla_h v, [v], [\tau n]) \in L^2(\Omega) \\ \times (L^2(\Omega))^2 \times \Pi_K H^{\frac{1}{2}}(\partial K) \times \Pi_K H^{-\frac{1}{2}}(\partial K) \, . \end{aligned} \tag{4.11}$$

Note that the jump terms have to be understood globally, i.e.,

$$\begin{aligned} \langle \hat{u}, [\tau n] \rangle &:= \textstyle\sum_K \langle \hat{u}, \tau n \rangle_{\partial K} \\ \langle \hat{t}, [v] \rangle &:= \textstyle\sum_K \langle \hat{t}, v \rangle_{\partial K} \, . \end{aligned} \tag{4.12}$$

The DPG ultraweak formulation (4.10) provides a natural setting for the PG method with optimal test functions. The test functions are discontinuous enabling local computation of approximate optimal test functions. Solution consists of several components: the original unknown u, the (continuous) flux σ, traces $\hat{u}$, and fluxes $\hat{t}$. At a first glance, it looks like the formulation based on the first order system is much more expensive than the one based on the second order equation discussed in Sect. 3. Actually, it is not. The L^2-variables u, σ are discretized with discontinuous elements and can be statically condensed out. After the condensation, we solve a global problem for traces $\hat{u}$ and fluxes $\hat{t}$ whose cost is identical to the one discussed in the previous section. Thus, if we disregard the cost of local computations (that are trivially parallelizable), the two methods are essentially equally expensive.

Abstract Setting. The DPG ultraweak variational formulation discussed above has been applied to and analyzed for a number of different problems: convection-diffusion [24], linear elasticity [7], and linear acoustics [26]. In the latter, we began to see the emerging general abstract

setting. We further formulated it and applied to Stokes problem in [40]. We shall attempt now to outline the main points of the study in [40].

The starting point is an operator representing a system of first order differential equations:

$$(L^2(\Omega))^N \supset H_A(\Omega) \ni u \to Au \in (L^2(\Omega))^N \,. \tag{4.13}$$

Here N denotes the total number of scalar unknowns, and the domain of the operator is the graph space equipped with the graph norm:

$$\|u\|^2_{H_A} = \|u\|^2 + \|Au\|^2 \tag{4.14}$$

where, as usual, $\|\cdot\|$ denotes the L^2-norm.

Integration by parts leads to the introduction of the formal L^2-adjoint and a sesquilinear form representing boundary terms:

$$(Au, v) = (u, A^* v) + c(\mathrm{tr}_A u, \mathrm{tr}_{A^*} v) \tag{4.15}$$

Here v comes from the graph space for the formal adjoint:

$$H_{A^*}(\Omega) := \{v \in (L^2(\Omega))^N \ : \ A^* v \in (L^2(\Omega))^N\} \,, \tag{4.16}$$

equipped with a graph norm. We assume that we have at our disposal trace operators along with trace spaces for both energy spaces:

$$\begin{aligned} \mathrm{tr}_A : H_A(\Omega) \ni u \quad &\to \mathrm{tr}_A u = \hat{u} \in \hat{H}_A(\Gamma) \\ \mathrm{tr}_{A^*} : H_{A^*}(\Omega) \ni v \quad &\to \mathrm{tr}_{A^*} v = \hat{v} \in \hat{H}_{A^*}(\Gamma), \end{aligned} \tag{4.17}$$

and that $c(\hat{u}, \hat{v})$ is a definite sesquilinear form. For the Poisson problem, we have (watch for overloaded symbols):

$$
\begin{aligned}
u := (u,\sigma), \quad v = (v,\tau)\,,\\
Au = A(u,\sigma) = (\sigma - \nabla u, \mathrm{div}\sigma), \quad A^* = A,\\
H_A(\Omega) = H^1(\Omega)\times H(\mathrm{div},\Omega), \quad H_{A^*}(\Omega) = H_A(\Omega), \quad \mathrm{tr}_{A^*} = \mathrm{tr}_A,\\
\hat{H}_A(\Gamma) = H^{\frac{1}{2}}(\Gamma)\times H^{-\frac{1}{2}}(\Gamma), \quad \mathrm{tr}_A u = (u,\sigma n), \quad \hat{H}_{A^*}(\Gamma) = \hat{H}_A(\Gamma),\\
c(\hat{u},\hat{v}) = c((u,\sigma n),(v,\tau n)) = \langle u,\tau n\rangle_{H^{\frac{1}{2}}(\Gamma)\times H^{-\frac{1}{2}}(\Gamma)} + \langle \sigma n, v\rangle_{H^{-\frac{1}{2}}(\Gamma)\times H^{\frac{1}{2}}(\Gamma)}.
\end{aligned}
\tag{4.18}
$$

Let C be a boundary operator generated by the boundary sesquilinear form:

$$
C \,:\, \hat{H}_A(\Gamma) \to (\hat{H}_{A^*}(\Gamma))', \quad \langle C\hat{u},\hat{v}\rangle = c(\hat{u},\hat{v}). \tag{4.19}
$$

We assume that C can be split into two operators,

$$
C = C_1 + C_2, \tag{4.20}
$$

in such a way that operators A and A^* restricted to spaces corresponding to homogeneous boundary conditions:

$$
\begin{aligned}
U_0 &:= \{u \in H_A(\Omega) \,:\, C_1 u = 0\}\,,\\
V_0 &:= \{v \in H_{A^*}(\Omega) \,:\, C_2' v = 0\}\,,
\end{aligned}
\tag{4.21}
$$

are L^2-adjoint. According to Banach Closed Range Theorem, the following four conditions are equivalent to each other:

$A|_{U_0}$ has a closed range,

$A^*|_{V_0}$ has a closed range,

$A|_{U_0}$ is bounded below in the orthogonal component of its null space $\mathcal{N}(A|_{U_0})$,

$A^*|_{V_0}$ is bounded below in the orthogonal component of its null space $\mathcal{N}(A^*|_{V_0})$.

For simplicity, we assume that both $A|_{U_0}$ and $A^*|_{V_0}$ are injective. Consequently, both operators are bounded below with the same constant γ,

$$
\|Au\| \geq \gamma\|u\| \quad \forall u \in U_0 \quad \text{and} \quad \|A^* v\| \geq \gamma\|v\| \quad \forall v \in V_0\,. \tag{4.22}
$$

Operator split (4.20) implies a split of the trace space $\hat{H}_A$,

$$
\hat{H}_A = X_1 \oplus X_2, \quad X_1 = \mathcal{N}(C_2), \quad X_2 = \mathcal{N}(C_1)\,. \tag{4.23}
$$

For the Poisson problem,

$$
C_1(u,\sigma n) = u, \quad C_2(u,\sigma n) = \sigma n, \quad C_2'(v,\tau n) = v, \quad U_0 = V_0\,, \tag{4.24}
$$

i.e. the operator $A|_{U_0}$ is self-adjoint. The trace space split is very simple:

$$
X_1 = H^{1/2}(\Gamma)\times\{0\} \sim H^{1/2}(\Gamma), \quad X_2 = \{0\}\times H^{-1/2}(\Gamma) \sim H^{-1/2}(\Gamma)\,. \tag{4.25}
$$

We are finally ready to write down the *abstract ultraweak variational formulation* for the boundary-value problem:

$$Au = f, \quad C_1 u = g \tag{4.26}$$

Assuming that $g \in \mathcal{R}(C_1)$, we have,

$$\left\{ \begin{array}{l} u \in L^2(\Omega), \quad \hat{u}_2 \in X_2 \\ \underbrace{(u, A^* v) + < \hat{u}_2, C_2' \mathrm{tr}_{A^*} v >}_{=: b((u,\hat{u}_2),v)} = (f, v) - < g, \mathrm{tr}_{A^*} v > \quad \forall v \in H_{A^*}(\Omega) \, . \end{array} \right. \tag{4.27}$$

It has been proved in [24, 40] that problem (4.27) is well posed with an inf-sup constant of order[**] γ. The conjugate operator corresponding to form $b((u, \hat{u}_2), v)$ is the strong operator:

$$H_{A^*}(\Omega) \ni v \to (A^* v, C_2' \mathrm{tr}_{A^*} v) \in (L^2(\Omega))^N \times \hat{H}_{A^*}(\Gamma) \, . \tag{4.28}$$

We refer also to [40] for the abstract DPG ultraweak variational formulation. As for the Poisson problem, the unknowns include additionally the (abstract) trace variable $\hat{u}$ defined on the whole interior mesh skeleton. Notice that the abstract notion of "trace" includes equivalents of both trace and flux for the Poisson problem.

There are two main points about the results presented in [40], generalizing the earlier results for particular problems in [7, 24, 26]. First of all, we prove that, under the assumptions outlined above, the DPG ultraweak variational formulation is well posed with a *mesh independent* inf-sup constant of order γ. Mesh independence is critical for h-convergence and it is not obvious at all as, with h mesh refinements, the skeleton grows and there are "more" traces (and fluxes). The second important observation is that the inf-sup constant is of order γ. If, for a singular perturbation problem involving a parameter, γ is independent of the parameter, then this uniform stability *automatically carries over* to the DPG ultraweak variational formulation. As the PG method with optimal test functions inherits the inf-sup constant from the continuous level,[††] our DPG method is automatically *uniformly stable*, i.e. the approximation error is bounded by the best approximation error times a stability constant independent of the perturbation parameter. Using the abstract notation we have:

$$\left(\|u - u_h\|^2 + \|\hat{u} - \hat{u}_h\|^2 \right)^{1/2} \leq C \inf_{(w_h, \hat{w}_h)} \left(\|u - w_h\|^2 + \|\hat{u} - \hat{w}_h\|^2 \right)^{1/2} \, . \tag{4.29}$$

The abstract trace $\hat{u}$ incorporates both all (abstract) traces on the interior skeleton and the unknown traces on boundary Γ. The minimum extension energy norm used to measure traces is mesh dependent but the *field*

[**] Under the assumption that the traces spaces are equipped with minimum energy extension norms.

[††] Neglecting the error due to the approximation of optimal test functions.

variables u are measured in mesh-independent L^2-norm. Two particular cases are of interest: convection-dominated diffusion and linear acoustics. For linear acoustics, under appropriate regularity assumptions on the domain Ω, constant γ is independent of wave number k [26]. For convection-dominated diffusion with specific boundary conditions and assumptions on the advection vector, constant γ is independent of diffusion parameter ϵ. For both classes of problems, the uniform stability result (4.29) should be approached with care (for different reasons). The issue of *robustness* for singular perturbation problems will be discussed in Sect. 5.

For additional studies of DPG method based on the test graph norm, see [36, 37] (convection-dominated diffusion) and [11, 35] (thin walled structures), [30] (2D cloaking problems). A relation between various versions of DPG and DG methods was studied in [10].

For a related study on well-posedness of DPG formulations for general Friedrichs' systems, see [9].

Accounting for Approximation of Optimal Test Functions. A general theory for taking into account the approximation of optimal test functions was put forth in [33]. Let $\tilde{V}$ be the *enriched* approximate test space in which the optimal test functions are approximated. Suppose we can identify a Fortin-like operator,

$$\Pi : V \to \tilde{V} \quad \|\Pi v\| \leq C\|v\| \,, \tag{4.30}$$

that satisfies the orthogonality property:

$$b((u_h, \hat{u}_h), v - \Pi v) = (u_h, A^*(v - \Pi v)) + \langle \hat{u}_h, (v - \Pi v))\rangle = 0 \quad \forall u_h, \hat{u}_h \,. \tag{4.31}$$

We have then:

$$\begin{aligned} \sup_{v\in V} \frac{|b((u_h,\hat{u}_h),v)|}{\|v\|} &= \sup_{v\in V}\left[\frac{|b((u_h,\hat{u}_h),v-\Pi v)|}{\|v\|} + \frac{|b((u_h,\hat{u}_h),\Pi v)|}{\|v\|}\right] \\ &= \sup_{v\in V} \frac{|b((u_h,\hat{u}_h),Piv)|}{\|v\|}\frac{\|\Pi v\|}{\|v\|} \leq C \sup_{v\in \tilde{V}} \frac{|b((u_h,\hat{u}_h),v)|}{\|v\|} \\ &= C \sup_{v_h\in V_h} \frac{|b((u_h,\hat{u}_h),v_h)|}{\|v_h\|} \end{aligned} \tag{4.32}$$

where the last equality follows from the fact that the approximate optimal test functions realize the supremum in the enriched space. Thus, at the expense of the additional C stability factor, the practical DPG method preserves the optimal stability. For examples of such Fortin operators for Poisson and elasticity problems, see [33]. In practice, the orthogonality condition (4.31) serves as a defining property for such operators. The Fortin operators constructed in [33] are defined for arbitrary polynomial order p, but the estimates of constant C are p-dependent. The results provide thus a basis for h-convergence analysis only.

REMARK 4.1. *Common variational formulations are based on a partial relaxation only. Consider, for example, a system of linear elasticity equations consisting of equilibrium equations and Cauchy strain-displacement relations combined with Hooke's law, formulated in terms of displacements and stresses. We can choose to relax, i.e. integrate by parts, the equilibrium equations. Satisfying the remaining equations in the strong, pointwise (almost everywhere) sense, we can eliminate the unknown stresses and we arrive at the classical Principle of Virtual Work. If we keep the equilibrium equations in the strong form and relax the geometrical relations, we arrive at the mixed formulation. This time, the displacements cannot be eliminated unless we consider the time-harmonic case with nonzero frequency. In the "ultraweak" formulation,* both *sets of equations are relaxed i.e., in the spirit of Schwartz's distributions, all derivatives are passed to test functions. Besides the name and the general idea of a maximal relaxation, a relation with the ultraweak variational formulation of Després and Cessenat [14], and a subsequent work of Monk, Buffa, Huttunen, Perugia, Hiptmair and others, is rather loose.*

5. Robustness. By now, the reader should realize that the DPG method is more a methodology than a single method. One can combine the method with different variational formulations and, most importantly, one can compute with *different test norms*. For each test norm, we get a different version of the method. For "standard" problems and formulation based on the first order systems, the adjoint operator graph norm is a natural choice. For singular perturbation problems, where we strive for uniform (with respect to the perturbation parameter) approximation properties, the so-called *robustness*, the optimal choice of the test norm is much more difficult.

The very definition of what we mean by a *robust method* for a singular perturbation problem is shaky. Rather than attempting to develop a general theory, we will focus in this section on an important model problem: convection-dominated diffusion. The "confusion" problem, as we call it, is an important stepping stone for compressible and incompressible fluid dynamics. To fix ideas, we shall consider a model problem illustrated in Fig. 1. We shall start right away with the first order system setting. The problem of interest is:

$$\left\{ \begin{array}{c} \frac{1}{\epsilon}\sigma - \nabla u = 0 \\ \mathrm{div}(\sigma - \beta u) = -f \end{array} \right. \tag{5.1}$$

where ϵ is the diffusion parameter and $\beta = \beta(x) \approx O(1)$ is a prescribed advection field. We shall consider two types of boundary conditions:

$$\begin{array}{cl} \sigma_n - \beta_n u = -\beta_n u_0 & \text{on } \Gamma_{\text{in}} := \{x \in \Gamma : \ \beta_n(x) < 0\} \\ u = 0 & \text{on } \Gamma_{\text{out}} := \{x \in \Gamma : \ \beta_n(x) \geq 0\}\,. \end{array} \tag{5.2}$$

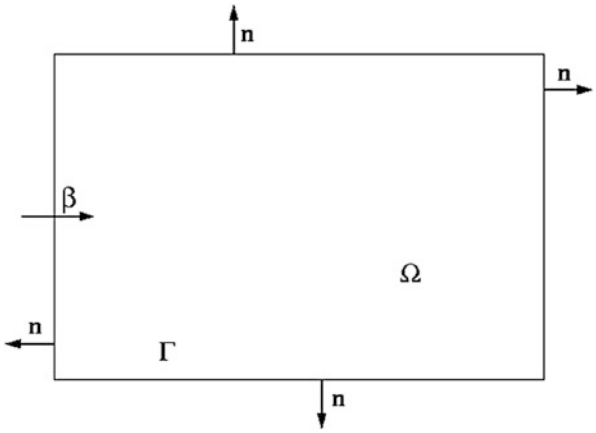


FIG. 1. A model convection-dominated diffusion problem

Here $\sigma_n = \sigma n$ and $\beta_n = \beta n$ represent normal components of flux σ and advection field β. The first boundary condition (BC) imposed on Γ represents an attempt at bringing from infinity condition $u = u_0$ where u_0 is a prescribed value. In presence of small diffusion ϵ, flux σ is expected to be small and the BC should approximate well the desired condition. The second BC is the main BC of interest as it produces a strong boundary layer on the outflow part of Γ.

Proceeding along lines discussed in Sect. 4, we obtain the following DPG ultraweak variational formulation for the problem:

$$\left\{\begin{array}{l} u \in L^2(\Omega),\ \sigma \in (L^2(\Omega))^2,\ \hat{u} \in \tilde{H}^{1/2}(\Gamma_h),\ \hat{t} \in \tilde{H}^{-1/2}(\Gamma_h) \\ \underbrace{(u, \mathrm{div}_h \tau + \beta \nabla_h v) + \left(\sigma, \frac{1}{\epsilon}\tau - \nabla_h v\right) - \langle \hat{u}, \tau_n \rangle + \langle \hat{t}, v \rangle}_{=:b((u,\sigma,\hat{u},\hat{t}),(v,\tau))} = -(f, v) + \langle \widetilde{\beta_n u_0}, v \rangle \\ \forall v \in H^1(\Omega_h),\ \tau \in H(\mathrm{div}, \Omega_h) \end{array}\right. \tag{5.3}$$

where $t = \sigma_n - \beta_n u$. Remember that all skeleton terms are global. Thus $\tilde{H}^{1/2}(\Gamma_h)$ are traces of functions from $H^1(\Omega)$, vanishing on Γ_{out}, and $\tilde{H}^{-1/2}(\Gamma_h)$ are traces of functions from $H(\mathrm{div}, \Omega)$ with (normal) trace vanishing on Γ_{in}. Finally, $\widetilde{\beta_n u_0}$ is an extension of $\beta_n u_0$ to the whole skeleton Γ_h.

If possible, we would like to have a robust behavior of L^2-error of the original unknown u. In other words, for a given mesh, we would like $\|u - u_h\|$ to be of the same order *uniformly* in $\epsilon \to 0$. The request is not so unreasonable, one can show that the L^2-norm of the solution u is bounded by the data f, u_0 *uniformly* in ϵ. This happens despite the fact that the solution develops a boundary layer on Γ_{out} which steepens up with $\epsilon \to 0$, the L^2-norm is simply insensitive to the developing boundary layer.

The critical question is:

How to define the test norm ?

The *ideal* DPG method delivers then the best approximation error in the energy norm (2.7). So one might want to solve an inverse problem: determine a test norm for which the corresponding energy norm coincides with the original norm used for the solution. This question can actually be answered at the abstract level. If both operators B and B' corresponding to the original sesquilinear form are injective, the ideal test norm is obtained by switching the role of spaces in the inf-sup condition (see [45] and Remark 5.2 below):

$$\|v\|_V := \sup_{u\in U, u\neq 0} \frac{|b(u,v)|}{\|u\|_U} \,. \tag{5.4}$$

The particular advantage of the ultraweak formulation (4.27) is that we can compute the ideal test norm explicitly:

$$\|v\|_V^2 = \|A^* v\|^2 + \left(\sup_{\hat{u}_2 \in X_2} \frac{|\langle \hat{u}_2, C_2' \mathrm{tr}_{A^*} v\rangle|}{\|\hat{u}_2\|_{\hat{H}_A}} \right)^2 \,. \tag{5.5}$$

For the model Poisson problem, we simply get:

$$\|(v,\tau)\|_V^2 = \|\tau - \nabla v\|^2 + \|\mathrm{div}\,\tau\|^2 + \|v|_\Gamma\|^2_{H^{1/2}(\Gamma)} \,. \tag{5.6}$$

The norm is very close to the adjoint operator graph norm used in our well-posedness analysis which prompted us to call in [26, 45] the graph norm a *quasi-optimal test norm.*

Remark 5.1. *A sesquilinear form $b(u,v)$, $u \in U, v \in V$ is called a* duality pairing *if*

$$\|u\|_U = \sup_{v\in V, v\neq 0} \frac{|b(u,v)|}{\|v\|_V} \quad \text{and} \quad \|v\|_V = \sup_{u\in U, u\neq 0} \frac{|b(u,v)|}{\|u\|_U} \,. \tag{5.7}$$

A class of problems with explicitly known norms $\|\cdot\|_U$, $\|\cdot\|_V$ for which the corresponding sesquilinear form is a duality pairing, was studied in [8].

We shall discuss now shortly a more general approach to the problem of determining an optimal test norm for singular perturbation problems that was proposed in [29]. Let us assume that we have found an optimal test norm $\|\cdot\|_V$. Consider a very special test function (v,τ) which, when substituted into the bilinear form (5.3) delivers the L^2-norm of solution u. This is obtained by requesting the conditions:

$$\begin{aligned} \mathrm{div}_h \tau + \beta \nabla_h v &= u \\ \tfrac{1}{\epsilon}\tau - \nabla_h v &= 0 \\ \langle \hat{u}, \tau_n\rangle &= 0 \quad \forall \hat{u} \\ \langle \hat{t}, v\rangle &= 0 \quad \forall \hat{t} \,. \end{aligned} \tag{5.8}$$

The last two conditions imply that the test function (v,τ) must be globally conforming, $v \in H^1(\Omega), \tau \in H(\mathrm{div},\Omega)$, and satisfy the homogeneous boundary conditions for the adjoint operator. This implies that the differential operators in the first two conditions can be understood globally. Simply, the test function (v,τ) solves the continuous adjoint problem with homogeneous BCs:

$$\left\{\begin{array}{l} v \in H^1(\Omega),\ \tau \in H(\mathrm{div},\Omega) \\ \frac{1}{\epsilon}\tau - \nabla v = 0, \quad \mathrm{div}\tau + \beta\nabla v = u \\ \tau_n = 0 \text{ on } \Gamma_{\mathrm{in}}, \quad v = 0 \text{ on } \Gamma_{\mathrm{out}}. \end{array}\right. \tag{5.9}$$

Notice that, contrary to the actual confusion problem, solution (v,τ) does not develop[††] a strong boundary layer on Γ_{in}, the outflow boundary for the adjoint problem.

We get:

$$\|u\|^2 = b((u,\sigma,\hat{u},\hat{t}),(v,\tau)) \leq \underbrace{\sup_{(v,\tau)} \frac{|b((u,\sigma,\hat{u},\hat{t}),(v,\tau))|}{\|(v,\tau)\|_V}}_{\|(u,\sigma,\hat{u},\hat{t})\|_E} \|(v,\tau)\|_V \,. \tag{5.10}$$

Thus, if the test norm has been selected in such a way that the solution of the adjoint problem (5.9) can be bounded by the data u *robustly*, i.e. *uniformly* in ϵ :

$$\|(v,\tau)\| \lesssim \|u\| \,, \tag{5.11}$$

the L^2-norm of the solution u is bounded robustly by the energy norm,

$$\|u\| \lesssim \|(u,\sigma,\hat{u},\hat{t})\|_E \,. \tag{5.12}$$

As the DPG method delivers the best approximation error in the energy norm, we obtain:

$$\begin{aligned} \|u-u_h\| &\lesssim \|(u-u_h,\sigma-\sigma_h,\hat{u}-\hat{u}_h,\hat{t}-\hat{t}_h)\|_E \\ &\leq \underbrace{\inf_{(u_h,\sigma_h,\tau u_h,\hat{t}_h)} \|(u-u_h,\sigma-\sigma_h,\hat{u}-\hat{u}_h,\hat{t}-\hat{t}_h)\|_E}_{\text{best approximation error in energy norm}} \,. \end{aligned} \tag{5.13}$$

The design of an optimal test norm leads thus to the stability analysis of the adjoint problem. Condition (5.12) is only necessary for the robustness. If the solution of the adjoint problem cannot be bounded in the test norm robustly by $\|u\|$, the robust estimate above is gone and the whole game is lost.

[††] Actually, BC $\tau_n = 0$ *does* produce a very weak boundary layer, hard to observe even with very accurate adaptive simulations, see [42].

The following stability estimates for the adjoint problem (5.9) have been proved in [18] (under some assumptions on advection β):

$$\|\beta\cdot\nabla v\|,\ \epsilon^{1/2}\|\nabla v\|,\ \|v\|,\ \frac{1}{\epsilon}\|\beta\cdot\boldsymbol{\tau}\|,\ \frac{1}{\epsilon^{1/2}}\|\boldsymbol{\tau}\|,\ \|\mathrm{div}\,\tau\|^2 \lesssim \|u\|\,. \tag{5.14}$$

The terms on the left are our "Lego blocks" to construct a test norm. In particular, the use of the graph norm:

$$\|(v,\tau)\|_V^2 = \|\frac{1}{\epsilon}\tau-\nabla v\|^2+\|\mathrm{div}\,v+\beta\cdot\nabla v\|^2+\|v\|^2+\|\tau\|^2 \tag{5.15}$$

is admissible. The main trouble with the graph norm and other possible test norms is that they inherit the main trouble of the original problem—optimal test functions may develop a boundary layer and, therefore, one may not be able to resolve them using the simple enrichment strategy. With unresolved optimal test functions, we cannot claim anymore the robust lower bound (5.14). In other words, our "Lego play" has to take into account another factor:

The optimal test functions should be easy to resolve.

The following mesh-dependent test norm has been studied extensively in [18]:

$$\|(v,\tau)\|_V^2 = \|\beta\cdot\nabla v\|^2+\epsilon\|\nabla\|^2+\|C_v\,v\|^2+\|\mathrm{div}\,\tau\|^2+\|C_\tau\,\tau\|^2 \tag{5.16}$$

with

$$C_v|_K = \min\left\{\sqrt{\frac{\epsilon}{|K|}},1\right\} \quad \text{and} \quad C_\tau|_K = \min\left\{\frac{1}{\sqrt{\epsilon}},\frac{1}{\sqrt{|K|}}\right\} \tag{5.17}$$

where $|K|$ denotes the element area. The zero order terms have been selected in such a way that they do not dominate the diffusion terms. Contrary to the graph norm, components v and τ have been separated, so the inversion of the Riesz operator can be done componentwise. The use of mesh-dependent zero order terms allows to employ the optimal powers of ϵ for small elements. Every time, we use suboptimal powers of ϵ in the construction of the test norm, we pay a price for that in the best approximation error estimate. With the mesh-dependent terms, we regain at least the optimality for small elements which appear in boundary layers and other places of high gradients. We refer once again to [18] for an extensive analysis and numerical experiments.

Finally, it is perhaps interesting to mention that, for general boundary conditions, it is impossible to avoid strong boundary layers in the solution of the adjoint problem. Robust estimates analogous to (5.16) are still possible but they employ *weighted* L^2-norms.[§§]The weighted test norms have been

[§§]Intuitively speaking, the weights are selected in such a way that they "kill" the effect of the boundary layers.

discovered in a purely experimental way in [27] and rediscovered through the theoretical analysis in [29].

An alternative to the presented philosophy is to work with the graph norm but employ special means (Shishkin meshes) to resolve the optimal test functions with boundary layers, see [36, 37].

REMARK 5.2 (*Connecting old dots...*). *A general approach for controlling convergence in a desired norm through the choice of optimal test functions was proposed a long time ago in [31, 32]. The idea is very simple. Let* $\|\cdot\|_U$ *be a trial norm in which we want the PG method to be optimal. Define optimal test functions* $v_h \in V$ *as follows:*

$$\begin{cases} v_h \in V \\ b(\Delta u, v_h) = (\Delta u, \delta u_h)_U \quad \forall \Delta u \in U \end{cases} \tag{5.18}$$

where $(\cdot,\cdot)_U$ *is the inner product corresponding to norm in* U*. The Galerkin orthogonality condition,*

$$b(u - u_h, v_h) = 0 \quad \forall v_h \,, \tag{5.19}$$

translates then [use $\Delta u := u - u_h$ *in (5.18)] into the corresponding orthogonality condition in terms of the inner product,*

$$(u - u_h, \delta u_h)_U = 0 \quad \forall \delta u_h \in U_h \,. \tag{5.20}$$

This implies that the PG scheme coincides with the orthogonal projection in the trial norm.

The optimal test space is given here by a different trial-to-test operator

$$U_h \ni \delta u_h \to (B')^{-1} R_U \delta u_h \in V \tag{5.21}$$

implied by the conjugate B' *and Riesz operator* R_U *for trial space* U*.*

For operators (5.21) and (2.13) to coincide, we need:

$$(B')^{-1} R_U = R_V^{-1} B \quad \Longleftrightarrow \quad R_V = B R_U^{-1} B' \,. \tag{5.22}$$

This implies the test norm:

$$\begin{aligned} \|v\|_V^2 &= (v,v)_V = \langle R_V v, v\rangle_{V'\times V} = \langle B R_U^{-1} v, v\rangle_{V'\times V} \\ &= \langle B'v, R_U^{-1} B' v\rangle_{U'\times U} = (R_U^{-1} B' v, R_U^{-1} B' v)_U \\ &= \|R_U^{-1} B' v\|_U^2 = \|B'\|_{U'}^2, \end{aligned} \tag{5.23}$$

which coincides with norm (5.4).

Note also that the ideal test norm is not readily available unless we can invert the Riesz operator R_U *explicitly. This is the case of the* L^2*-norm used in the ultraweak formulations but it is not available if the* U*-norm includes derivatives. This may be considered to be an additional advantage of the ultraweak formulations.*

6. Global Optimal Test Functions. In this section we discuss an alternate interpretation of the DPG method based on the concept of global optimal test functions.

Suppose that we pursue the PG method with optimal test functions for the global ultraweak variational formulation (4.27). In other words, for each trial function $(u_h, \hat{u}_{2,h})$, we determine the corresponding *globally optimal test functions* by solving the global problem:

$$\begin{cases} v \in H_{A^*}(\Omega) \\ (v, \delta v)_V = (u_h, A^* \delta v) + \langle \hat{u}_{2,h}, C_2' \mathrm{tr} A^* \delta v \rangle \quad \forall \delta v \in H_{A^*}(\Omega) \, . \end{cases} \tag{6.1}$$

The relation between the DPG (local) *approximate* optimal test functions and the global optimal test functions is quite revealing.

In order to study a relation between the approximate local and global test functions, we will resort to the following abstract notation for the sesquilinear form:

$$b((w, \hat{w}), v) = b_0(w, v) + \langle\langle \hat{w}, v \rangle\rangle. \tag{6.2}$$

By w above we mean the group unknown corresponding to the ultra-weak formulation with globally conforming test functions (6.1). It consists of field variable $u \in L^2(\Omega)$ and the unknown part $\hat{u}_2$ of trace on Γ. The second variable $\hat{w}$ denotes the unknown abstract trace defined on the *internal* skeleton only. For conforming test functions, the second term vanishes.

Let $W_p, \hat{W}_p$ be now appropriate discrete spaces for the two sets of variables, and let $V_r(\Omega_h)$ denote the enriched approximate broken test space $(r = p + \Delta p)$ used to determine the *practical trial-to-test operator:*

$$\begin{cases} T^r(w, \hat{w}) \in V_r(\Omega_h) \\ (T^r(w, \hat{w}), \delta v)_V = b((w, \hat{w}), \delta v) \quad \forall \delta v \in V_r(\Omega_h) \, . \end{cases} \tag{6.3}$$

The *local optimal test space* of the practical DPG method is

$$V_r^p(\Omega_h) := T^r(W_p \times \hat{W}_p) \, . \tag{6.4}$$

To study its relation with a weakly conforming approximate test space, define

$$\tilde{V}_r(\Omega_h) := \{ v \in V_r(\Omega_h) \, : \, \langle\langle \hat{w}_p, v \rangle\rangle = 0 \quad \forall \hat{w} \in \hat{W}_p \} \, . \tag{6.5}$$

Then, let $\tilde{T}^r w$ in $\tilde{V}_r(\Omega_h)$ be defined by

$$(\tilde{T}^r w, \delta v)_V = b_0(w, \delta v) \quad \forall \delta v \in \tilde{V}_r(\Omega_h) \, . \tag{6.6}$$

The *weakly conforming global optimal test space* is defined by

$$\tilde{V}_r^p(\Omega_h) := \tilde{T}^r(W_p) \, . \tag{6.7}$$

We have (comp. also [27]).

PROPOSITION 6.1.

$$\tilde{V}_r^p(\Omega_h) \subset V_p^r(\Omega_h)\,. \tag{6.8}$$

Proof. Since $V_r^p(\Omega_h) \subset V_r(\Omega_h)$, we have $V_r(\Omega_h) = V_r^p(\Omega_h) + V_r^\perp(\Omega_h)$ where $V_r^\perp(\Omega_h)$ is the V-orthogonal component of $V_r^p(\Omega_h)$ in $V_r(\Omega_h)$.

Let $\tilde{v} = \tilde{T}^r w_p$. Since $\tilde{v}$ is in $V_r(\Omega_h)$, we can apply the decomposition above to get

$$\tilde{v} = v^p + \tilde{v}^\perp, \quad v^p \in V_r^p(\Omega_h),\ \tilde{v}^\perp \in V_r^\perp(\Omega_h)\,. \tag{6.9}$$

Since $\tilde{v}^\perp$ is V-orthogonal to $V_r^p(\Omega_h)$, for every $\hat{w}_p \in \hat{W}_p$, we have

$$0 = (T^r(0,\hat{w}_p), \tilde{v}^\perp)_V = b((0,\hat{w}_p), \tilde{v}^\perp) = \langle\langle \hat{w}_p, \tilde{v}^\perp \rangle\rangle\,, \tag{6.10}$$

hence $\tilde{v}^\perp \in \tilde{V}_r(\Omega_h)$. Therefore, we may substitute $\tilde{v}^\perp$ for v in (6.6) to get

$$(\tilde{T}^r w_p, \tilde{v}^\perp)_V = b_0(w_p, \tilde{v}^\perp)\,. \tag{6.11}$$

Now, the right-hand side above must vanish because

$$b_0(w_p, \tilde{v}^\perp) = ((w_p,0), \tilde{v}^\perp) = (T^r(w_p,0), \tilde{v}^\perp)_V = 0\,. \tag{6.12}$$

Therefore, $0 = (\tilde{T}^r w_p, \tilde{v}^\perp)_V = (\tilde{v}, \tilde{v}^\perp)_V = \|\tilde{v}^\perp\|_V$. Returning to (6.9), we find that $\tilde{v} = v^p \in V_r^p(\Omega_h)$, thus finishing the proof. □

Now consider two DPG methods corresponding to the two test spaces defined previously. The first is the standard practical DPG method that defines $w_p \in W_p$ and $\hat{w}_p \in \hat{W}_p$ satisfying

$$b((w_p,\hat{w}_p), v) = l(v) \quad \forall v \in V_r^p(\Omega_h)\,. \tag{6.13}$$

The second is the DPG method with the weakly conforming globally optimal test space, which finds $\tilde{w}_p \in W_p$ satisfying

$$b_0(\tilde{w}_p, \tilde{v}) = l(\tilde{v}) \quad \forall \tilde{v} \in \tilde{V}_r^p(\Omega_h) \tag{6.14}$$

PROPOSITION 6.2. *If both methods are uniquely solvable, then*

$$w_p = \tilde{w}_p\,. \tag{6.15}$$

Proof. By virtue of Proposition 6.1, we can substitute any $\tilde{v} \in \tilde{V}_r^p(\Omega_h)$ into (6.13), which then immediately reduces to (6.14). □

The moral of the story is that the actual DPG method may be interpreted simply as a *localization* of the corresponding global PG methodology. This fact has a number of important implications.

The most important one is the fact that we do not have to resolve the element local DPG optimal test functions but only element restrictions of

global optimal test functions. For instance, if we insist on using the graph norm for the confusion problem discussed in Sect. 5, the global optimal test functions *do not form (strong) boundary layers*. If we trade the flux BC for the BC on u, the global optimal test functions do develop a strong layer on the inflow boundary. Consequently, if we can resolve (by whatever means) optimal test functions adjacent to the inflow boundary (and use the standard enriched spaces with $\Delta p = 2$ elsewhere), we observe a robust behavior of the method. On the contrary, if we go after the DPG (local) optimal test functions, we need to resolve boundary layers within each element. For concrete examples illustrating the discussion, see [15].

From the analysis point of view, it looks like one can deemphasize convergence of traces (and fluxes) and focus on studying the convergence of the field variables only. In particular, a discretization of traces with discontinuous elements is non-conforming form the point of view of the DPG method (traces live in $H^{1/2}$ space) but it is perfectly OK from the point of view of the global PG method and non-conforming discretization of optimal test functions.

As the approximation of optimal test functions in non-conforming, one has to account for both approximation and consistency errors. For 1D problems, weak conformity is equivalent to conformity and there is no consistency error. The DPG method delivers optimal convergence in the L^2-error. This explains, in particular, why the 1D DPG method for linear acoustics is pollution-free [45].

7. Generalizations and Conclusions. The DPG method is a minimum residual method minimizing the residual in a dual norm. It generalizes the classical least squares approach based on L^2-residuals. The general philosophy is straightforward—we minimize the residual in hope of the corresponding error (measured in a desired norm) converging to zero as well. It is thus clear from the very beginning that the choice of the norm used to measure the residual is critical. We need what Wolfgang Dahmen calls "the right mapping property." This leads to the task of selecting the right test norm. The issue is especially critical for singular perturbation problems.

In the paper, we attempted to review the main ideas behind the DPG method and review our and our collaborators' work in the last three and a half years, since the inception of the main idea of computing (approximate) optimal test functions on the fly [23, 25]. Despite several success stories, the method is still in its infancy. We hope that this overview will help to propagate an interest in the DPG methodology.

We will finish the paper with a few additional comments on current work and open research problems.

Adaptivity. Once the resolution of optimal test functions is secured, the DPG methods guarantee stability for any well-posed linear problem and *any discretization*, in particular *hp* elements. This promises superior

convergence rates for problems with singular solutions and boundary layers. The method comes with *a built-in error evaluator*. We have made the point of using higher order elements, *h*- and *hp*-adaptive meshes in most of our papers to demonstrate that the method remains stable and delivers optimal approximation properties for arbitrary *hp* meshes. Whereas the element contributions to the global residual serve as perfect element error indicators, the subject of automatic *hp*-adaptivity remains completely open. In all our examples of *hp*-adaptivity, we have used so far simple marking strategies only.

DPG for Nonlinear Problems. The idea of minimizing the dual residual can be extended to nonlinear problems. This was first pursued in [34]. If operator B in (2.10) is nonlinear, the corresponding formula (2.11) for the first Gâteaux derivative must be modified:

$$\langle \delta J(u_h); \delta u_h \rangle = \text{Re } (R_V^{-1} \underbrace{(Bu_h - l)}_{=:r_h}, R_V^{-1} B'(u_h; \delta u_h))_V \, . \tag{7.1}$$

Above, $B'(u_h; \delta u_h)$ denotes the first derivative of operator B at u_h in the direction of δu_h, and r_h is the residual.

The second derivative (Hessian) consists of two terms:

$$\begin{aligned} \langle \delta^2 J(u_h); \delta u_h, \Delta u_h \rangle = \text{Re } \big[&(R_V^{-1}(B'(u_h; \Delta u_h)), R_V^{-1} B'(u_h; \delta u_h))_V \\ &+ (R_V^{-1} r_h, R_V^{-1} B''(u_h; \delta u_h, \Delta u_h))_V \big] \, . \end{aligned} \tag{7.2}$$

The trial-to-test operator depends now upon u_h:

$$v_{\delta u_h} = T(u_h)\delta u_h = R_V^{-1} B'(u_h; \delta u_h) \, . \tag{7.3}$$

Introducing the optimal test functions into the formulas, we obtain:

$$\begin{aligned} \langle \delta J(u_h); \delta u_h \rangle &= \text{Re } \big[b(u_h, T(u_h)\delta u_h) - l(T(u_h)\delta u_h) \big] \\ \langle \delta^2 J(u_h); \delta u_h, \Delta u_h \rangle &= \text{Re } \big[(T(u_h)\Delta u_h, T(u_h)\delta u_h)_V \\ &\qquad + \langle B''(u_h; \delta u_h, \Delta u_h), R_V^{-1} r_h \rangle \big] \, . \end{aligned} \tag{7.4}$$

With the elementwise inversion of the Riesz operator, we can compute not only the gradient but also the hessian of the residual and use it to solve the nonlinear minimum residual problem. Note that the formula for the hessian assumes that the test inner product is fixed. In practice, an optimal inner product depends upon the first derivative and it also evolves with u_h.

For preliminary attempts to apply the DPG method to nonlinear problems, see also [16].

Element Conservation Properties. In general, the DPG method does not assure element conservation properties. For problems involving conservation laws, enforcement of conservation properties is deemed desirable. For instance, in the convection-dominated diffusion, the second equation

in (5.1) represents a conservation law. The element conservation property translates into the requirement that the test space includes functions $(v,\tau)=(1_K,0)$, where 1_K denotes indicator function for element K. There is no reason why, in general, the optimal test space should include such functions.

It was though also noticed by the MIT colleagues [34] that it is relatively easy to enforce the element conservation property. The main idea is to turn the residual minimization problem (2.10) into a *constrained minimization problem*. We form the Lagrangian:

$$L(u_h,\lambda_K):=\frac{1}{2}\|R_V^{-1}(Bu_h-l)\|_V^2-\sum_K \text{Re}\ \left[b(u_h,\lambda_K(1_K,0))-l(\lambda_K(1_K,0))\right] \tag{7.5}$$

and seek its stationary points. We arrive at the mixed problem:

$$\left\{\begin{array}{lll} b(u_h,T\delta u_h)-\sum_K \lambda_K b(\delta u_h,(1_K,0)) & =l(T\delta u_h) & \forall \delta u_h \\ b(u_h,(1_K,0)) & =l((1_K,0)) & \forall K\ . \end{array}\right. \tag{7.6}$$

Thus, at the expense of introducing an extra scalar unknown per element for each conserved quantity, the minimum residual approach allows naturally for enforcing conservation laws at the element level, a critical property in CFD simulations. We refer to [17] for additional details.

Preconditioning. Numerical tests indicate that the DPG method based on the ultraweak formulation delivers stiffness matrices with same condition numbers as conforming finite elements. The work on preconditioners and iterative solvers for the DPG method is in its infancy, see [2] for the only results we are aware of at this point.

Maxwell Problems. For an application of DPG methodology to a 2D Maxwell cloaking problem, see [30]. A theoretical groundwork for the 3D DPG Maxwell method has recently been laid down by Wieners and Wohlmut [43].

Implementation of DPG Method. We conclude with a short discussion on implementational issues.

The DPG method is definitely more expensive than standard, conforming finite elements or hybridizible DG elements [19]. Even if we disregard element interior degrees-of-freedom (d.o.f.) ("bubbles"), the number of d.o.f. is of the same order as for mixed and standard DG methods, i.e. it is doubled. On top of it, the element computations are essentially more expensive. The extra computational cost is balanced with the extraordinary stability properties and the possibility of controling the norm in which we converge through the selection of test norm. This is a unique property that distinguishes DPG from other methods.

The method comes with a built-in a posteriori *error evaluator*. A routine evaluating the residual error for an element is a calc copy of the element routine. Anyone who had to implement even a simple, explicit

a-posteriori error estimate will appreciate this point. The methodology provides thus a very natural framework for adaptive methods including *hp*-adaptivity.

The main difficulty in implementing the DPG method comes from the fact that it is a hybrid FE method. As the field variables are discretized with L^2-conforming elements, trace variables with traces of H^1-conforming elements, and flux variables with traces of $H(\mathrm{div})$-conforming elements, the DPG method is naturally implementable within any framework supporting the exact sequence elements. Solution of Maxwell problems in 3D will require traces of $H(\mathrm{curl})$-conforming elements as well. Two implementations of this type have been built at ICES and are available for interested parties: a parallel implementation built on top of Sandia's Trilinos [41], and a workstation Fortran 90 version based on our earlier work on *hp* methods [22].

As we have tried to convince the reader in Remark 3.2, with the enriched spaces approach to the computation of optimal test functions, implementation of the element routine is rather straightforward. Otherwise, the rest of the code (assembling, interfacing with solvers, graphical postprocessing, etc.) uses the standard FE technology.

Acknowledgements. The work of the author Leszek F. Demkowicz was supported by the Department of Energy under Award Number DE-FC52-08NA28615.

REFERENCES

[1] I. Babuška. Error-bounds for finite element method. *Numer. Math*, 16, 1970/1971.
[2] A.T. Barker, S.C. Brenner, E.-H. Park, and L.-Y. Sung. A one-level additive Schwarz preconditioner for a discontinuous Petrov–Galerkin method. Technical report, Dept. of Math., Luisiana State University, 2013. http://arxiv.org/abs/1212.2645.
[3] P. Bochev and M.D. Gunzburger. *Least-Squares Finite Element Methods*, volume 166 of *Applied Mathematical Sciences*. Springer Verlag, 2009.
[4] C.L. Bottasso, S. Micheletti, and R. Sacco. The discontinuous Petrov-Galerkin method for elliptic problems. *Comput. Methods Appl. Mech. Engrg.*, 191: 3391–3409, 2002.
[5] C.L. Bottasso, S. Micheletti, and R. Sacco. A multiscale formulation of the discontinuous Petrov-Galerkin method for advective-diffusive problems. *Comput. Methods Appl. Mech. Engrg.*, 194:2819–2838, 2005.
[6] J.H. Bramble, R.D. Lazarov, and J.E. Pasciak. A least-squares approach based on a discrete minus one inner product for first order systems. *Math. Comp*, 66, 1997.
[7] J. Bramwell, L. Demkowicz, J. Gopalakrishnan, and W. Qiu. A locking-free *hp* DPG method for linear elasticity with symmetric stresses. *Num. Math.*, 2012. accepted.
[8] T. Bui-Thanh, L. Demkowicz, and O. Ghattas. Constructively well-posed approximation methods with unity inf-sup and continuity. *Math. Comp.* accepted.
[9] T. Bui-Thanh, L. Demkowicz, and O. Ghattas. A unified discontinuous Petrov-Galerkin Method and its analysis for Friedrichs' systems. Technical Report 34, ICES, 2011. SIAM J. Num. Anal., revised version submitted.

[10] T. Bui-Thanh, O. Ghattas, and L. Demkowicz. A relation between the discontinuous Petrov–Galerkin method and the Discontinuous Galerkin Method. Technical Report 45, ICES, 2011.
[11] V.C. Calo, N.O. Collier, and A.H. Niemi. Analysis of the discontinuous Petrov–Galerkin method with optimal test functions for the Reissner–Mindlin plate bending model. In preparation.
[12] P. Causin and R. Sacco. A discontinuous Petrov-Galerkin method with Lagrangian multipliers for second order elliptic problems. *SIAM J. Numer. Anal.*, 43, 2005.
[13] P. Causin, R. Sacco, and C.L. Bottasso. Flux-upwind stabilization of the discontinuous Petrov-Galerkin formulation with Lagrange multipliers for advection-diffusion problems. *M2AN Math. Model. Numer. Anal.*, 39:1087–1114, 2005.
[14] O. Cessenat and B. Després. Application of an ultra weak variational formulation of elliptic pdes to the two-dimensional helmholtz problem. *SIAM J. Numer. Anal.*, 35(1):255–299, 1998.
[15] J. Chan and L. Demkowicz. Global properties of DPG test spaces for convection–diffusion problems. Technical report, ICES, 2013. In preparation.
[16] J. Chan, L. Demkowicz, R. Moser, and N. Roberts. A class of Discontinuous Petrov–Galerkin methods. Part V: Solution of 1D Burgers and Navier–Stokes equations. Technical Report 25, ICES, 2010.
[17] J. Chan, T. Ellis, L. Demkowicz, and N. Roberts. Element conservation properties in DPG method. Technical report, ICES, 2013. In preparation.
[18] J. Chan, N. Heuer, Tan Bui-Thanh B., and L. Demkowicz. Robust DPG method for convection-dominated diffusion problems II: Natural inflow condition. Technical Report 21, ICES, June 2012. submitted to Comput. Math. Appl.
[19] B. Cockburn, J. Gopalakrishnan, and R. Lazarov. Unified hybridization of Discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems. *SIAM J. Numer. Anal.*, 47(2):1319–1365, 2009.
[20] Al Cohen, W. Dahmen, and G. Welper. Adaptivity and variational stabilization for convection-diffusion equations. Technical Report 323, Institut fuer Geometrie und Praktische Mathematik, 2011.
[21] W. Dahmen, Ch. Huang, Ch. Schwab, and G. Welper. Adaptive Petrov Galerkin methods for first order transport equations. Technical Report 321, Institut fuer Geometrie und Praktische Mathematik, 2011.
[22] L. Demkowicz. *Computing with hp Finite Elements. I.One- and Two-Dimensional Elliptic and Maxwell Problems.* Chapman & Hall/CRC Press, Taylor and Francis, October 2006.
[23] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. Part I: The transport equation. *Comput. Methods Appl. Mech. Engrg.*, (23–24):1558–1572, 2010. see also ICES Report 2009–12.
[24] L. Demkowicz and J. Gopalakrishnan. Analysis of the DPG method for the Poisson problem. *SIAM J. Num. Anal.*, 49(5):1788–1809, 2011.
[25] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. Part II: Optimal test functions. *Numer. Meth. Part. D. E.*, 27: 70–105, 2011. see also ICES Report 9/16.
[26] L. Demkowicz, J. Gopalakrishnan, I. Muga, and J. Zitelli. Wavenumber explicit analysis for a DPG method for the multidimensional Helmholtz equation. *Comput. Methods Appl. Mech. Engrg.*, 213–216:126–138, 2012.
[27] L. Demkowicz, J. Gopalakrishnan, and A. Niemi. A class of discontinuous Petrov-Galerkin methods. Part III: Adaptivity. *Appl. Numer. Math.*, 62(4):396–427, 2012. see also ICES Report 2010/1.
[28] L. Demkowicz, J. Gopalakrishnan, and J. Wang. A primal DPG method with no numerical traces. 2013. In preparation.
[29] L. Demkowicz and N. Heuer. Robust DPG method for convection-dominated diffusion problems. Technical report, ICES, 2011. SIAM J. Num. Anal., in review.
[30] L. Demkowicz and J. Li. Numerical simulations of cloaking problems using a DPG method. *Comp. Mech.*, 2012. In print.

[31] L. Demkowicz and J. T. Oden. An adaptive characteristic Petrov-Galerkin finite element method for convection-dominated linear and nonlinear parabolic problems in one space variable. *Journal of Computational Physics*, 68(1):188–273, 1986.

[32] L. Demkowicz and J. T. Oden. An adaptive characteristic Petrov-Galerkin finite element method for convection-dominated linear and nonlinear parabolic problems in two space variables. *Comput. Methods Appl. Mech. Engrg.*, 55(1–2):65–87, 1986.

[33] J. Gopalakrishnan and W. Qiu. An analysis of the practical DPG method. *Math. Comp.*, 2012. accepted.

[34] D. Moro, N.C. Nguyen, and J. Peraire. A hybridized discontinuous Petrov–Galerkin scheme for scalar conservation laws. *Int.J. Num. Meth. Eng.*, 2011. in print.

[35] A.H. Niemi, J.A. Bramwell, and L.F. Demkowicz. Discontinuous Petrov–Galerkin method with optimal test functions for thin-body problems in solid mechanics. *Comput. Methods Appl. Mech. Engrg.*, 200:1291–1300, 2011.

[36] A.H. Niemi, N.O. Collier, and V.M. Calo. Automatically stabilized discontinuous Petrov-Galerkin methods for stationary transport problems: Quasi-optimal test space norm. 2011. In preparation.

[37] A.H. Niemi, N.O. Collier, and V.M. Calo. Discontinuous Petrov-Galerkin method based on the optimal test space norm for one-dimensional transport problems. *Journal of Computational Science*, 2011. In press.

[38] J.T. Oden, L. Demkowicz, T. Strouboulis, and Ph. Devloo. *Accuracy Estimates and Adaptive Refinements in Finite Element Computations*, chapter Adaptive Methods for Problems in Solid and Fluid Mechanics. John Wiley and Sons Ltd., London, 1986.

[39] J.T. Oden and L.F. Demkowicz. *Applied Functional Analysis for Science and Engineering*. Chapman & Hall/CRC Press, Boca Raton, 2010. Second edition.

[40] N. Roberts, Tan Bui-Thanh B., and L. Demkowicz. The DPG method for the Stokes problem. Technical Report 22, ICES, June 2012. submitted to Comput. Math. Appl.

[41] N.V. Roberts, D. Ridzal, P.B. Bochev, and L. Demkowicz. A toolbox for a class of discontinuous Petrov-Galerkin methods using Trilinos. Technical Report SAND2011-6678, Sandia National Laboratories, 2011.

[42] H. Roos, M. Stynes, and L. Tobiska. *Robust Numerical Methods for Singularly Perturbed Differential Equations: Convection-Diffusion-Reaction and Flow Problems*. Springer, 2008.

[43] Ch. Wieners and B. Wohlmuth. Robust operator estimates. Technical report, Oberwolfach Reports, 2013.

[44] K. Yosida. *Functional Analysis*. Springer-Verlag, New York, 4 edition, 1974.

[45] J. Zitelli, I. Muga, L. Demkowicz, J. Gopalakrishnan, D. Pardo, and V. Calo. A class of discontinuous Petrov-Galerkin methods. Part IV: Wave propagation problems. *J. Comp. Phys.*, 230:2406–2432, 2011.

DISCONTINUOUS GALERKIN FOR THE RADIATIVE TRANSPORT EQUATION

JEAN-LUC GUERMOND*, GUIDO KANSCHAT†, AND JEAN C. RAGUSA‡

Abstract. This note presents some recent results regarding the approximation of the linear radiative transfer equation using discontinuous Galerkin methods. The locking effect occurring in the diffusion limit with the upwind numerical flux is investigated and a correction technique is proposed.

Key words. Finite elements, Discontinuous Galerkin, Neutron transport, Diffusion limit

AMS(MOS) subject classifications. 65N35, 65N22, 65F05, 35J05

1. Introduction. The linear radiative transfer equation describes the processes by which particles (photons, neutrons, ...) interact with a background medium. Such processes play a crucial role in stellar atmospheres, nuclear reactor analysis, and shielding applications. The Discontinuous Galerkin (DG) finite element technique has been introduced by Reed and Hill [16] and Lesaint and Raviart [13] in the early 1970s to specifically solve this equation. It has been observed in the literature that the DG approximation with the upwind flux locks when the physical medium is optically thick. In this case the width of the medium is many mean free paths and the interaction processes are scattering-dominated. In the present paper we adopt the terminology of Babuška and Suri [3]: *a numerical scheme for the approximation of a parameter-dependent problem is said to exhibit locking if the accuracy of the approximations deteriorates as the parameter tends to a limiting value. A robust numerical scheme for the problem is one that is essentially uniformly convergent for all values of the parameter.* The objective of this paper is to review the influence of the definition of the numerical flux of the DG method when the medium is optically thick.

The paper is organized as follows. Section 2 introduces notation and recalls the S_N transport equation. Section 3 describes the discrete formulation which is obtained when applying a discontinuous Galerkin technique to the S_N equations. The origin of the locking phenomenon occurring when the DG method is equipped with the upwind flux in identified in Sect. 4.

*Department of Mathematics, Texas A&M University, College Station, TX 77843, USA. guermond@math.tamu.edu

†Heidelberg University, Im Neuenheimer Feld 368 69120 Heidelberg, USA. guido.kanschat@.iwr.uni-heidelberg.de

‡Department of Nuclear Engineering, Texas A&M University, College Station, TX 77843, USA. ragusa@ne.tamu.edu

X. Feng et al. (eds.), *Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations*, The IMA Volumes in Mathematics and its Applications 157, DOI 10.1007/978-3-319-01818-8_7,

A modified numerical flux is analyzed in Sect. 5. Numerical results illustrating the performance of the modified numerical flux are presented at the end of this section.

2. Formulation of the Problem and S_N Discretization. We recall in this section the transport equation and we provide some notations for angular discretization. To keep the discussion simple, we limit ourselves to the one-group discrete-ordinates equations; these equations model one-group neutron transport and grey radiative transfer.

2.1. The Transport Equation. Let $\mathcal{D}$ be the spatial domain in $\mathbb{R}^d$ (with $d = 1, 2, 3$), $\partial\mathcal{D}$ be the boundary of $\mathcal{D}$, $\mathbf{n}$ be the outward unit normal vector on $\partial\mathcal{D}$, and S^2 be the unit sphere in $\mathbb{R}^3$. The set of propagation directions $\mathcal{S}$ is defined as S^2 for $d = 3$ and as the projection of S^2 onto $\mathbb{R}^d$ when $d = 1, 2$. For instance, $\mathcal{S}$ is the unit disk if $d = 2$ and $\mathcal{S}$ is the unit segment $[-1, +1]$ if $d = 1$. Denoting meas($\mathcal{S}$) the measure of $\mathcal{S}$, we have meas($\mathcal{S}$) $= 4\pi$ if $d = 3$, meas($\mathcal{S}$) $= \pi$ if $d = 2$, and meas($\mathcal{S}$) $= 2$ if $d = 1$. This convention, which is common in the radiation transport community, means that radiation is accounted for as a three-dimensional effect even in lower dimensional geometries. The transport of particles is then modeled by the linear Boltzmann equation:

$$\mathbf{\Omega}\cdot\nabla\Psi(\mathbf{\Omega},\mathbf{x}) + \sigma_t(\mathbf{x})\Psi(\mathbf{\Omega},\mathbf{x}) - \sigma_s(\mathbf{x})\,\overline{\Psi}(\mathbf{x}) = q(\mathbf{x}), \qquad \forall(\mathbf{\Omega},\mathbf{x}) \in \mathcal{S}\times\mathcal{D}, \tag{2.1a}$$

where $\overline{\Psi} = \frac{1}{4\pi}\int_{\mathcal{S}} \Psi(\mathbf{\Omega},\mathbf{x})\,\mathrm{d}\mu(\mathbf{\Omega})$ is the scalar flux and the boundary conditions are

$$\Psi(\mathbf{\Omega},\mathbf{x}) = \Psi^{\mathrm{inc}}(\mathbf{\Omega},\mathbf{x}), \qquad \forall(\mathbf{\Omega},\mathbf{x}) \in \mathcal{S}\times\partial\mathcal{D},\ \mathbf{\Omega}\cdot\mathbf{n}(\mathbf{x}) < 0. \tag{2.1b}$$

where $\mathbf{n}$ is the outward unit normal vector on $\partial\mathcal{D}$. The measure $\mathrm{d}\mu(\mathbf{\Omega})$ is such that $\mathrm{d}\mu(\mathbf{\Omega}) = \mathrm{d}\mathbf{\Omega}$ if $d = 3$, $\mathrm{d}\mu(\mathbf{\Omega}) = 2(1 - |\mathbf{\Omega}|^2)^{-\frac{1}{2}}\,\mathrm{d}\mathbf{\Omega}$ if $d = 2$ and $\mathrm{d}\mu(\mathbf{\Omega}) = 2\pi|\mathbf{\Omega}|(1-|\mathbf{\Omega}|^2)^{-\frac{1}{2}}\,\mathrm{d}\mathbf{\Omega}$ if $d = 1$, where $\mathrm{d}\mathbf{\Omega}$ is the Lebesgue measure over the unit sphere in $\mathbb{R}^3$ or the Lebesgue measure in $\mathbb{R}^d$ if $d = 1, 2$. For simplicity, we have assumed that the scattering and the extraneous sources are isotropic; this assumption does not affect the conclusions of the analysis. The dependent variable is the angular flux $\Psi(\mathbf{\Omega},\mathbf{x})$, and the independent variables $(\mathbf{\Omega},\mathbf{x})$ span $\mathcal{S}\times\mathcal{D}$. The given data are the extraneous source term $q(\mathbf{x})$, the incoming boundary radiation $\Psi^{\mathrm{inc}}(\mathbf{\Omega},\mathbf{x})$, the scattering cross section $\sigma_s(\mathbf{x})$, and the absorption cross section $\sigma_a(\mathbf{x}) := \sigma_t(\mathbf{x}) - \sigma_s(\mathbf{x})$.

2.2. The S_N Discretization. A traditional way to approximate the Eq. (2.1a) consists of dealing with $\mathcal{S}$ and $\mathcal{D}$ separately. In this paper the approximation with respect to the angles is done by using the so-called S_N-method. The S_N, or discrete-ordinates, version of (2.1a) is obtained by solving the transport equation along discrete directions (or ordinates) and

by replacing the integrals over the unit sphere $\mathcal{S}$ by quadratures. In the rest of the paper we assume that we have at hand a quadrature rule $\{(\mathbf{\Omega}_j, \omega_j), j = 1, \ldots, n_\Omega\}$

$$\frac{1}{4\pi}\int_{\mathcal{S}} f(\mathbf{\Omega}, \mathbf{x})\, \mathrm{d}\mu(\mathbf{\Omega}) \approx \sum_{j=1}^{n_\Omega} \omega_j f(\mathbf{\Omega}_j, \mathbf{x}), \tag{2.2}$$

satisfying the following properties:

$$\sum_{j=1}^{n_\Omega} \omega_j = 1, \qquad \sum_{j=1}^{n_\Omega} \omega_j \mathbf{\Omega}_j = 0, \tag{2.3}$$

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^3, \qquad \sum_{j=1}^{n_\Omega} \omega_j (\mathbf{\Omega}_j \cdot \mathbf{a})(\mathbf{\Omega}_j \cdot \mathbf{b}) = \frac{1}{3}\mathbf{a}\cdot\mathbf{b}, \tag{2.4}$$

$$\exists c_0 > 0, \forall n_\Omega, \qquad c_{\mathbf{n}} := \sum_{\mathbf{\Omega}_j \cdot \mathbf{n} < 0} \omega_j |\mathbf{\Omega}_j \cdot \mathbf{n}| \geq c_0. \tag{2.5}$$

Although it is a standard result that $\frac{1}{4\pi}\int_{\mathbf{\Omega}\in S^2, \mathbf{\Omega}\cdot\mathbf{n}<0} |\mathbf{\Omega}\cdot\mathbf{n}|\, \mathrm{d}\mathbf{\Omega} = \frac{1}{4}$ for any unit vector $\mathbf{n}$, this equality may not exactly hold for any numerical quadrature at hand. However, reasonable sets of quadrature rules are such that this limit value is approached as the number of directions in the quadrature increases ($\lim_{n_\Omega \to \infty} c_{\mathbf{n}} = \frac{1}{4}$). In any case the hypothesis (2.5) holds whenever one can find d linearly independent vectors among the quadrature points $\mathbf{\Omega}_j$.

The S_N method consists of replacing the angular flux $\Psi(\mathbf{\Omega}, \mathbf{x})$ by a discrete angular flux $\psi(\mathbf{x}) = (\psi_1(\mathbf{x}), \psi_2(\mathbf{x}), \ldots, \psi_{n_\Omega}(\mathbf{x}))$, and to convert the integro-differential equation (2.1) over $\mathcal{S}\times\mathcal{D}$ into a system of n_Ω coupled partial differential equations over $\mathcal{D}$ for all the directions j as follows :

$$\mathbf{\Omega}_j \cdot \nabla \psi_j(\mathbf{x}) + \sigma_t(\mathbf{x})\psi_j(\mathbf{x}) - \sigma_s(\mathbf{x})\,\overline{\psi}(\mathbf{x}) = q(\mathbf{x}), \qquad \text{in } \mathcal{D}, \tag{2.6a}$$

with the inflow boundary condition

$$\psi_j(\mathbf{x}) = \Psi_j^{\mathrm{inc}}(\mathbf{x}), \qquad \forall \mathbf{x} \in \partial\mathcal{D} \text{ with } \mathbf{\Omega}_j \cdot \mathbf{n}(\mathbf{x}) < 0. \tag{2.6b}$$

The discrete scalar flux is defined by:

$$\overline{\psi}(\mathbf{x}) = \sum_{j=1}^{n_\Omega} \omega_j \psi_j(\mathbf{x}). \tag{2.7}$$

The discrete angular flux ψ is said to be isotropic when $\psi_j = \overline{\psi}$, for all $j \in \{1, \ldots, n_\Omega\}$. In order to simplify the notation in subsequent sections, we introduce the discrete current vector $\mathbf{J}(\psi)$, also known as the first angular moment of ψ, as follows:

$$\mathbf{J}(\psi) = \sum_{j=1}^{n_\Omega} \omega_j \psi_j(\mathbf{x}) \mathbf{\Omega}_j. \tag{2.8}$$

Note that $\mathbf{J}(\psi) = 0$ whenever ψ is isotropic.

2.3. Diffusion Limit. We say that the medium is optically thick when it takes many mean free paths for particles to cross the domain. In order to better understand the behavior of the solutions of the linear Boltzmann equation in this regime, we rescale the equation under the assumption that the ratio between the mean free path between two scattering events and the characteristic size (diameter) of the domain goes to zero. A measure of this ratio is given by

$$\varepsilon = \frac{1}{\sigma_s \operatorname{diam}(\mathcal{D})}. \tag{2.9}$$

This parameter is well known to characterize the diffusivity of the problem, see, for instance, Larsen et al. [12] and Dautray and Lions [5, Chap. XXI]. We assume throughout this section that σ_s is constant over the domain to simplify the analysis. Then, we assume the following behaviors

$$\sigma_s = \varepsilon^{-1}\tilde{\sigma}_s, \qquad \sigma_a = \varepsilon\tilde{\sigma}_a, \qquad q = \varepsilon\tilde{q}, \tag{2.10}$$

where the tilde quantities are independent of ε [note in particular that $\tilde{\sigma}_s = 1/\operatorname{diam}(\mathcal{D})$]. As ε goes to zero, the scattering and total cross sections take large values and the absorption cross section becomes small, rendering the configuration optically thick and diffusive.

Using (2.10), the scaled version of the transport equation (2.1) becomes

$$\boldsymbol{\Omega}\cdot\nabla\Psi(\boldsymbol{\Omega},\mathbf{x}) + \left(\frac{\tilde{\sigma}_s}{\varepsilon} + \varepsilon\tilde{\sigma}_a\right)\Psi(\boldsymbol{\Omega},\mathbf{x}) - \frac{\tilde{\sigma}_s}{\varepsilon}\overline{\Psi}(\mathbf{x}) = \varepsilon\tilde{q}(\mathbf{x}). \tag{2.11}$$

It is now well understood (see, e.g., Chandrasekhar [4], Larsen et al. [12], and Dautray and Lions [5, Chap. XXI]) that $\lim_{\varepsilon\to 0}\Psi(\boldsymbol{\Omega},\mathbf{x}) = \lim_{\varepsilon\to 0}\overline{\Psi}(\mathbf{x}) = \varphi(\mathbf{x})$, where the scalar flux φ satisfies the diffusion problem

$$-\nabla\cdot\left(\frac{1}{3\tilde{\sigma}_s}\nabla\varphi\right) + \tilde{\sigma}_a\varphi = \tilde{q}, \tag{2.12a}$$

$$\varphi(\boldsymbol{x}) = \frac{1}{2\pi}\int_{\boldsymbol{\Omega}\in S^2,\,\boldsymbol{\Omega}\cdot\mathbf{n}(\boldsymbol{x})<0} W(|\boldsymbol{\Omega}\cdot\mathbf{n}(\boldsymbol{x})|)\Psi^{\mathrm{inc}}(\boldsymbol{\Omega},\mathbf{x})\,\mathrm{d}\boldsymbol{\Omega}, \qquad \forall \boldsymbol{x}\in\partial D, \tag{2.12b}$$

where $W(\mu) = \frac{\sqrt{3}}{2}\mu H(\mu)$ is defined in terms of Chandrasekhar's H-function for isotropic scattering in a conservative medium (see Malvagi and Pomraning [14] for the asymptotic analysis and Chandrasekhar [4] for details on the H-function). It is shown in Malvagi and Pomraning [14] that $\lim_{\varepsilon\to 0}\Psi = \varphi$, and the convergence is not uniform unless the incident flux is isotropic.

It is remarkable that under the assumptions made for the angular quadrature, the diffusion limit of the solution to the semi-discrete problem (2.6) (discrete-ordinate transport equation) has the same limit properties, i.e.,

$$\lim_{\varepsilon\to 0}\psi_j(\mathbf{x}) = \lim_{\varepsilon\to 0}\overline{\psi}(\mathbf{x}) = \varphi(\mathbf{x}), \qquad \forall j\in\{1,\dots,n_\Omega\}. \tag{2.13}$$

The goal of the present paper is to determine when the above property holds when space is approximated using discontinuous Galerkin methods.

3. DG Discretization. We now proceed with the spatial discretization of the S_N transport equation using DG finite elements.

3.1. The Mesh. Let $\mathbb{T}_h$ be a subdivision of $\mathcal{D}$ into disjoint (open) cells K such that the closure of $\mathcal{D}$ is equal to $\cup_{K\in\mathbb{T}_h}\overline{K}$. The meshes are assumed to be affine to avoid unnecessary technicalities; i.e., $\mathcal{D}$ is assumed to be a polyhedron. The diameter of $K \in \mathbb{T}_h$ is denoted by h_K, and we set $h = \max_{K\in\mathbb{T}_h} h_K$. We suppose that we have at hand a family of meshes $\{\mathbb{T}_h\}$ and that this family is uniformly shape-regular. We also assume that the mesh is quasi-uniform; i.e., there is $c > 0$ so that

$$c\,h \le h_K \le h \qquad \forall K \in \mathbb{T}_h. \tag{3.1}$$

This hypothesis is used when invoking inverse inequalities. It could be avoided by localizing the inverse estimate arguments, but we shall refrain from doing so to steer clear of unnecessary technicalities.

We denote $\mathbb{F}_h^i$ the set of interior faces (also called interfaces); each face $F \in \mathbb{F}_h^i$ is the intersection of the boundaries of two mesh cells. We assign a normal vector $\mathbf{n}$ for each face $F \in \mathbb{F}_h^i$. While the choice of the normal vector is arbitrary for interior faces, all the weak formulations considered below are independent of this choice and thus well defined. The set of faces on the domain boundary, $\partial\mathcal{D}$, is denoted $\mathbb{F}_h^b$. The set of interfaces and boundary faces is denoted $\mathbb{F}_h = \mathbb{F}_h^i \cup \mathbb{F}_h^b$.

3.2. The Discontinuous Galerkin Setting. We define a discontinuous approximation space for the scalar flux based on the mesh $\mathbb{T}_h$ as follows:

$$V_h = \left\{ v \in L^2(\mathcal{D}) \;\middle|\; \forall K \in \mathbb{T}_h,\, v|_K \in P_K, \right\}, \tag{3.2}$$

where, denoting $\mathbb{P}_k$ the set of polynomials of degree at most k, the finite-dimensional space P_K is assumed to contain $\mathbb{P}_k$, i.e.,

$$\mathbb{P}_k \subset P_K, \quad \forall K \in \mathbb{T}_h. \tag{3.3}$$

The discrete space for the angular flux, W_h, simply consists of copies of V_h for each of the discrete ordinates:

$$W_h = (V_h)^{n_\Omega}. \tag{3.4}$$

We finally introduce the spaces with zero boundary conditions

$$V_{0,h} = \left\{ v \in V_h \middle|\; v|_{\partial\mathcal{D}} = 0 \right\}, \qquad W_{0,h} = \left(V_{0,h}\right)^{n_\Omega}. \tag{3.5}$$

3.3. The DG Weak Formulation. The DG formulation of the problem (2.6) consists of seeking $\psi \in W_h$ so that the following holds for all cells $K \in \mathbb{T}_h$, for all test functions $v_j \in V_h$ supported on K, and for all direction $j \in \{1, \ldots, n_\Omega\}$:

$$\int_K \left(-\psi_j \mathbf{\Omega}_j \cdot \nabla v_j + (\sigma_s + \sigma_a)\, \psi_j v_j - \sigma_s \overline{\psi} v_j\right) \mathrm{d}\mathbf{x} + \int_{\partial K} \widehat{\mathbf{F}_j}(\mathbf{x}) \cdot \mathbf{n} v_j \,\mathrm{d}\mathbf{x} = \int_K q v_j \,\mathrm{d}\mathbf{x}. \tag{3.6}$$

where the numerical flux*$\widehat{\mathbf{F}_j}$ has yet to be defined. The purpose of the numerical flux $\widehat{\mathbf{F}_j}(\mathbf{x})\cdot\mathbf{n}$ is to approximate the quantity $\psi_j \mathbf{\Omega}_j \cdot \mathbf{n}$ at the mesh interfaces since this quantity is double-valued due to the discontinuous nature of the approximation. The above system is obtained by (i) multiplying the S_N equations for direction j with test function v_j, (ii) integrating the results by parts, and (iii) replacing the two-valued function $\psi_j \mathbf{\Omega}_j \cdot \mathbf{n}$ by the numerical flux $\widehat{\mathbf{F}_j} \cdot \mathbf{n}$.

3.4. Jumps and Averages. Due to the discontinuous nature of the spatial approximation, functions $v \in V_h$ are double-valued on interior faces. let $F \in \mathbb{F}_h^i$ be an interior face separating two mesh cells, K_1 and K_2. The mean value and jump of a function $v \in V_h$ across F are defined as follows:

$$\{\!\{v\}\!\} = \tfrac{1}{2}(v_1 + v_2), \qquad [\![v]\!] = v_1 - v_2, \tag{3.7}$$

where $v_1 := v|_{K_1}$ and $v_2 := v|_{K_2}$ are the restrictions of v on the mesh cells K_1 and K_2, respectively. Obviously, $\{\!\{v\}\!\}$ does not depend on the numbering of the cells K_1 and K_2, but the jump does (there is a sign change when exchanging the cells K_1 and K_2). However, since the weak bilinear forms (to be defined further below) contain the product of two jumps, the orientation of the unit normal vector does not matter. Let $\mathbf{n}_1$ and $\mathbf{n}_2$ be the unit normal vectors on F pointing towards K_2 and K_1, respectively. The mean value of quantities containing a normal vector is actually a jump since

$$\{\!\{v\mathbf{n}\}\!\} = \tfrac{1}{2}(v_1\mathbf{n}_1 + v_2\mathbf{n}_2) = \tfrac{1}{2}(v_1 - v_2)\mathbf{n}_1 = \tfrac{1}{2}(v_2 - v_1)\mathbf{n}_2.$$

For any v in V_h and any interior face $F \in \mathbb{F}_h^i$, we introduce the so-called upwind and downwind values of v at $\mathbf{x} \in F$, $v^{\uparrow}(\mathbf{x})$ and $v^{\downarrow}(\mathbf{x})$, respectively, as follows:

*The term "flux" is used in two different contexts. In the radiation transport context, we use the terms "angular flux" and "scalar flux." In the DG context, we use the notion of "numerical flux." These two notions are unfortunately unrelated but commonly employed in the radiation transport and DG literature, respectively. To avoid confusion, we always try to use the proper adjective in this paper.

$$v^{\uparrow}(\mathbf{x}) = \begin{cases} v_1(\mathbf{x}), & \text{if } \boldsymbol{\Omega}\cdot\mathbf{n}_1(\mathbf{x}) \geq 0 \\ v_2(\mathbf{x}), & \text{if } \boldsymbol{\Omega}\cdot\mathbf{n}_1(\mathbf{x}) < 0, \end{cases}$$
$$v^{\downarrow}(\mathbf{x}) = \begin{cases} v_2(\mathbf{x}) & \text{if } \boldsymbol{\Omega}\cdot\mathbf{n}_1(\mathbf{x}) \geq 0 \\ v_1(\mathbf{x}) & \text{if } \boldsymbol{\Omega}\cdot\mathbf{n}_1(\mathbf{x}) < 0. \end{cases} \tag{3.8}$$

Observing that the following holds for any positive number ($\gamma \geq 0$):

$$\boldsymbol{\Omega}\cdot\mathbf{n}_1 \{\!\{v\}\!\} + \tfrac{1}{2}\gamma|\boldsymbol{\Omega}\cdot\mathbf{n}_1|[\![v]\!] = \boldsymbol{\Omega}\cdot\mathbf{n}_1 \left(v^{\uparrow}(\mathbf{x}) + \tfrac{1}{2}(\gamma-1)(v^{\uparrow}(\mathbf{x}) - v^{\downarrow}(\mathbf{x}))\right), \tag{3.9}$$

we obtain that

$$\boldsymbol{\Omega}\cdot\mathbf{n}_1 \{\!\{v\}\!\} + \frac{\gamma}{2}|\boldsymbol{\Omega}\cdot\mathbf{n}_1|[\![v]\!] = \begin{cases} \boldsymbol{\Omega}\cdot\mathbf{n}_1 v^{\uparrow}(\mathbf{x}) & \text{if } \gamma = 1, \\ \boldsymbol{\Omega}\cdot\mathbf{n}_1 \{\!\{v\}\!\} & \text{if } \gamma = 0. \end{cases} \tag{3.10}$$

The so-called upwind DG numerical flux is obtained with (3.9) by using $\gamma = 1$, and the centered numerical flux is obtained by using $\gamma = 0$. The representation (3.9) gives an easy way to construct numerical fluxes by modifying the coefficient γ.

4. The Upwind Approximation. In the radiative transfer literature it is common to replace $\widehat{\mathbf{F}_j}(\mathbf{x})$ in (3.6) by the upwind flux

$$\widehat{\mathbf{F}_j}\cdot\mathbf{n} = \boldsymbol{\Omega}_j\cdot\mathbf{n}\psi_j^{\uparrow}(\mathbf{x}). \tag{4.1}$$

We focus in this section on the consequences of this choice. We show in particular that it leads to locking in the diffusive regime for some families of approximation spaces.

4.1. Locking in the Diffusion Regime. It has been observed in the literature that the DG approximation (3.6) equipped with the upwind flux locks when the medium is optically thick. For instance, it is pointed out in Larsen [8, 9] that the so-called step scheme, a finite volume scheme (i.e., a piecewise constant DG scheme) with standard upwind, locks in the diffusion limit. A modification of the "step scheme" depending upon the total mean free path was proposed in Larsen [8] to correct the locking of the method in the diffusion limit, but this required modifying the streaming term and abandoning particle balance. Several other variations of the "step scheme" have been analyzed in Larsen et al. [12]: it was shown that the "Lund–Wilson" and the "Castor" variants of the step scheme yielded cell-edge angular fluxes that lock in the diffusion limit, and that the auxiliary relations linking the outgoing edge angular flux to the cell-average angular flux employ a multiplicative factor that depends on the mesh cell optical thickness in the direction traveled. Furthermore, the cell-edge fluxes for these schemes cannot reproduce the infinite medium solution. A "new" scheme was proposed in Larsen et al. [12] but was subsequently dismissed due its

poor behavior at the boundaries. For many years, the diamond-difference scheme was found to be the best performing finite-difference scheme, even though its cell-edge fluxes lock in the thick diffusion limit. In Larsen and Morel [11], most of the previous schemes have been set aside in favor of the Linear Discontinuous finite element scheme (the piecewise linear DG technique with standard upwinding).

Adams [1] analyzed multi-dimensional DG approximations and showed that some schemes lock in the diffusion limit because the upwind method forces the scalar flux, and thus the angular flux, to be continuous across mesh cells. This observation is essential to understand what happens.

4.2. Convergence Analysis. In the rest of the paper we adopt the scaling defined in (2.10) and consider the rescaled transport equation (2.11).

The analysis of Adams [1] have been confirmed in Guermond and Kanschat [7], where the equivalence of the limit problem to a mixed discretization for the Laplacian was proved and the nature of the boundary layers was discussed. To better formulate the conclusions from Guermond and Kanschat [7], we introduce the subspace of V_h composed of the functions that are continuous:

$$C_h = V_h \cap \mathcal{C}^0(\overline{\mathcal{D}}), \tag{4.2}$$

and we define $m(\boldsymbol{x}) := \frac{1}{\pi} \int_{\boldsymbol{\Omega}\cdot\mathbf{n}(\mathbf{x})<0} \Psi^{\text{inc}}(\boldsymbol{\Omega}, \mathbf{x}) |\boldsymbol{\Omega}\cdot\mathbf{n}(\mathbf{x})| \, \mathrm{d}\boldsymbol{\Omega}$. The first key result is the following:

LEMMA 4.1. *Assume that m is the trace of a function in C_h. Then the solution of (3.6) with the upwind flux (4.1) is such that*

$$\lim_{\varepsilon\to 0} \psi_j \in C_h, \qquad \forall j \in \{0, \ldots, n_\Omega\}. \tag{4.3}$$

REMARK 4.1. *An immediate consequence of this result is that while piecewise constant approximation is admissible for solving the transport problem (2.1a) (or (2.11)), the continuity condition (4.3) forces the diffusion limit solution to be globally constant. This leads to locking, i.e., $\lim_{\varepsilon\to 0} \psi_j$ does not converge to φ when using DG0 with the upwind flux, unless φ is constant.*

Let us further assume that the following approximation properties hold for all $\phi \in H^l(D)$ and all $l \in [1, 2]$:

$$\inf_{v_h \in \mathcal{C}_{h,0}} \|\phi - v_h\|_{H^p(D)} \le c h^{l-p} \|\phi\|_{H^l(D)}, \quad \forall p \in [0, 1], \tag{4.4}$$

$$\inf_{v_h \in \mathcal{C}_{h,0}} \left(\|\phi - v_h\|_{L^2(\partial D)} + h \|\partial_n(\phi - v_h)\|_{L^2(\partial D)} \right) \le c h^{l-\frac{1}{2}} \|\phi\|_{H^l(D)}. \tag{4.5}$$

The following result is then proved in Guermond and Kanschat [7]:

THEOREM 4.1. *Assume that* Ψ^{inc} *is isotropic and smooth enough and that* (4.4) *and* (4.5) *hold. Then the solution of* (3.6) *with the upwind flux* (4.1) *converges in* $H^1(\mathcal{D})$ *to* φ*, solution of* (2.12a) *and* (2.12b)*, and the following error estimate holds:*

$$\| \lim_{\varepsilon \to 0} \psi_j - \varphi \|_{H^1(\mathcal{D})} \leq c \inf_{v_h \in \mathcal{C}_{h,0}} \| \varphi - v_h \|_{H^1(\mathcal{D})}, \qquad \forall j \in \{0, \ldots, n_\Omega\}. \quad (4.6)$$

The critical assumption here is (4.4), which requires the spaces $\mathcal{C}_h$ to be rich enough so as to have reasonable approximation properties. This is a condition on the mesh family $\{\mathbb{T}_h\}_{h>0}$ and the associated discrete space family $\{V_h\}_{h>0}$. More precisely (4.4) holds if the following two conditions are satisfied:

(i) The meshes are conforming; i.e., each face of a cell is either the face of a neighboring cell or at the boundary. This condition can be weakened to accommodate for local refinement, and in this case each face of any cell may be a subset of a face of its neighbor.

(ii) The polynomial spaces on each cell must allow continuity across interfaces of neighboring cells without loosing approximation properties. This is usually achieved by using multidimensional polynomial spaces $\mathbb{P}_k$ of total order $k \geq 1$ for triangles and tetrahedra or mapped tensor product spaces $\mathbb{Q}_k$ of order $k \geq 1$ in each coordinate direction on quadrilaterals and hexahedra.

REMARK 4.2. *For instance, condition* (ii) *is violated if piecewise constant elements are used.*

REMARK 4.3. *Conditions* (i) *and* (ii) *have been identified in Adams [1] and termed "locality" and "surface-matching" properties. We think though that the condition* (4.4) *gives a complementary rational to that given in Adams [1]. Lists of admissible and nonadmissible finite elements are given in Tables I and II in Adams [1].*

When the incoming flux at the boundary is not isotropic some boundary layer effect occur as mentioned in Adams [1] and Larsen and Keller [10]. To formulate a precise result we introduce the function $\mathbf{M}(\boldsymbol{x}) := \frac{1}{4\pi} \int_{\boldsymbol{\Omega}\cdot\mathbf{n}(\mathbf{x})<0} \Psi^{\text{inc}}(\boldsymbol{\Omega}, \mathbf{x}) |\boldsymbol{\Omega}\cdot\mathbf{n}(\mathbf{x})| \boldsymbol{\Omega} \, \mathrm{d}\boldsymbol{\Omega}$.

THEOREM 4.2. *Assume that* $\mathbf{M}(\boldsymbol{x})\cdot\mathbf{n}$ *is the trace of a function in* C_h *and* (4.4) *and* (4.5) *hold. Then the solution of* (3.6) *with the upwind flux* (4.1) *converges to a limit* $\psi_{\lim}$ *in* $H^s(\mathcal{D})$ *for all* $s \in [0, \frac{1}{2})$ *and*

$$\| \lim_{\varepsilon \to 0} \psi_j - \psi_{\lim} \|_{L^2(\mathcal{D})} \leq c\, h^{\frac{s}{3}}, \qquad \forall s \in [0, \tfrac{1}{2}), \ \forall j \in \{0, \ldots, n_\Omega\} \quad (4.7)$$

That the above convergence occurs in a space $H^s(\mathcal{D})$ with $s < \frac{1}{2}$ is the signature of boundary layer effects developing at the boundary when the incoming flux is not isotropic.

5. Robust DG Approximation. The asymptotic analysis in Adams [1] and Guermond and Kanschat [7] suggests that the problem could be alleviated by modifying the upwind numerical flux. As pointed out in Ayuso

and Marini [2], Ern and Guermond [6], the upwind numerical flux is only one particular choice among many for stabilization. By making the amount of stabilization dependent on the scattering cross section so that the amount of upwinding decreases as the scattering cross section increases, it is shown in Ragusa et al. [15] that locking can indeed be avoided in the thick diffusive limit.

5.1. Modified Numerical Flux. The new numerical flux proposed by Ragusa et al. [15] is based on (3.9). Before giving its expression we define the following stabilization parameters

$$\gamma(\mathbf{x}) = \frac{\gamma_0}{\max(\gamma_0, \sigma_s(\mathbf{x})\,\mathrm{diam}\,\mathcal{D})}, \qquad \delta(\mathbf{x}) = \delta_0 \frac{1-\gamma(\mathbf{x})}{\gamma(\mathbf{x})}, \tag{5.1}$$

where the parameters $\gamma_0 > 0$, $\delta_0 > 0$ are assumed to be of order one. The rationale for these definitions is as follows: γ tends to 0 in the diffusive limit, whereas γ converges 1 in the optically thin regions.

The following definition for the numerical flux across the interface $F \in \mathbb{F}_h^i$ from K_1 to K_2 is proposed in Ragusa et al. [15]:

$$\widehat{\mathbf{F}_j}(\mathbf{x})\cdot\mathbf{n}_1 = \boldsymbol{\Omega}_j\cdot\mathbf{n}_1 \{\!\{\psi_j\}\!\} + \frac{\gamma(\mathbf{x})}{2}|\boldsymbol{\Omega}_j\cdot\mathbf{n}_1|[\![\psi_j]\!] + \frac{\delta(\mathbf{x})}{2}\{\!\{\mathbf{J}(\psi)\cdot\mathbf{n}\}\!\}\boldsymbol{\Omega}_j\cdot\mathbf{n}_1. \tag{5.2}$$

We use the standard upwind definition of the numerical flux for any boundary face $F \in F_h^b$:

$$\widehat{\mathbf{F}_j}(\mathbf{x})\cdot\mathbf{n} = \begin{cases} \boldsymbol{\Omega}_j\cdot\mathbf{n}\,\Psi_j^{\mathrm{inc}} & \text{if } \boldsymbol{\Omega}_j\cdot\mathbf{n}(\mathbf{x}) < 0 \\ \boldsymbol{\Omega}_j\cdot\mathbf{n}\,\psi_j & \text{otherwise.} \end{cases} \tag{5.3}$$

Note that the definition of $\gamma(\mathbf{x})$ is such that, on the one hand, $\gamma \to 0$ when the ratio of the scattering mean free path to the diameter of the domain is small (i.e., $\sigma_s(\mathbf{x})\mathcal{D}$ is large); on the other hand, γ is bounded away from zero when the mean free path is a non-negligible fraction of the diameter of the domain (the γ_0 constant assures that $\gamma(\mathbf{x}) \to 1$ when $\sigma_s(\mathbf{x})\mathcal{D}$ is small. The parameter δ is designed so that it goes to zero when $\gamma \to 1$ and behaves like $1/\gamma$ when $\gamma \to 0$. This behavior is dictated from the forthcoming asymptotic analysis. The intuitive motivations for the first and second terms in (5.2) are the expressions (3.9) and (3.10). The standard upwind numerical flux is obtained by setting $\gamma = 1$, which also implies $\delta = 0$. The justification for the third term $\{\!\{\mathbf{J}(\psi)\cdot\mathbf{n}\}\!\}\boldsymbol{\Omega}_j\cdot\mathbf{n}_1$ comes from the asymptotic analysis; this term turns out to be necessary for the limit problem to be well posed.

5.2. Convergence Analysis. In the rest of this section we assume that $\Psi^{\mathrm{inc}} = 0$ and we refer the reader to [7] for the handling of inhomogeneous Dirichlet boundary conditions. The main result from [15] is the following:

PROPOSITION 1. *Let $\psi \in W_h$ be the solution to the S_N-DG problem (3.6) equipped with the numerical flux (5.2). Then ψ converges to an isotropic function $\varphi \in V_{0,h}$ as $\varepsilon \to 0$. Furthermore, there is a vector field $\mathbf{J} \in (V_h)^d$ so that the pair $(\varphi, \mathbf{J})$ solves the following DG system for all $v \in V_{0,h}$ and all $\mathbf{L} \in (V_h)^d$:*

$$\begin{aligned}
&\sum_{K\in\mathbb{T}_h} \int_K (\nabla\cdot\mathbf{J} + \tilde{\sigma}_a\varphi)\, v \,\mathrm{d}\mathbf{x} \\
&\quad + \sum_{F\in\mathbb{F}_h^i} \int_F \left(c_{\mathbf{n}_F} \frac{\gamma_0}{2} [\![\varphi]\!][\![v]\!] - 2\{\!\{\mathbf{J}\cdot\mathbf{n}\}\!\}\{\!\{v\}\!\} \right) \mathrm{d}\mathbf{x} = \int_{\mathcal{D}} \tilde{q}\, v \,\mathrm{d}\mathbf{x}, \\
&\sum_{K\in\mathbb{T}_h} \int_K \left(\frac{1}{3}\nabla\varphi + \tilde{\sigma}_s\mathbf{J} \right) \cdot \mathbf{L} \,\mathrm{d}\mathbf{x} \\
&\quad + \sum_{F\in\mathbb{F}_h^i} \int_F \left(-\frac{2}{3}\{\!\{\varphi\mathbf{n}\}\!\}[\![\mathbf{L}]\!] + \frac{\delta_0}{3\gamma_0}\{\!\{\mathbf{J}\cdot\mathbf{n}\}\!\}[\![\mathbf{L}\cdot\mathbf{n}]\!] \right) \mathrm{d}\mathbf{x} = 0,
\end{aligned} \tag{5.4}$$

where $c_{\mathbf{n}_F} := \sum_{\mathbf{\Omega}_j\cdot\mathbf{n}_F\le 0} \omega_j |\mathbf{\Omega}_j\cdot\mathbf{n}_F|$ is bounded away from zero uniformly with respect to $F \in \mathbb{F}_h^i$, h, and n_Ω.

The above result may seem obscure, but the limit problem (5.4) coincides exactly with the method from Ern and Guermond [6] (see Sect. 5.3 therein) that has been proposed to solve the limit problem (2.12a) and (2.12b) in mixed form:

$$\nabla\cdot\mathbf{J} + \tilde{\sigma}_a\varphi = \tilde{q} \tag{5.5a}$$

$$\frac{1}{3}\nabla\varphi + \tilde{\sigma}_s\mathbf{J} = 0 \tag{5.5b}$$

$$\varphi|_{\partial\mathcal{D}} = 0. \tag{5.5c}$$

The theoretical convergence analysis from Ern and Guermond [6] implies that (5.4) is a consistent and convergent approximation of (2.12a) and (2.12b). That is, the discrete transport formulation (3.6) with the numerical flux (5.2) is robust and yields a convergent approximation of the diffusion equation as ε goes to zero.

5.3. Numerical Experiments. We finish this paper by numerically illustrating the above method. We solve the problem of local energy equilibrium in the domain $D = (-1,1)^2\times\mathbb{R}$ with zero incoming flux,

$$\Omega\cdot\nabla\psi(\Omega,x) + \frac{1}{\varepsilon}\big(\psi(\Omega,x) - \overline{\psi}(x)\big) = \frac{\varepsilon}{3}\frac{\pi^2}{4}\prod_{i=1}^{2}\cos\left(\tfrac{\pi x_i}{2}\right).$$

The solution is independent of x_3. We study the limit of DG approximations using the upwind flux (4.1) and the modified flux (5.2) as $\varepsilon \to 0$. The solution to the diffusion limit is $\varphi(\mathbf{x}) = \prod_{i=1}^{d}\cos\left(\tfrac{\pi x_i}{2}\right)$, with $d = 2$.

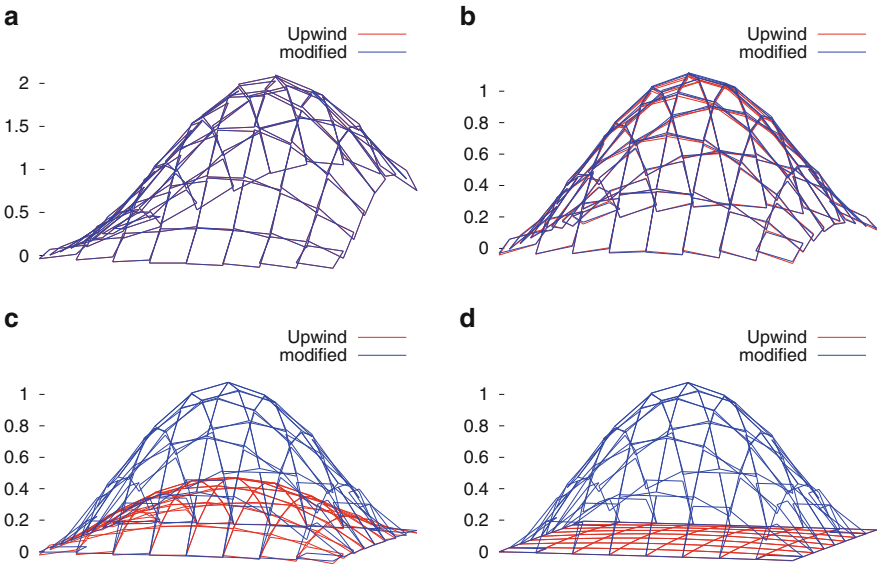


FIG. 1. Comparison of the solutions with upwind and modified flux with quadrangular $\mathbb{P}_1$ finite elements, respectively. As the scattering cross section increases, the upwind flux solution locks, while the other converges to the correct diffusion limit (**a**) $\varepsilon = 1$. (**b**) $\varepsilon = 2^{-6}$. (**c**) $\varepsilon = 2^{-10}$. (**d**) $\varepsilon = 2^{-14}$

We use piecewise linear polynomials in space, and we choose γ and δ as in (5.1) with $\gamma_0 = 4$ and $\delta_0 = 1$. The results computed on a quadrangular mesh composed of 64 cells are shown in Fig. 1 for $\varepsilon = 1, 2^{-6}, 2^{-10}$, and 2^{-14}. We observe that the solution obtained with the upwind flux locks when $\varepsilon \to 0$, whereas the solution computed with the modified flux converges to the correct diffusion limit as expected.

Acknowledgement. This material is based upon work supported by the Department of Homeland Security under agreement 2008-DN-077-ARI018-02, National Science Foundation grants DMS-0713829, DMS-0810387, and CBET-0736202, and is partially supported by award KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST).

REFERENCES

[1] M. L. Adams. Discontinuous finite element methods in thick diffusive problems. *Nucl. Sci. Eng.*, 137:298–333, 2001.
[2] B. Ayuso and L. D. Marini. Discontinuous Galerkin methods for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, 47(2):1391–1420, 2009.
[3] I. Babuška and M. Suri. On locking and robustness in the finite element method. *SIAM J. Numer. Anal.*, 29(5):1261–1293, 1992.
[4] S. Chandrasekhar. *Radiative Transfer*. Oxford University Press, 1950.

[5] R. Dautray and J.-L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology. Vol. 5. Evolution problems, I.* Springer-Verlag, Berlin, Germany, 1992. ISBN 3-540-50205-X; 3-540-66101-8.

[6] A. Ern and J.-L. Guermond. Discontinuous Galerkin methods for Friedrichs' systems. I. General theory. *SIAM J. Numer. Anal.*, 44(2):753–778, 2006.

[7] J.-L. Guermond and G. Kanschat. Asymptotic analysis of upwind discontinuous Galerkin approximation of the radiative transport equation in the diffusive limit. *SIAM J. Numer. Anal.*, 48(1):53–78, 2010.

[8] E. W. Larsen. Deterministic transport methods. Technical Report LA-9533-PR, Los Alamos Scientific Laboratory, Los Alamos, NM, 1982.

[9] E. W. Larsen. On numerical solutions of transport problems in the diffusion limit. *Nucl. Sci. Engr.*, 83:90–99, 1983.

[10] E. W. Larsen and J. B. Keller. Asymptotic solution of neutron transport problems for small mean free paths. *J. Mathematical Phys.*, 15:75–81, 1974.

[11] E. W. Larsen and J. E. Morel. Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes. II. *J. Comput. Phys.*, 83(1):212–236, 1989.

[12] E. W. Larsen, J. E. Morel, and W. F. Miller, Jr. Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes. *J. Comput. Phys.*, 69 (2):283–324, 1987.

[13] P. Lesaint and P.-A. Raviart. On a finite element method for solving the neutron transport equation. In *Mathematical Aspects of Finite Elements in Partial Differential Equations*, pages 89–123. Publication No. 33. Math. Res. Center, Univ. of Wisconsin-Madison, Academic Press, New York, 1974.

[14] F. Malvagi and G. C. Pomraning. Initial and boundary conditions for diffusive linear transport problems. *J. Math. Phys.*, 32(3):805–820, 1991.

[15] J. C. Ragusa, J.-L. Guermond, and G. Kanschat. A robust S_N-DG-approximation for radiation transport in optically thick and diffusive regimes. *J. Comput. Phys.*, 231(4):1947–1962, 2012.

[16] W. Reed and T. Hill. Triangular mesh methods for the neutron transport equation. Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.

ERROR CONTROL FOR DISCONTINUOUS GALERKIN METHODS FOR FIRST ORDER HYPERBOLIC PROBLEMS

EMMANUIL H. GEORGOULIS*, EDWARD HALL†, AND CHARALAMBOS MAKRIDAKIS‡

Abstract. An a posteriori error bound for a first order linear hyperbolic problem, with constant advection coefficient, discretized by the discontinuous Galerkin method is presented. The bound is derived using a suitable reconstruction framework, but it is essentially of residual-type. For simplicity, the special case of the mesh having one characteristic face per element is the focus of discussion, although some comments on possible extensions to general meshes are given. Numerical experiments verify the reliability and the efficiency of the estimator.

Key words. A posteriori error bounds, Discontinous Galerkin method, Hyperbolic problem

1. Introduction. Discontinuous Galerkin (dG) methods for advection problems have gained considerable popularity in the literature since their introduction in 1971 by Reed and Hill [32] (see, e.g., [5, 8, 11, 13–15, 19], the volume [12] and the references therein.) The a priori error for the dG method has been considered in [5, 22, 26, 29, 33]. In [31], it was shown numerically that the dG method suffers from slightly suboptimal rates of convergence with respect to the mesh parameter when the error is measured in the L^2-norm. Optimal error bounds in the L^2-norm for various classes of structured meshes have been shown in [9, 10, 33]. Recently, Burman [6] proposed an a posteriori bound for the case of constant elements under a saturation assumption. Other works dealing with error control of various types for first order hyperbolic problems include [2, 3, 24, 34]. Overall, it seems that this rather interesting problem requires further study.

A posteriori error bounds for dG methods for elliptic problems have been considered in [4, 7, 17, 18, 21, 27, 28]. Such bounds are based on suitable recoveries/post-processing theoretical tools of the dG solution.

In this short note, we present some recent results regarding the a posteriori error analysis of the classical dG method of Reed and Hill [32] for a

*Department of Mathematics, University of Leicester, University Road, Leicester LE1 7RH, UK, Georgoulis@le.ac.uk

†School of Mathematical Sciences University of Nottingham, University Park, Nottingham, NG7 2RD, UK, edward.hall@nottingham.ac.uk

‡Department of Applied Mathematics, University of Crete, L. Knosou GR 71409, Heraklion-Crete, Greece.
Current Address: School of Mathematical and Physical Sciences, University of Sussex, BN1 9QH, UK, makr@tem.uoc.gr

X. Feng et al. (eds.), *Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations*, The IMA Volumes in Mathematics and its Applications 157, DOI 10.1007/978-3-319-01818-8_8,

scalar first order linear hyperbolic problem. For simplicity, the special case of a two-dimensional computational domain with constant advection fields is considered. We first consider triangular meshes having one characteristic face per element, where the analysis is simpler. The analysis will be extended to the general case in a forthcoming work; here, we briefly discuss the main ideas for this case also.

The rest of this work is structured as follows. In Sect. 2, we describe the model problem considered, along with its discretization by the discontinuous Galerkin method. Section 3 is devoted to the derivation of energy norm a posteriori bounds for the discontinuous Galerkin method, under the assumption of a mesh containing one characteristic face per element. A brief discussion on relaxing the characteristic face assumption is given in Sect. 4. We conclude with some numerical experiments in Sect. 5.

2. The Problem and Its Discretization. We start by assuming the notion of a Sobolev space $W_p^k(\omega)$, based on the Lebesgue space $L^p(\omega)$, for some open domain $\omega \subset \mathbb{R}^d$, $d = 1, 2$ (for more on Sobolev spaces, see, e.g., [1]). We shall also denote the Hilbertian Sobolev spaces by $H^k(\omega) := W_2^k(\omega)$.

In $\Omega \subset \mathbb{R}^2$, we consider the first order Cauchy problem

$$\mathcal{L}_0 u \equiv b \cdot \nabla u + cu = f \quad \text{in } \Omega, \tag{2.1}$$

$$u = g \quad \text{on } \partial_- \Omega, \tag{2.2}$$

where

$$\partial_- \Omega := \{x \in \partial\Omega : b(x) \cdot n(x) < 0\}$$

is the inflow part of the domain boundary $\partial\Omega$, $b := (b_1, b_2) \in \mathbb{R}^2$ and $g \in L^2(\partial_- \Omega)$.

We assume further that there exists a positive constant γ_0 such that

$$c(x) - \frac{1}{2}\nabla \cdot b(x) \geq \gamma_0 \quad \text{for almost every } x \in \Omega, \tag{2.3}$$

and we define $c_0 := (c - 1/2\nabla \cdot b)^{1/2}$.

The discontinuous Galerkin method. We consider a mesh $\mathcal{T}$ of the domain Ω into shape-regular triangular elements $T \in \mathcal{T}$. We define

$$\partial_- T := \{x \in \partial T : b(x) \cdot n(x) < 0\}, \quad \partial_+ T := \{x \in \partial T : b(x) \cdot n(x) > 0\},$$
$$\partial_0 T := \{x \in \partial T : b(x) \cdot n(x) = 0\}$$

for each element T; we call these the *inflow*, *outflow* and *characteristic* parts of ∂T, respectively. Note that for each element $T \in \mathcal{T}$, at least one face is inflow and one is outflow; $\partial_0 T$ can, in general, be of one-dimensional Lebesgue measure zero.

For $T \in \mathcal{T}$, and a (possibly discontinuous) element-wise smooth function v, we consider the *upwind jump* across the inflow boundary $\partial_- T$, by

$$\lfloor v \rfloor(x) := \lim_{t \to 0^+} \big(u(x + tb) - u(x - tb)\big),$$

for almost all $x \in \partial_- T$, when $\partial_- T \subset \Gamma_{\text{int}}$, and by $\lfloor v \rfloor(x) := v(x)$ for almost all $x \in \partial_- T$, when $\partial_- T \subset \partial_- \Omega$. Let also $\Gamma := \cup_{T \in \mathcal{T}} \partial T$ be the *skeleton* of the mesh (i.e., the union of all one-dimensional element faces). Let also $\Gamma_{\text{int}} := \Gamma \backslash \partial\Omega$, so that $\Gamma = \partial\Omega \cup \Gamma_{\text{int}}$.

We define the discontinuous Galerkin finite element space by

$$S_h := \{w_h \in L^2(\Omega) : w_h|_T \in \mathcal{P}_p(T),\ T \in \mathcal{T}\},$$

with $\mathcal{P}_p(T)$ denoting the space of polynomials of total degree p on T.

We also define the space $\mathcal{S} := G_b + S_h$, where

$$G_b := \{w \in L^2(\Omega) : b \cdot \nabla w \in L^2(\Omega)\},$$

is the graph space of the PDE (2.1) [16, 25].

We require some more notation to describe the method. Let $u \in \mathcal{S}$; then, for every element $T \in \mathcal{T}$, we denote by u_T^+ the trace of u on ∂T taken from within the element T (interior trace). We also define the exterior trace u_T^- of $u \in \mathcal{S}$ for almost all $x \in \partial_- T \backslash \partial_- \Omega$ to be the interior trace $u_{T'}^+$ of u on the element(s) T' that share the edges contained in $\partial_- T \backslash \partial_- \Omega$ of the boundary of element T. Then, the *jump* of u across $\partial_- T \backslash \partial_- \Omega$ is defined by

$$\lfloor u \rfloor_T := u_T^+ - u_T^-;$$

the subscripts will be suppressed when no confusion is likely to occur.

Setting

$$\begin{aligned} B(u, v) &:= \sum_{T \in \mathcal{T}} \int_T (\mathcal{L}_0 u) v \, \mathrm{d}x - \sum_{T \in \mathcal{T}} \int_{\partial_- T \cap \partial_- \Omega} (b \cdot n) u^+ v^+ \, \mathrm{d}s \\ &\quad - \sum_{T \in \mathcal{T}} \int_{\partial_- T \backslash \partial_- \Omega} (b \cdot n) \lfloor u \rfloor v^+ \, \mathrm{d}s, \qquad (2.4) \\ l(v) &:= \sum_{T \in \mathcal{T}} \int_T f v \, \mathrm{d}x - \sum_{T \in \mathcal{T}} \int_{\partial_- T \cap \partial_- \Omega} (b \cdot n) g v^+ \, \mathrm{d}s, \end{aligned}$$

the discontinuous Galerkin method for the problem (2.1) then reads:

$$\text{Find } u_h \in S_h \text{ such that } B(u_h, v_h) = l(v_h) \quad \forall v_h \in S_h. \qquad (2.5)$$

A natural error notion is the *energy norm* defined by

$$|||w||| := \left(\|c_0 w\|_\Omega^2 + \frac{1}{2} \|\beta \lfloor w \rfloor\|_{\Gamma_{\text{int}}}^2 + \frac{1}{2} \|\beta w^+\|_{\partial\Omega}^2 \right)^{1/2},$$

where $\beta(x) := \sqrt{|b(x) \cdot n(x)|}$, with n on ∂T denoting the outward normal to ∂T. The choice of the above energy norm is related to the coercivity of the bilinear form $B(\cdot, \cdot)$. Indeed, we have

$$|||w|||^2 = B(w, w), \tag{2.6}$$

for all $w \in \mathcal{S}$ (see, e.g., [23] for details). Standard error analysis [5, 22, 26, 29, 33] yields the a priori error bound

$$|||u - u_h||| \leq Ch^{\min\{p+1,r\}-1/2}|u|_{H^r(\Omega)}, \tag{2.7}$$

for $p \geq 0$ and $r \geq 1$. In general, we do not observe improved rate of convergence in the (weaker) L^2-norm (see, e.g., [31]). Optimal L^2-bounds have been shown for a variety of structured meshes [9, 10, 33].

3. A Posteriori Error Bounds. We begin by assuming that one face per element is characteristic, i.e., there is one face e per element $T \in \mathcal{T}$ such that $e = \partial_0 T$, (resulting to $b \cdot n = 0$ for that face). Our analysis is based on the introduction of an intermediate function, which we call reconstruction in the spirit of [30]. The definition of this function, and hence our analysis, is simpler when one face per element is characteristic. The more involved general case is briefly discussed in the next section.

We begin by defining the reconstruction space $\tilde{S}_h$ by

$$\tilde{S}_h := \{w_h \in C(\bar{\Omega} \setminus \cup_{T\in\mathcal{T}} \partial_0 T) : w_h|_T \in \mathcal{P}_{p+1}(T)\}. \tag{3.1}$$

Definition 3.1. *We define the reconstruction $\hat{u}|_T \in \mathcal{P}_{p+1}(T)$ element-wise by the relations*

$$\int_T b \cdot \nabla \hat{u}\, v_h \,\mathrm{d}x = \int_T b \cdot \nabla u_h\, v_h \,\mathrm{d}x - \int_{\partial_- T} (b \cdot n) \lfloor u_h \rfloor v_h^+ \,\mathrm{d}s, \tag{3.2}$$

for all $v_h \in S_h$ and we set

$$\pi_p \hat{u} = u_h^- \tag{3.3}$$

on the unique inflow face, where π_p is the L^2-projection onto the element face. Finally, we require that $\hat{u}$ is continuous at the "inflow" element vertex (i.e., the only vertex that is not on the boundary of the characteristic face) with

$$\hat{u}(x_-) = (\mathrm{card}\{e : e \cap \partial_+ T \neq \emptyset\})^{-1} \sum_{e: e\cap\partial_+ T\neq\emptyset} u_h(x_-)|_e, \tag{3.4}$$

with card *being the cardinality of a set. Here, the face e is assumed* not *to contain its zero-dimensional boundary. Also, we define $\hat{u} : \Omega \to \mathbb{R}$ to be the function equal to the reconstruction $\hat{u}$ on each $T \in \mathcal{T}$.*

Remark 3.1. *The construction of the nodal values takes into account only the value down-wind, so it is local by nature.*

LEMMA 3.1. *$\hat{u} \in \tilde{S}_h$ and it is well defined.*

Proof. As $\hat{u}$ is continuous on the inflow element vertices by construction, it will suffice to show that the remaining p faces degrees of freedom coincide across each element face. The conditions $\pi_p \hat{u} = u_h^-$ on the inflow face will imply the result if $\pi_p \hat{u} = u_h^+$ on the (unique) outflow face.

The reconstruction relation (3.2) implies

$$-\int_T b\cdot\nabla v_h\,(\hat{u}-u_h)\,\mathrm{d}x+\int_{\partial T}(b\cdot n)(\pi_p\hat{u}-u_h^+)v_h^+\,\mathrm{d}s=-\int_{\partial_- T}(b\cdot n)\lfloor u_h\rfloor v_h^+\,\mathrm{d}s, \tag{3.5}$$

which, in turn gives

$$-\int_T b\cdot\nabla v_h\,(\hat{u}-u_h)\,\mathrm{d}x+\int_{\partial_+ T}(b\cdot n)(\pi_p\hat{u}-u_h^+)v_h^+\,\mathrm{d}s=0, \tag{3.6}$$

using the conditions $\pi_p \hat{u} = u_h^-$. Upon considering $v_h \in S_h$ such that $b \cdot \nabla v_h = 0$ on T, we deduce

$$\int_{\partial_+ T}(b\cdot n)(\pi_p\hat{u}-u_h^+)v_h^+\,\mathrm{d}s=0,$$

which, readily implies $\pi_p \hat{u} = u_h^+$ on the outflow face, as the outflow face is not characteristic. The proof that the reconstruction is well defined (i.e., uniquely determined, given u_h) is given in detail in [20]. □

LEMMA 3.2 ([20]). *We have*

$$\|\hat{u}-u_h\| \le C\|\sqrt{(b\cdot n)h}[u_h]\|_{\Gamma_-},\ \text{and}\ \|\hat{u}-u_h\|_{\Gamma_-},\le C\|\sqrt{(b\cdot n)}[u_h]\|_{\Gamma_-},$$

We are now ready to show an a posteriori bound in the energy norm for the dG method.

THEOREM 3.1. *Let u be the solution of (2.1), u_h its approximation by the dG method (2.5) and let $\hat{u}$ as in Definition 3.1. Then, we have the following bound*

$$\begin{aligned}|||u-u_h|||^2 \le & C\|\sqrt{b\cdot n}[u_h]\|_{\Gamma_-}^2+C\sum_{T\in\mathcal{T}}\Big(\|\beta(g-u_h^+)\|_{\partial_- T\cap\partial_-\Omega}^2\\ &+\|f-cu_h-\Pi_p(f-cu_h)\|_T^2\Big).\end{aligned} \tag{3.7}$$

with $\Pi_p : L^2(\Omega) \to \mathbb{R}$ denoting the (local) orthogonal L^2-projection onto S_h, where $C > 0$ is independent of u_h, u, $\hat{u}$, h and $\mathcal{T}$.

Proof. Let $\rho := u - \hat{u}$. We have, respectively,

$$\begin{aligned}
&\sum_{T\in\mathcal{T}} \int_T (b\cdot\nabla\rho + c(u-u_h))\,\rho\,\mathrm{d}x \\
&= \sum_{T\in\mathcal{T}} \int_T (f - cu_h)\,\rho\,\mathrm{d}x - \sum_{T\in\mathcal{T}} \int_T b\cdot\nabla\hat{u}\,\Pi_p\rho\,\mathrm{d}x \\
&\quad - \sum_{T\in\mathcal{T}} \int_T cu_h\,(\rho - \Pi_p\rho)\,\mathrm{d}x \\
&= \sum_{T\in\mathcal{T}} \int_T \big((f-cu_h) - \Pi_p(f-cu_h)\big)\,\rho\,\mathrm{d}x,
\end{aligned} \tag{3.8}$$

from (3.2), with $\Pi_p : L^2(\Omega) \to \mathbb{R}$ denoting the orthogonal L^2-projection onto S_h. This implies

$$\begin{aligned}
|||\rho|||^2 &= \sum_{T\in\mathcal{T}} \int_T (b\cdot\nabla\rho + c\rho)\,\rho\,\mathrm{d}x \\
&= \sum_{T\in\mathcal{T}} \int_T \big((f-cu_h) - \Pi_p(f-cu_h)\big)\,\rho\,\mathrm{d}x - \sum_{T\in\mathcal{T}} \int_T c(\hat{u}-u_h)\,\rho\,\mathrm{d}x.
\end{aligned} \tag{3.9}$$

The result already follows by Cauchy–Schwarz inequality, the triangle inequality and the previous lemma. □

4. The General Case. We continue by assuming that no element face is characteristic, i.e.,

$$|b(x)\cdot n(x)| > 0 \quad \text{for all} \quad x \in \partial\tau,\ T \in \mathcal{T}. \tag{4.1}$$

Note that (4.1) implies that for each element $T \in \mathcal{T}$, either one or two whole faces are contained in $\partial_- T$.

The reconstruction space $\tilde{S}_h$ is now of higher order, defined by

$$\tilde{S}_h := \{w_h \in C(\bar{\Omega}) : w_h|_T \in \mathcal{P}_{p+2}(T)\}. \tag{4.2}$$

The reconstruction can be defined along the same lines as before, but its construction is more involved.

Definition 4.1. *We define the reconstruction $\hat{u}|_T \in \mathcal{P}_{p+2}(T)$ element-wise by the relations*

$$\int_T b\cdot\nabla\hat{u}\,v_h\,\mathrm{d}x = \int_T b\cdot\nabla u_h\,v_h\,\mathrm{d}x - \int_{\partial_- T} (b\cdot n)\lfloor u_h\rfloor v_h^+\,\mathrm{d}s, \tag{4.3}$$

for all $v_h \in S_h$ and we set

$$\pi_p\hat{u} = u_h^- \tag{4.4}$$

on each inflow face (with $u_h^- = g$ on $\partial_-\Omega$), if the element T has two inflow boundary faces, where π_p is the L^2-projection onto the element face, or $\pi_p\hat{u} = u_h^-$ and

$$\pi_p\hat{u} = u_h^+ \tag{4.5}$$

on the unique inflow and one of the two outflow faces, if the element T has two outflow faces. Finally, we require that $\hat{u}$ is continuous at the element vertices with

$$\hat{u}(x_i) = (\mathrm{card}\{e : e \cap \partial_+ T \neq \emptyset\})^{-1} \sum_{e: e\cap\partial_+T\neq\emptyset} u_h(x_i)|_e, \tag{4.6}$$

with x_i, $i = 1, \ldots, d+1$ denoting the vertices of T, $e \subset \partial T$ a (generic) element face and card *being the cardinality of a set. Here, the face e is assumed* not *to contain its zero-dimensional boundary. Also, we define $\hat{u} : \Omega \to \mathbb{R}$ to be the function equal to the reconstruction $\hat{u}$ on each $T \in \mathcal{T}$.*

Using this reconstruction, one can show the following a posteriori bound [20].

THEOREM 4.1. *Let u be the solution of (2.1), u_h its approximation by the dG method (2.5) and let $\hat{u}$ as in Definition 3.1. Then, we have the following bound*

$$\begin{aligned} |||u - u_h|||^2 \leq & C\|\sqrt{b\cdot n}[u_h]\|^2_{\Gamma_-} + C\sum_{T\in\mathcal{T}} \Big(\|\beta(g - u_h^+)\|^2_{\partial_- T\cap\partial_-\Omega} \\ & + \|f - b\cdot\nabla\hat{u} - cu_h - \Pi_p(f - b\cdot\nabla\hat{u} - cu_h)\|^2_T \Big). \end{aligned} \tag{4.7}$$

where $C > 0$ is independent of u_h, u, $\hat{u}$, h and $\mathcal{T}$.

5. Numerical Experiments. We present some numerical experiments whereby the a posteriori bound (3.7) is used within a mesh-adaptive algorithm. In all experiments we use a fixed fraction refinement strategy with 25 % refinement and 0 % coarsening. After each step of the adaptive strategy a green refinement is carried out to ensure no hanging nodes are present in the resultant mesh.

Example 1. In this example, we consider a problem with rough solution, constant advection field, whereby the elements in the mesh have each a characteristic face. In particular, we consider the Cauchy problem (2.1), (2.2) in $\Omega = [0,1]^2$, $b = (1,1)^T$ and $c = 1$. In this case the boundary condition g and right-hand side f are chosen so that the exact solution is $u = \tanh(100(x + y - 0.5))$ which has an internal layer centred on the line $y = -x + 0.5$. Experiments begin on an initial uniform grid comprising 128 right angled triangles with positive slope so that every element (except those formed by a green refinement) has an edge which is characteristic (Figs. 1 and 2; Table 1).

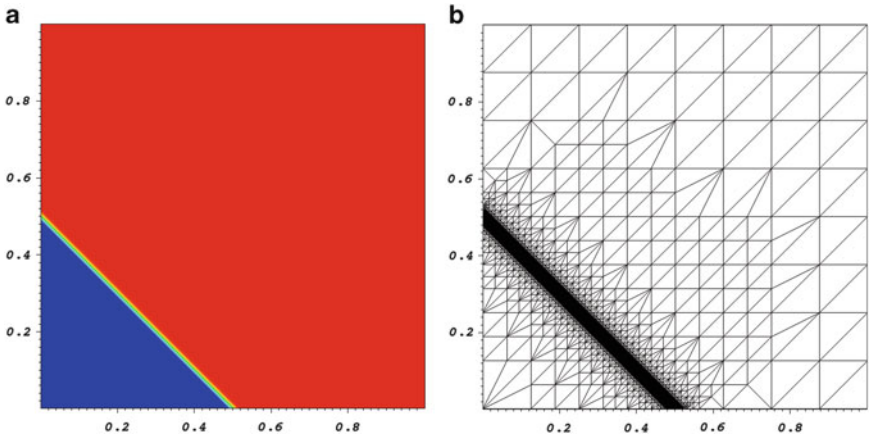


FIG. 1. Example 1. (**a**) Solution, (**b**) mesh after seven refinement steps

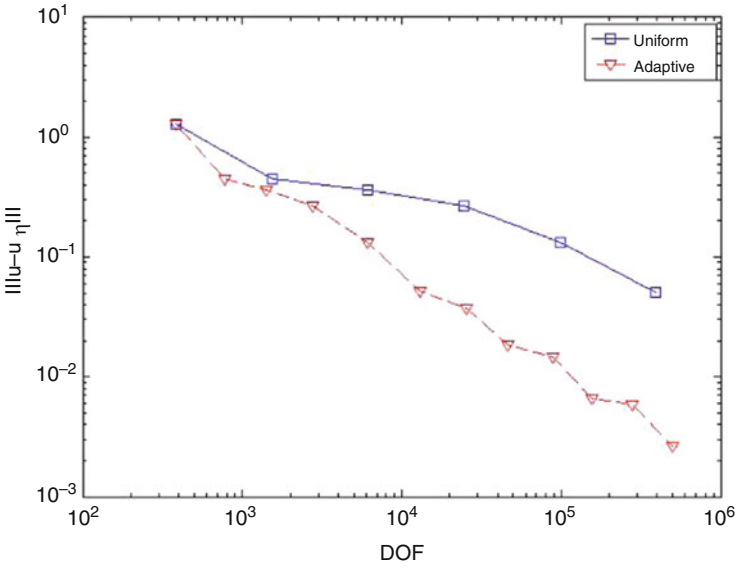


FIG. 2. Example 1. Error convergence

Table 1
Example 1. Error convergence and effectivities

DOF	$\|\|\|u-u_h\|\|\|$	Error indicator	Effectivity
384	1.2996	20.795	16.00
783	4.5209E−1	13.158	29.11
1,413	3.6374E−1	8.8324	24.28
2,778	2.6582E−1	4.3646	16.42
6,093	1.3230E−1	1.2117	9.16
13,035	5.1778E−2	0.3.4670E−1	6.70
25,578	3.7128E−2	1.6754E−1	4.51
46,116	1.8479E−2	8.3647E−2	4.53
88,827	1.4568E−2	4.4617E−2	3.06
155,433	6.6246E−3	2.4235E−2	3.66
279,162	5.8329E−3	1.4263E−2	2.45
500,859	2.6474E−3	8.0287E−3	3.03

Example 2. We consider the same problem as in Example 1, except $b = (-1,1)^T$, but again g and f are chosen so that the exact solution is $u = \tanh(100(x+y-0.5))$. The same initial mesh is used from Example 1, so that *no* elements have edges which are characteristic (Figs. 3 and 4; Table 2).

Example 3. In our final example, we let $r = (x^2+y^2)^{1/2}$, $\theta = \tan^{-1}(y/x)$ and consider the Cauchy problem (2.1), (2.2) in $\Omega = [0,1]^2$, $b = (-r\sin(\theta),$
$r\cos(\theta))^T$ and $c = \mathrm{e}^r$. g and f are chosen so that the exact solution is given by $\tanh(200(r-0.75))$ (Figs. 5 and 6; Table 3).

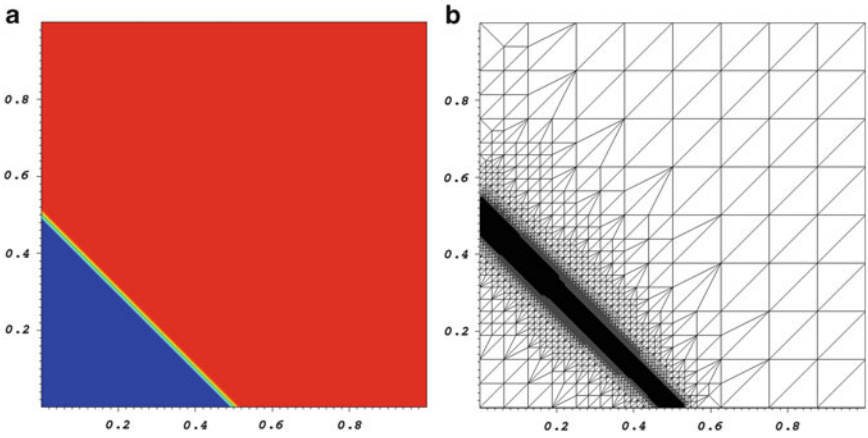


Fig. 3. Example 2. (**a**) Solution, (**b**) mesh after six refinement steps

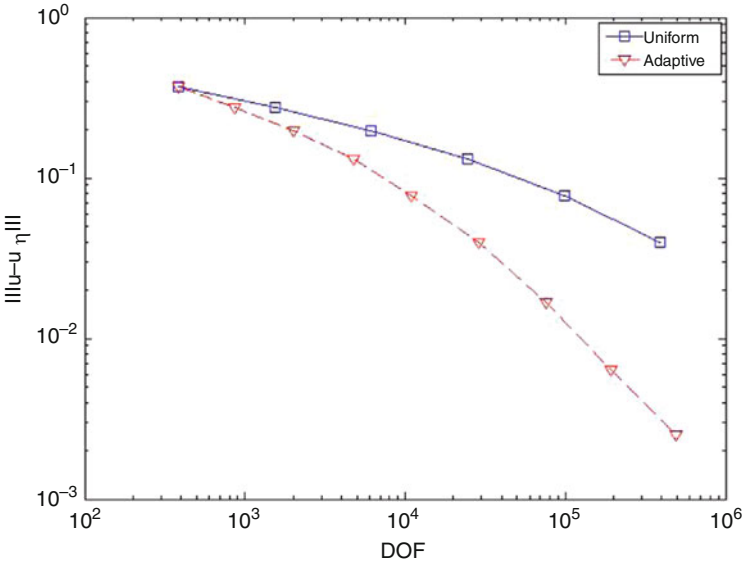


FIG. 4. Example 2. Error convergence

TABLE 2
Example 2. Error convergence and effectivities

DOF	$\|\|\|u-u_h\|\|\|$	Error indicator	Effectivity
384	3.6771E−1	3.9094E−1	1.06
864	2.7572E−1	2.9408E−1	1.07
2,019	1.9688E−1	2.1425E−1	1.09
4,785	1.3058E−1	1.4728E−1	1.13
10,923	7.7532E−2	9.2398E−2	1.19
28,797	3.9397E−2	5.0271E−2	1.28
75,450	1.6817E−2	2.2800E−2	1.36
191,673	6.3610E−3	8.8995E−3	1.40
491,598	2.5283E−3	3.5585E−3	1.41

REFERENCES

[1] R. A. ADAMS AND J. J. F. FOURNIER, *Sobolev spaces*, vol. 140 of Pure and Applied Mathematics (Amsterdam), Elsevier/Academic Press, Amsterdam, second ed., 2003.

[2] S. ADJERID AND T. C. MASSEY, *A posteriori discontinuous finite element error estimation for two-dimensional hyperbolic problems*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 5877–5897.

[3] R. BECKER, D. CAPATINA, AND R. LUCE, *Reconstruction-based a posteriori error estimators for the transport equation*, in Numerical Mathematics and Advanced Applications 2011: Proceedings of ENUMATH 2011, Leicester, September 2011, Springer, 2013.

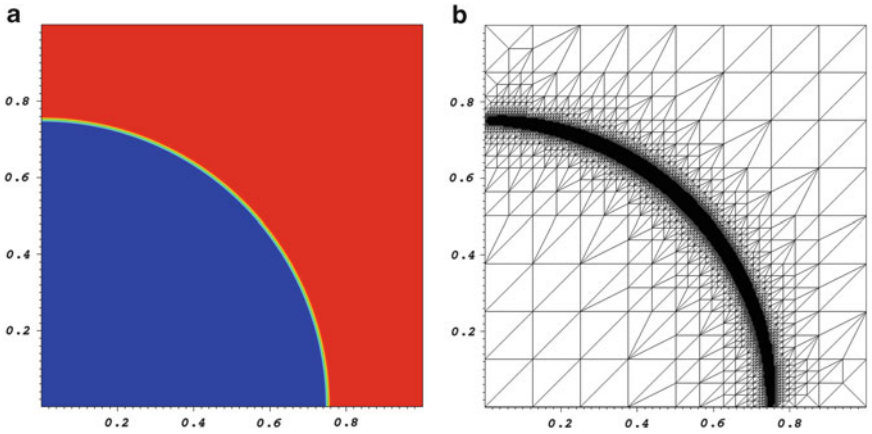


FIG. 5. Example 3. (**a**) Solution, (**b**) mesh after seven refinement steps

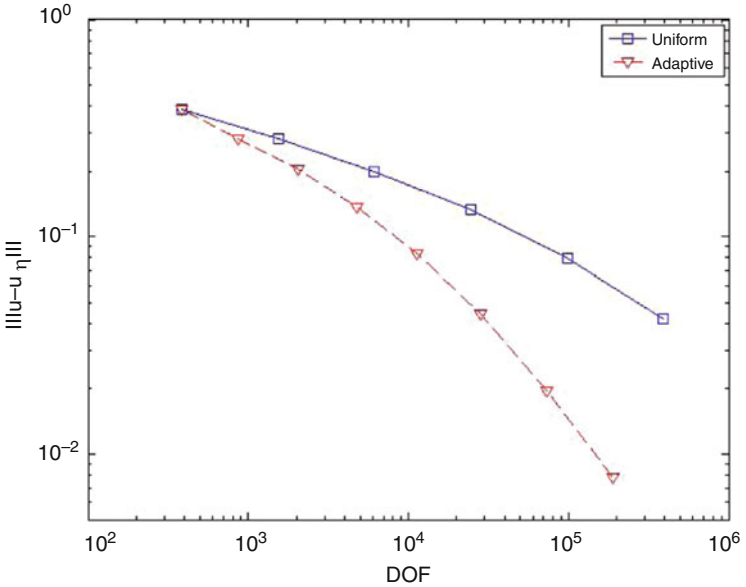


FIG. 6. Example 3. Error convergence

[4] R. BECKER, P. HANSBO, AND M. G. LARSON, *Energy norm a posteriori error estimation for discontinuous Galerkin methods*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 723–733.

[5] K. S. BEY AND T. ODEN, *hp-version discontinuous Galerkin methods for hyperbolic conservation laws*, Comput. Methods Appl. Mech. Engrg., 133 (1996), pp. 259–286.

[6] E. BURMAN, *A posteriori error estimation for interior penalty finite element approximations of the advection-reaction equation*, SIAM J. Numer. Anal., 47 (2009), pp. 3584–3607.

Table 3
Example 3. Error convergence and effectivities

DOF	$\|\|\|u - u_h\|\|\|$	Error indicator	Effectivity
384	3.8544E−1	4.4612E−1	1.16
858	2.8452E−1	3.1540E−1	1.11
2,034	2.0374E−1	2.1831E−1	1.07
4,764	1.3710E−1	1.4379E−1	1.05
11,217	8.3927E−2	9.1930E−2	1.10
28,353	4.4468E−2	5.3819E−2	1.21
73,212	1.9684E−2	2.5946E−2	1.32
190,656	7.7997E−3	1.0708E−2	1.37

[7] C. Carstensen, T. Gudi, and M. Jensen, *A unifying theory of a posteriori error control for discontinuous Galerkin FEM*, Numer. Math., 112 (2009), pp. 363–379.
[8] B. Cockburn, *Discontinuous Galerkin methods for convection-dominated problems*, in High-order methods for computational physics, Springer, Berlin, 1999, pp. 69–224.
[9] B. Cockburn, B. Dong, and J. Guzmán, *Optimal convergence of the original DG method for the transport-reaction equation on special meshes*, SIAM J. Numer. Anal., 46 (2008), pp. 1250–1265.
[10] B. Cockburn, B. Dong, J. Guzmán, and J. Qian, *Optimal convergence of the original DG method on special meshes for variable transport velocity*, SIAM J. Numer. Anal., 48 (2010), pp. 133–146.
[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 (1990), pp. 545–581.
[12] B. Cockburn, G. E. Karniadakis, and C.-W. Shu, eds., *Discontinuous Galerkin methods*, Springer-Verlag, Berlin, 2000. Theory, computation and applications, Papers from the 1st International Symposium held in Newport, RI, May 24–26, 1999.
[13] 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 (1989), pp. 90–113.
[14] 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 (1989), pp. 411–435.
[15] ———, *The Runge-Kutta discontinuous Galerkin method for conservation laws. V. Multidimensional systems*, J. Comput. Phys., 141 (1998), pp. 199–224.
[16] A. Ern and J.-L. Guermond, *Discontinuous Galerkin methods for Friedrichs' systems. I. General theory*, SIAM J. Numer. Anal., 44 (2006), pp. 753–778.
[17] A. Ern and A. F. Stephansen, *A posteriori energy-norm error estimates for advection-diffusion equations approximated by weighted interior penalty methods*, J. Comp. Math., 26 (2008), pp. 488–510.
[18] A. Ern, A. F. Stephansen, and P. Zunino, *A discontinuous Galerkin method with weighted averages for advection-diffusion equations with locally small and anisotropic diffusivity*, IMA J. Numer. Anal., 29 (2009), pp. 235–256.
[19] R. S. Falk and G. R. Richter, *Local error estimates for a finite element method for hyperbolic and convection-diffusion equations*, SIAM J. Numer. Anal., 29 (1992), pp. 730–754.
[20] E. H. Georgoulis, E. Hall, and C. Makridakis, *A posteriori error control for the RKDG method for transport problems*, to appear.

[21] P. Houston, D. Schötzau, and T. P. Wihler, *Energy norm a posteriori error estimation of hp-adaptive discontinuous Galerkin methods for elliptic problems*, Math. Models Methods Appl. Sci., 17 (2007), pp. 33–62.

[22] P. Houston, C. Schwab, and E. Süli, *Stabilized hp-finite element methods for first-order hyperbolic problems*, SIAM J. Numer. Anal., 37 (2000), pp. 1618–1643 (electronic).

[23] ———, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2002), pp. 2133–2163 (electronic).

[24] P. Houston and E. Süli, *hp-adaptive discontinuous Galerkin finite element methods for first-order hyperbolic problems*, SIAM J. Sci. Comput., 23 (2001), pp. 1226–1252.

[25] M. Jensen, *Discontinuous Galerkin methods for friedrichs systems*, D.Phil. Thesis, University of Oxford, (2005).

[26] C. Johnson and J. Pitkäranta, *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation*, Math. Comp., 46 (1986), pp. 1–26.

[27] O. A. Karakashian and F. Pascal, *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 2374–2399 (electronic).

[28] ———, *Convergence of adaptive discontinuous Galerkin approximations of second-order elliptic problems*, SIAM J. Numer. Anal., 45 (2007), pp. 641–665 (electronic).

[29] P. Lesaint and P.-A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical aspects of finite elements in partial differential equations (Proc. Sympos., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1974), Math. Res. Center, Univ. of Wisconsin-Madison, Academic Press, New York, 1974, pp. 89–123. Publication No. 33.

[30] C. Makridakis and R. H. Nochetto, *A posteriori error analysis for higher order dissipative methods for evolution problems*, Numer. Math., 104 (2006), pp. 489–514.

[31] T. E. Peterson, *A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation*, SIAM J. Numer. Anal., 28 (1991), pp. 133–140.

[32] 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).

[33] G. R. Richter, *An optimal-order error estimate for the discontinuous Galerkin method*, Math. Comp., 50 (1988), pp. 75–88.

[34] 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 (Freiburg/Littenweiler, 1997), vol. 5 of Lect. Notes Comput. Sci. Eng., Springer, Berlin, 1999, pp. 123–194.

VIRTUAL ELEMENT AND DISCONTINUOUS GALERKIN METHODS

F. BREZZI* AND L. D. MARINI†

Abstract. Virtual element methods (VEM) are the latest evolution of the Mimetic Finite Difference Method and can be considered to be more close to the Finite Element approach. They combine the ductility of mimetic finite differences for dealing with rather weird element geometries with the simplicity of implementation of Finite Elements. Moreover, they make it possible to construct quite easily high-order and high-regularity approximations (and in this respect they represent a significant improvement with respect to both FE and MFD methods). In the present paper we show that, on the other hand, they can also be used to construct DG-type approximations, although numerical tests should be done to compare the behavior of DG-VEM versus DG-FEM.

Key words. Discontinuous Galerkin, Virtual elements, Mimetic finite differences

AMS(MOS) subject classifications. 65N30, 65N12, 65G99, 76R99

1. Introduction. The aim of this paper is to present a possible way to introduce the virtual element method (VEM) in the discontinuous Galerkin (DG) framework. From several points of view VEM can be considered as the natural extension of Finite Element Methods to more general geometries and continuity requirements. Apparently, their extension to the Discontinuous Galerkin world could be seen as useless, as DG methods can already deal with rather general geometries. However, in a certain number of their applications there is some need of a conforming interpolant that for general geometries or for higher order continuity (as for plate problems, among others) will not be easily available within the usual DG framework. Here, however, to start with, we will deal with the simplest possible case, that is the discretization of the Poisson problem in two dimensions. The idea is to start understanding what are the most convenient ways to deal with Discontinuous Virtual Elements. We shall see that a direct application of the DG technology cannot be done, but some simple variants are available that still ensure uniqueness, stability, and convergence with optimal error bounds.

As a first step we will recall the basic concepts of VEM. This will be done with some details, taking into account that the introduction of VEM is quite recent, and we cannot expect many readers to be familiar with them. In the next section we will present the basic assumptions (on the element geometry, on the discrete spaces) and recall an abstract convergence result.

*IUSS-Pavia, Pavia, Italy
and
King Abdulaziz University, Jeddah, Saudi Arabia, brezzi@imati.cnr.it
†Università di Pavia, Pavia, Italy, marini@imati.cnr.it

X. Feng et al. (eds.), *Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations*, The IMA Volumes in Mathematics and its Applications 157, DOI 10.1007/978-3-319-01818-8_9,

Then we will recall the general way to construct the discrete bilinear form, in Sect. 3, and the discrete right-hand side, in Sect. 4. In Sect. 5 we will recall the classical instruments and concepts of DG formulations (in a much less detailed way, this time). The novelty of the paper will appear in Sect. 6, where VEM will be adapted to DG formulations, and in Sect. 7 where optimal error bounds will be proved.

Throughout the paper, we will follow the usual notation for Sobolev spaces and norms (see, e.g., [6]). In particular, for an open bounded domain $\mathcal{D}$, we will use $|\cdot|_{s,\mathcal{D}}$ and $\|\cdot\|_{s,\mathcal{D}}$ to denote seminorm and norm, respectively, in the Sobolev space $H^s(\mathcal{D})$, while $(\cdot,\cdot)_{0,\mathcal{D}}$ will denote the $L^2(\mathcal{D})$ inner product. Often the subscript will be omitted when $\mathcal{D}$ is the computational domain Ω. For a nonnegative integer k, the space of polynomials of degree less than or equal to k will be denoted by $\mathbb{P}_k$. Following a common convention, we will also use $\mathbb{P}_{-1} := \{0\}$.

Finally, C will be a generic constant independent of the decomposition that could change from an occurrence to the other.

2. Basic Assumptions and an Abstract Convergence Result. We first recall the general idea of continuous VEM, underlying the similarities with classical Finite Element Methods (we refer to [3] for a more detailed presentation).

For this we consider, as usual, the simplest possible problem: *find* $u \in V \equiv H^1_0(\Omega)$ *such that* $-\Delta u = f$. Written in variational form, the problem becomes

$$\text{find } u \in V \equiv H^1_0(\Omega) \text{ such that } \; a(u,v) = (f,v) \quad \forall v \in V, \tag{2.1}$$

where (as usual):

$$a(u,v) := \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}\mathbf{x}, \qquad (f,v) = \int_\Omega f\, v \, \mathrm{d}x. \tag{2.2}$$

Let $\mathcal{T}_h$ be a decomposition of Ω into polygons of almost arbitrary shape (see, as an example, Fig. 1). On $\mathcal{T}_h$ we make the following assumptions

H1 There exists an integer N and a positive real number ζ such that for every h and for every $K \in \mathcal{T}_h$:

- The number of edges of K is $\le N$,
- The ratio between the shortest edge and the diameter h_K of K is bigger than ζ, and
- K is star-shaped with respect to every point of a ball of radius ζh_K.

Remark 2.1. *We point out that from assumption* **H1** *we can easily deduce that there exists an* $s* > 3/2$*, depending on* ζ*, such that for every smooth* g *on* ∂K *and for every smooth* f *in* K *the solution* φ *of the problem* $\Delta\varphi = f$ *in* K *with* $\varphi = g$ *on* ∂K *belongs to* $H^{s*}(K)$.

Next, we fix an integer $k \ge 1$ (that will be our order of accuracy) and define for each $K \in \mathcal{T}_h$

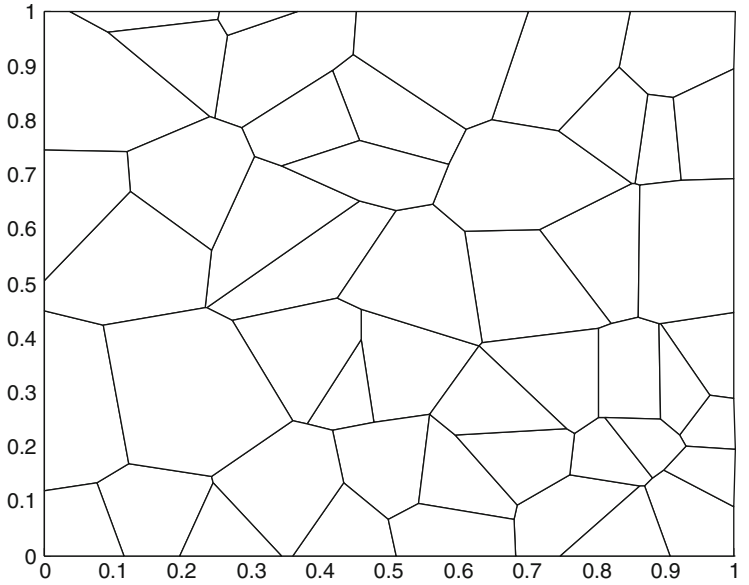


FIG. 1. Example of a Voronoi tessellation

$$V_k^K := \{v : v_{|e} \in \mathbb{P}_k(e) \ \forall \text{ edge } e \text{ of } K, \Delta v \in \mathbb{P}_{k-2}(K)\}, \tag{2.3}$$

where we recall that $\mathbb{P}_k$ denotes the space of polynomials of degree $\leq k$, and $\mathbb{P}_{-1} := \{0\}$.

Denoting by NV the number of vertices of K (obviously equal, as well, to the number of edges), the dimension of V_k^K will clearly be

$$N^K := NV + NV * (k-1) + \frac{k(k-1)}{2} = NV * k + \frac{k(k-1)}{2}$$

An element v of V_k^K can be identified by

(a) The values of v at the vertices;

(b) The moments $\displaystyle\int_e v\, p_{k-2}\, \mathrm{d}s$ on each edge e, $\quad k \geq 2$;

(c) The moments $\displaystyle\int_K v\, p_{k-2}\, \mathrm{d}\mathbf{x}$, $\quad k \geq 2$.

THEOREM 2.1. *For every $k \geq 1$ the set of degrees of freedom (a)–(c) are unisolvent for the space V_k^K.*

Proof. The number of degrees of freedom (a)–(c) equals the dimension of V_k^K. Hence we have only to check that every $v \in V_k^K$ having the d.o.f.'s equal to zero is identically zero. For this we first observe that if the degrees of freedom (a) and (b) are equal to zero, then $v = 0$ on ∂K. Remember that v, being in V_k^K, has Δv in $\mathbb{P}_{k-2}$. Hence, if the d.o.f. (c) are equal to 0, we have

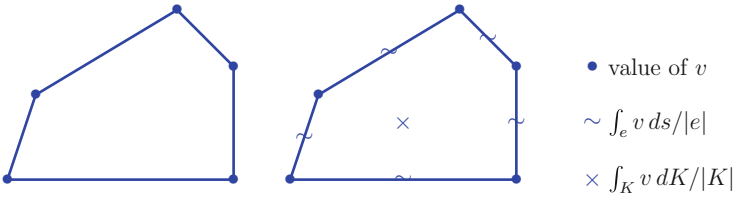


FIG. 2. Example of d.o.f. for $k = 1$ (*left*), and $k = 2$ (*right*)

$$0 = \int_K (-\Delta v) v \, \mathrm{d}\mathbf{x} = |v|^2_{1,K}$$

implying that $v \equiv 0$. □

For later use, it will be, however, more convenient to define the degrees of freedom in a more precise way. For this, for a geometric object $\mathcal{O} \subset \mathbb{R}^d$ (as an edge, a face, a d-dimensional domain, etc.) we define first its barycenter $\mathbf{x}_\mathcal{O}$ and its diameter $d_\mathcal{O}$. Then we consider, for every integer $r \geq 0$, the set $\mathcal{M}_r(\mathcal{O})$ of all *monomials*, in $\mathbb{R}^d$, of the type

$$\mathcal{M}_r(\mathcal{O}) := \left\{ \frac{(\mathbf{x} - \mathbf{x}_\mathcal{O})^\alpha}{d_\mathcal{O}^{|\alpha|}} \right\} \qquad \text{for } |\alpha| \leq r, \tag{2.4}$$

where for the multi-integer $\alpha \in \mathbb{N}^d$ we followed the usual notation

$$(x_1, \ldots, x_d)^\alpha \equiv x_1^{\alpha_1} \cdot x_2^{\alpha_2} \cdots x_d^{\alpha_d} \qquad \text{and} \qquad |\alpha| = \sum_{i=1}^{d} \alpha_i.$$

Now we can make precise the actual degrees of freedom that we want to use in V_k^K:

- The values of v at the vertices;

and for $k \geq 2$

- The moments $\displaystyle\int_e v\, m_{k-2}\, \mathrm{d}s/|e|, \quad m_{k-2} \in \mathcal{M}_{k-2}(e),$ on each edge e,
- The moments $\displaystyle\int_K v\, m_{k-2}\, \mathrm{d}\mathbf{x}/|K|, \quad m_{k-2} \in \mathcal{M}_{k-2}(K).$

Figure 2 shows an example of d.o.f for the cases $k = 1$ and $k = 2$.

For each h and for each k we then define the VEM space as

$$V_h := \{ v \in V : \ v_{|K} \in V_k^K \ \forall K \in \mathcal{T}_h \}. \tag{2.5}$$

Following [3], we need now to define an element $f_h \in V_h'$, and a bilinear form $a_h(\cdot, \cdot)$ from $V_h \times V_h$ to $\mathbb{R}$ satisfying the following assumptions:

H2 • k-consistency: for all h, and for all K in $\mathcal{T}_h$

$$\forall p \in \mathbb{P}_k,\ \forall v_h \in V_h \qquad a_h^K(p, v_h) = a^K(p, v_h). \tag{2.6}$$

- Stability: $\exists$ twopositive constants α_* and α^*, independent of h and of K, such that

$$\forall v_h \in V_h \quad \alpha_*\, a^K(v_h, v_h) \le a_h^K(v_h, v_h) \le \alpha^*\, a^K(v_h, v_h). \tag{2.7}$$

In (2.6) and (2.7) $a^K(\cdot,\cdot)$ denotes the restriction of the bilinear form $a(\cdot,\cdot)$ defined in (2.2) to the element K. We point out that, due to the symmetry of $a(\cdot,\cdot)$, (2.7) implies as well continuity:

$$\begin{aligned} a_h^K(v_h, w_h) \le\ & \Big(a_h^K(v_h, v_h)\Big)^{1/2}\Big(a_h^K(w_h, w_h)\Big)^{1/2} \\ \le\ & \alpha^*\,(a^K(v_h, v_h))^{1/2}(a^K(w_h, w_h))^{1/2} \le \alpha^*\,|v_h|_{1,K}\,|w_h|_{1,K}. \end{aligned} \tag{2.8}$$

Then, the approximate problem is, as usual,

$$\text{find } u_h \in V_h \text{ such that} \quad a_h(u_h, v_h) = \langle f_h, v_h\rangle \quad \forall v_h \in V_h. \tag{2.9}$$

The following convergence result is proved in [3].

THEOREM 2.2. *Under Assumptions* **H2**, *the discrete problem* (2.9) *has a unique solution* u_h. *Moreover, for every approximation* u_I *of* u *in* V_h, *and for every approximation* u_π *of* u *that is piecewise in* $\mathbb{P}_k$, *we have*

$$\|u - u_h\|_V \le C\Big(\|u - u_I\|_V + \|u - u_\pi\|_{h,V} + \|f - f_h\|_{V_h'}\Big),$$

where C *is a constant independent of* h *and*

$$\|f - f_h\|_{V_h'} := \sup_{v_h \in V_h} \frac{\langle f - f_h, v_h\rangle}{\|v_h\|_V}.$$

3. Construction of the Bilinear Form $a_h(u_h, v_h)$**.** First of all, we observe that the local degrees of freedom allow us to compute exactly $a^K(p, v)$ for any $p \in \mathbb{P}_k(K)$ and for any $v \in V_k^K$. Indeed, observe first that the value of each function $v \in V_h$ at the boundary of each element is known (it is a polynomial!), even when the value inside the element is not. Then consider the following integration by parts

$$a^K(p, v) = \int_K \nabla p \cdot \nabla v \mathrm{d}x = -\int_K \Delta p\, v \mathrm{d}x + \int_{\partial K} \frac{\partial p}{\partial n} v \mathrm{d}s, \tag{3.1}$$

and observe that since $\Delta p \in \mathbb{P}_{k-2}(K)$ and $\partial p/\partial n \in \mathbb{P}_{k-1}(e)$ for all $e \subset \partial K$, the last two integrals can be computed exactly knowing only the degrees of freedom associated with v (and without necessarily knowing v in the interior of K).

This allows us to define (and compute!) the (projection) operator $\Pi_k^K : V_k^K \longrightarrow \mathbb{P}_k(K) \subset V_k^K$ as follows: for all $v \in V_k^K$ we define $\Pi_k^K v$ as the solution of

$$\begin{cases} (\nabla \Pi_k^K v, \nabla q)_{0,K} = (\nabla v, \nabla q)_{0,K} \quad \forall q \in \mathbb{P}_k(K) \\ \displaystyle\int_{\partial K} \Pi_k^K v \, \mathrm{d}s = \int_{\partial K} v \, \mathrm{d}s. \end{cases} \tag{3.2}$$

We note that (3.2) clearly implies

$$\Pi_k^K q = q, \quad \forall q \in \mathbb{P}_k(K), \tag{3.3}$$

since the first equation in (3.2) tells us that q and $\Pi_k^K q$ have the same gradient, and the second equation takes care of the constant part.

At this point, we observe that choosing $a_h^K(u,v) = a^K(\Pi_k^K u, \Pi_k^K v)$ would easily ensure property (2.6). However this choice would not, in general, satisfy (2.7). Therefore we need to add a term able to ensure (2.7). Let then $S^K(u,v)$ be a symmetric positive definite bilinear form (to be chosen) that verifies

$$c_* a^K(v,v) \le S^K(v,v) \le c^* a^K(v,v) \quad \forall v \in V_k^K \quad \text{with } \Pi_k^K v = 0 \tag{3.4}$$

for some positive constants c_*, c^* independent of K and h_K. Then we set

$$a_h^K(u,v) = a^K(\Pi_k^K u, \Pi_k^K v) + S^K(u - \Pi_k^K u, v - \Pi_k^K v) \quad \forall u, v \in V_k^K. \tag{3.5}$$

THEOREM 3.1. *The bilinear form* (3.5) *satisfies the consistency property* (2.6) *and the stability property* (2.7).

Proof. Property (2.6) follows immediately from (3.3) and (3.2): for $p \in \mathbb{P}_k(K)$ (3.3) implies $S^K(p - \Pi_k^K p, v - \Pi_k^K v) = 0$. Hence, for all $v \in V_k^K$, using (3.5) and (3.2), we have

$$a_h^K(p,v) = a^K(\Pi_k^K p, \Pi_k^K v) = a^K(p,v). \tag{3.6}$$

Then we observe first that, since $a_h^K(v - \Pi_k^K v, \Pi_k^K v) \equiv 0$ for all v, we easily have

$$a_h^K(v,v) = a_h^K(\Pi_k^K v, \Pi_k^K v) + a_h^K(v - \Pi_k^K v, v - \Pi_k^K v) \quad \forall v \in V_k^K. \tag{3.7}$$

Property (2.7) now follows from (3.4) and (3.7) with $\alpha^* := \max\{1, c^*\}$ and $\alpha_* := \min\{1, c_*\}$: indeed for all $v \in V_k^K$

$$\begin{aligned} a_h^K(v,v) &\le a^K(\Pi_k^K v, \Pi_k^K v) + c^* a^K(v - \Pi_k^K v, v - \Pi_k^K v) \\ &\le \max\{1, c^*\} \Big(a^K(\Pi_k^K v, \Pi_k^K v) + a^K(v - \Pi_k^K v, v - \Pi_k^K v) \Big) \\ &= \alpha^* a^K(v,v), \end{aligned}$$

and similarly

$$\begin{aligned} a_h^K(v,v) &\ge \min\{1, c_*\} \Big(a^K(\Pi_k^K v, \Pi_k^K v) + a^K(v - \Pi_k^K v, v - \Pi_k^K v) \Big) \\ &= \alpha_* a^K(v,v). \end{aligned}$$

□

3.1. Choice of S^K. In general, the choice of the bilinear form S^K would depend on the problem and on the degrees of freedom. From (3.4) it is clear that S^K must scale like $a^K(\cdot,\cdot)$ on the kernel of Π_k^K. Denoting by χ_i, $i = 1, \ldots, N^K$ the ith d.o.f. in V_k^K, and choosing then the canonical basis $\varphi_1, \ldots, \varphi_{N^K}$ as

$$\chi_i(\varphi_j) = \delta_{ij}, \quad i, j = 1, 2, \ldots, \mathrm{N}^K, \tag{3.8}$$

the local stiffness matrix is given by

$$a_h^K(\varphi_i, \varphi_j) = a^K(\Pi_k^K \varphi_i, \Pi_k^K \varphi_j) + S^K(\varphi_i - \Pi_k^K \varphi_i, \varphi_j - \Pi_k^K \varphi_j). \tag{3.9}$$

In the present case it is easy to check that, on a "reasonable" polygon (like the ones that satisfy assumptions **H1**) we have $a^K(\varphi_i, \varphi_i) \simeq 1$. Hence, a possible choice for S^K is simply

$$S^K(\varphi_i - \Pi_k^K \varphi_i, \varphi_j - \Pi_k^K \varphi_j) = \sum_{r=1}^{N^K} \chi_r(\varphi_i - \Pi_k^K \varphi_i)\chi_r(\varphi_j - \Pi_k^K \varphi_j). \tag{3.10}$$

REMARK 3.1. *This explains why, in defining the d.o.f. in V_k^K, we used, instead of the more usual $\mathbb{P}_k$, the set $\mathcal{M}_k$. With the latter choice all the d.o.f. scale like* 1*, and this allows to choose S^K as simple as in* (3.10).

4. Construction of the Right-Hand Side. We consider first the case $k \geq 2$, and define f_h on each element K as the $L^2(K)$-projection of f onto the space $\mathbb{P}_{k-2}$, that is,

$$f_h = P_{k-2}^K f \qquad \text{on each } K \in \mathcal{T}_h.$$

Consequently, the associated right-hand side

$$\begin{aligned} \langle f_h, v_h \rangle &= \sum_{K \in \mathcal{T}_h} \int_K f_h \, v_h \, \mathrm{d}\mathbf{x} \equiv \sum_{K \in \mathcal{T}_h} \int_K (P_{k-2}^K f) \, v_h \, \mathrm{d}\mathbf{x} \\ &= \sum_{K \in \mathcal{T}_h} \int_K f \, (P_{k-2}^K v_h) \, \mathrm{d}\mathbf{x} \end{aligned}$$

can be exactly computed using the degrees of freedom for V_h that represent the internal moments. Then, standard L^2-orthogonality and approximation estimates on star-shaped domains yield

$$\begin{aligned}\langle f_h, v_h\rangle - (f, v_h) &= \sum_{K\in\mathcal{T}_h} \int_K (P^K_{k-2} f - f)\, v_h \,\mathrm{d}\mathbf{x} \\ &= \sum_{K\in\mathcal{T}_h} \int_K (P^K_{k-2} f - f)(v_h - P^K_0 v_h)\,\mathrm{d}\mathbf{x} \\ &\le C \sum_{K\in\mathcal{T}_h} h_K^{k-1} |f|_{k-1,K}\ h_K |v_h|_{1,K} \\ &\le C\, h^k \left(\sum_{K\in\mathcal{T}_h} |f|^2_{k-1,K} \right)^{1/2} |v_h|_1,\end{aligned} \tag{4.1}$$

and thus,

$$\|f - f_h\|_{V'} \le C h^k \Big(\sum_{K\in\mathcal{T}_h} |f|^2_{k-1,K} \Big)^{1/2}. \tag{4.2}$$

For the case $k = 1$ we can first, on each element K, define $\overline{v_h}$ as

$$\overline{v_h} := \frac{1}{|\partial K|} \int_{\partial K} v_h \,\mathrm{d}s$$

and then define

$$\langle f_h, v_h\rangle := \sum_{K\in\mathcal{T}_h} \int_K f\, \overline{v_h}\,\mathrm{d}\mathbf{x}$$

to obtain

$$\langle f_h, v_h\rangle - (f, v_h) = \sum_{K\in\mathcal{T}_h} (f, \overline{v_h} - v_h)_{0,K} \le C\, h \|f\|_{0,\Omega}\, |v_h|_{1,\Omega}.$$

5. Basic Concepts of DG Methods. The extension of what we have presented in the previous sections to DG is almost straightforward. The first difference is, obviously, in the definition of the space V_h, which is now made of discontinuous functions. Let V_{DG} be such a space:

$$V_{\mathrm{DG}} := \{v \in L^2(\Omega) :\ v_{|K} \in V_k^K\ \forall K \in \mathcal{T}_h\}, \tag{5.1}$$

where the local spaces V_k^K are still defined as in (2.3). We recall the definition of jumps and averages for scalar and vector-valued functions ($v, \boldsymbol{\tau}$, respectively) on an edge e common to two elements $K_1,\ K_2$ (see [2]):

$$\begin{aligned}\{v\} &= \frac{v^1 + v^2}{2}, \quad [\![\, v \,]\!] = v^1 \mathbf{n^1} + v^2 \mathbf{n^2} \\ \{\boldsymbol{\tau}\} &= \frac{\boldsymbol{\tau}^1 + \boldsymbol{\tau}^2}{2}, \quad [\![\, \boldsymbol{\tau} \,]\!] = \boldsymbol{\tau}^1 \cdot \mathbf{n^1} + \boldsymbol{\tau}^2 \cdot \mathbf{n^2},\end{aligned}$$

where $\mathbf{n}^1,\ \mathbf{n}^2$ are the outward normal unit vectors to $K_1,\ K_2$. On a boundary edge we only need $[\![\, v \,]\!] = v\mathbf{n}$ and $\{\boldsymbol{\tau}\} = \boldsymbol{\tau}$.

We also define, for every $t \geq 0$, the space $H^t(\mathcal{T}_h) := \prod_K H^t(K)$ of piecewise regular functions. We recall from Remark 2.1 that

$$V_{\mathrm{DG}} \subset H^{s*}(\mathcal{T}_h) \tag{5.2}$$

for some $s* > 3/2$ depending on the value of ζ. Then for v, $w \in H^{s*}(\mathcal{T}_h)$ we set

$$(\nabla v, \nabla w)_h = \sum_K \int_K \nabla v \cdot \nabla w \, \mathrm{d}\mathbf{x}, \quad \langle \{\nabla v\}, [\![w]\!] \rangle = \sum_e \int_e \{\nabla v\} \cdot [\![w]\!] \, \mathrm{d}s$$
$$\langle [\![v]\!], [\![w]\!] \rangle = \sum_e \frac{1}{h_e} \int_e [\![v]\!] \cdot [\![w]\!] \, \mathrm{d}s, \quad \|[\![v]\!]\|^2_{0,\partial K} = \sum_{e \subset \partial K} \frac{1}{h_e} \int_e |[\![v]\!]|^2 \, \mathrm{d}s.$$

For $v \in H^2(\mathcal{T}_h)$ we define

$$\|v\|^2_{2,\mathrm{DG}} = \sum_{K \in \mathcal{T}_h} \left(\|\nabla v\|^2_{0,K} + h_K^2 |\nabla v|^2_{1,K} \right) + \langle [\![v]\!], [\![v]\!] \rangle. \tag{5.3}$$

We remark that, for functions v_h that are piecewise polynomials, by the usual inverse inequality we have

$$\|v_h\|^2_{2,\mathrm{DG}} \simeq \|v_h\|^2_{1,\mathrm{DG}} := \sum_{K \in \mathcal{T}_h} \|\nabla v_h\|^2_{0,K} + \langle [\![v_h]\!], [\![v_h]\!] \rangle. \tag{5.4}$$

We also set, for functions v, $w \in H^1(\mathcal{T}_h)$

$$\tilde{a}(v, w) := \sum_{K \in \mathcal{T}_h} a^K(v, w) = \sum_{K \in \mathcal{T}_h} \int_K \nabla v \cdot \nabla w \, \mathrm{d}\mathbf{x}.$$

We observe that the solution u of (2.1) verifies $[\![\nabla u]\!] = 0$ so that, integrating by parts on each element and recalling that $f = -\Delta u$, we have

$$\tilde{a}(u, v) - \langle \{\nabla u\}, [\![v]\!] \rangle = (f, v) \quad \forall v \in V_{\mathrm{DG}}. \tag{5.5}$$

On the other hand, the solution u of (2.1) obviously verifies $[\![u]\!] = 0$ as well, so that adding terms that are identically zero for $[\![u]\!] = 0$ we also have, for every $v \in V_{\mathrm{DG}}$ and for every real numbers δ and γ:

$$\tilde{a}(u, v) - \langle \{\nabla u\}, [\![v]\!] \rangle - \delta \langle \{\nabla v\}, [\![u]\!] \rangle + \gamma \langle [\![u]\!], [\![v]\!] \rangle = (f, v). \tag{5.6}$$

In what follows (as usual for DG methods) we will actually consider only the values $\delta = 1$, $\delta = -1$ and $\delta = 0$, while γ will be assumed to be positive and represents the usual penalty parameter.

6. Discontinuous VEM. All this will help us in constructing the discrete problem. To start with, for every $K \in \mathcal{T}_h$ we consider again the operator Π_k^K defined in (3.2), and assume that S^K is a bilinear form satisfying the stability property (3.4), that we recall here to avoid confusion with the new notation:

$$c_*(\nabla v, \nabla v)_{0,K} \le S^K(v,v) \le c^*(\nabla v, \nabla v)_{0,K} \quad \forall v \in V_k^K \quad \text{with } \Pi_k^K v = 0. \tag{6.1}$$

We set, for v, $w \in H^{s*}(\mathcal{T}_h)$,

$$\begin{aligned} \tilde{a}_h(v,w) &:= \sum_{K\in\mathcal{T}_h} \tilde{a}_h^K(v,w) \\ \tilde{a}_h^K(v,w) &= (\nabla\Pi_k^K v, \nabla\Pi_k^K w)_{0,K} + S^K(v-\Pi_k^K v, w-\Pi_k^K w). \end{aligned} \tag{6.2}$$

THEOREM 6.1. *The bilinear form* (6.2) *satisfies the consistency property* (2.6) *and the stability property* (2.7).

Proof. The proof is exactly the same of Theorem 3.1 and gives (in the new notation)

$$\tilde{a}_h^K(p,v) = a^K(p,v) \quad \forall p \in \mathbb{P}_k(K),\ \forall v \in (V_{\mathrm{DG}})_{|K}, \tag{6.3}$$

$$\alpha_* a^K(v,v) \le \tilde{a}_h^K(v,v) \le \alpha^* a^K(v,v)\ \forall v \in (V_{\mathrm{DG}})_{|K}. \tag{6.4}$$

□

Finally, for w, $v \in H^1(\mathcal{T}_h)$ we define the discrete bilinear form as

$$B_h(w,v) := \tilde{a}_h(w,v) - \langle\{\nabla\Pi_k w\}, [\![v]\!]\rangle - \delta\langle\{\nabla\Pi_k v\}, [\![w]\!]\rangle + \gamma\langle [\![w]\!], [\![v]\!]\rangle. \tag{6.5}$$

In (6.5) δ is, as already said, a parameter to include different DG-schemes. Precisely, for $\delta = 1$ we have the Virtual Element analogue of the *SIPG* (see [1, 10]), for $\delta = -1$ the analogue of the *NIPG* (see [8]), and for $\delta = 0$ the analogue of the *IIPG* [7, 9]. On the other hand, as we already said, γ is a stabilization parameter that will be assumed to be *big enough*, as usual for DG methods. We also point out that, with an abuse of notation, in (6.5) we denoted by Π_k the operator which, on each element K, coincides with Π_k^K.

THEOREM 6.2. *There exist positive constants M_s and C, independent of h, such that:*

$$B_h(v,v) \ge M_s \|v\|_{1,\mathrm{DG}}^2 \qquad v \in V_{\mathrm{DG}}, \tag{6.6}$$

$$\tilde{a}_h(v,w) + \langle [\![v]\!], [\![w]\!]\rangle \le C\|v\|_{1,\mathrm{DG}}\|w\|_{1,\mathrm{DG}} \quad v,w \in V_{\mathrm{DG}}, \tag{6.7}$$

$$\langle\{\nabla v\}, [\![w]\!]\rangle \le C\|v\|_{2,\mathrm{DG}}\|w\|_{1,\mathrm{DG}} \quad v \in H^2(\mathcal{T}_h), w \in H^1(\mathcal{T}_h). \tag{6.8}$$

Proof. Following the typical analysis of DG methods (see also [5] in this book) we recall that, using the trace inequality

$$\|v\|_{0,\partial K}^2 \le C\Big(\ell^{-1}\|v\|_{0,K}^2 + \ell|v|_{1,K}^2\Big),$$

(ℓ being a characteristic length of K, for instance its diameter) we immediately deduce (6.8). We also notice that, if v is a piecewise polynomial, then (6.8) becomes

$$\langle\{\nabla v\}, [\![w]\!]\rangle \leq C\|v\|_{1,\mathrm{DG}}\|w\|_{1,\mathrm{DG}}, \quad v \text{ p.w. polynomial}, \ w \in H^1(\mathcal{T}_h), \tag{6.9}$$

thanks to (5.4). Inequality (6.7) is an immediate consequence of (6.4). Finally, from (6.4), (6.9) and Cauchy–Scharwz inequality we deduce (6.6) for γ big enough. □

We are now ready to define the discrete problem as follows.

$$\begin{cases} \text{Find } u_h \in V_{\mathrm{DG}} \text{ such that} \\ B_h(u_h, v_h) = \langle f_h, v_h\rangle \ \forall v_h \in V_{\mathrm{DG}}. \end{cases} \tag{6.10}$$

7. Convergence of DG-VEM. We have the following convergence result.

THEOREM 7.1. *Under Assumptions* **H2**, *for γ big enough and for $\delta = 0, 1, -1$ the discrete problem (6.10) has a unique solution u_h. Moreover, for every approximation u_I of u in V_{DG} and for every approximation u_π of u that is piecewise in $\mathbb{P}_k$, we have*

$$\|u - u_h\|_{1,\mathrm{DG}} \leq C\Big(\|u - u_I\|_{1,\mathrm{DG}} + \|u - u_\pi\|_{2,\mathrm{DG}} + \|f - f_h\|_{V'_{1,\mathrm{DG}}}\Big), \tag{7.1}$$

where C is a constant independent of h.

Proof. Stability (6.6) implies that problem (6.10) has a unique solution $u_h \in V_{\mathrm{DG}}$, and

$$\|u_h\|_{1,\mathrm{DG}} \leq \frac{\|f\|_0}{M_s}.$$

In order to prove (7.1), set $\eta_h := u_h - u_I$. From (6.6), and then using (6.10) and (6.5), we have:

$$\begin{aligned} M_s\,\|\eta_h\|^2_{1,\mathrm{DG}} \leq & B_h(\eta_h, \eta_h) = B_h(u_h, \eta_h) - B_h(u_I, \eta_h) \\ = & \Big(\langle f_h, \eta_h\rangle - \tilde{a}_h(u_I, \eta_h) + \langle\{\nabla\Pi^k u_I\}, [\![\eta_h]\!]\rangle\Big) \\ & + \Big(\delta\langle\{\nabla\Pi^k\eta_h\}, [\![u_I]\!]\rangle - \gamma < [\![u_I]\!], [\![\eta_h]\!]\Big) =: I + II. \end{aligned} \tag{7.2}$$

Adding and subtracting u_π, and then using (6.3) we have:

$$\begin{aligned} I = & \langle f_h, \eta_h\rangle - \sum_K \Big(\tilde{a}_h^K(u_I - u_\pi, \eta_h) + \tilde{a}_h^K(u_\pi, \eta_h)\Big) \\ & + \langle\{\nabla\Pi_k^K(u_I - u_\pi)\}, [\![\eta_h]\!]\rangle + \langle\{\nabla\Pi_k^K u_\pi\}, [\![\eta_h]\!]\rangle \\ = & \langle f_h, \eta_h\rangle - \sum_K \Big(\tilde{a}_h^K(u_I - u_\pi, \eta_h) + a^K(u_\pi, \eta_h)\Big) \\ & + \langle\{\nabla\Pi_k^K(u_I - u_\pi)\}, [\![\eta_h]\!]\rangle + \langle\{\nabla\Pi_k^K u_\pi\}, [\![\eta_h]\!]\rangle. \end{aligned} \tag{7.3}$$

Then we add the term $\tilde{a}(u,\eta_h) - \langle\{\nabla u\}, [\![\,\eta_h\,]\!]\rangle - (f,\eta_h)$ that, thanks to (5.5), is equal to zero, and in the last term we remember that, thanks to (3.3), $\Pi_k^K u_\pi = u_\pi$. We obtain

$$\begin{aligned} I =& \langle f_h,\eta_h\rangle - \sum_K \Big(\tilde{a}_h^K(u_I - u_\pi,\eta_h) + a^K(u_\pi,\eta_h)\Big) \\ &+ \tilde{a}(u,\eta_h) - \langle\{\nabla u\}, [\![\,\eta_h\,]\!]\rangle - (f,\eta_h) \\ &+ \langle\{\nabla\Pi_k^K(u_I - u_\pi)\}, [\![\,\eta_h\,]\!]\rangle + \langle\{\nabla u_\pi\}[\![\,\eta_h\,]\!]\rangle, \end{aligned} \tag{7.4}$$

that rearranging terms we write as

$$\begin{aligned} I =& \langle f_h,\eta_h\rangle - (f,\eta_h) - \sum_K \Big(\tilde{a}_h^K(u_I - u_\pi,\eta_h) + a^K(u_\pi - u,\eta_h)\Big) \\ &+ \langle\{\nabla\Pi_k^K(u_I - u_\pi)\}, [\![\,\eta_h\,]\!]\rangle + \langle\{\nabla(u_\pi - u)\}, [\![\,\eta_h\,]\!]\rangle. \end{aligned} \tag{7.5}$$

Using (6.9) and (6.8) in (7.5) we have then

$$|I| \le C\left(\|f - f_h\|_{V'_{1,\mathrm{DG}}} + \|u_I - u_\pi\|_{1,\mathrm{DG}} + \|u_\pi - u\|_{2,\mathrm{DG}}\right)\|\eta_h\|_{1,\mathrm{DG}}. \tag{7.6}$$

On the other hand, recalling first that $[\![\,u\,]\!] = 0$, and then using (5.4) we have

$$\begin{aligned} |II| &= \left|\delta\langle\{\nabla\Pi^k\eta_h\}, [\![\,u_I - u\,]\!]\rangle - \gamma < [\![\,u_I - u\,]\!], [\![\,\eta_h\,]\!]\right| \\ &\le C\,\|u - u_I\|_{1,\mathrm{DG}}\|\eta_h\|_{1,\mathrm{DG}}. \end{aligned} \tag{7.7}$$

Using (7.6) and (7.7) in (7.2) we have then

$$\|\eta_h\|_{1,\mathrm{DG}} \le C\Big(\|f - f_h\|_{V'_{1,\mathrm{DG}}} + \|u - u_I\|_{1,\mathrm{DG}} + \|u_\pi - u\|_{2,\mathrm{DG}}\Big), \tag{7.8}$$

and estimate (7.1) follows by triangle inequality. □

According to the classical Scott–Dupont theory (see, e.g., [4]) we have the following result.

PROPOSITION 7.1. *Assume that Assumption* **H1** *is satisfied. Then there exists a constant C, depending only on k and ζ, such that for every* $w \in H^{k+1}(K)$ *there exist a* $w_\pi \in \mathbb{P}_k(K)$, *and a* $w_I \in V_k^K$ *such that*

$$\begin{aligned} |w - w_\pi|_{r,K} &\le C\,h_K^{k+1-r}\,|w|_{k+1,K} \quad 0 \le r \le k+1, \\ |w - w_I|_{r,K} &\le C\,h_K^{k+1-r}\,|w|_{k+1,K} \quad r = 0, 1. \end{aligned} \tag{7.9}$$

This, together with (4.2), inserted in (7.1) gives the optimal estimate

$$\|u - u_h\|_{1,\mathrm{DG}} \le C\,h^k|u|_{k+1,\Omega}.$$

REFERENCES

[1] D.N. Arnold, *An interior penalty finite element method with discontinuous elements,* SIAM J. Numer. Anal. **19** (1982), pp. 742–760.

[2] D.N. Arnold, F. Brezzi, B. Cockburn, and L.D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems,* SIAM J. Numer. Anal. **39** (2001/02), pp. 1749–1779.

[3] L. Beirão da Veiga, F. Brezzi, A. Cangiani, G. Manzini, L.D Marini, and A. Russo, *Basic Principles of Virtual Element Methods,* Math. Models Methods Appl. Sci. **1** (2013), pp. 199–214.

[4] S. C. Brenner and R. L. Scott, *The mathematical theory of finite element methods,* Texts in Applied Mathematics, 15. Springer-Verlag, New York, 2008.

[5] F. Brezzi and L.D. Marini, *A quick tutorial on DG methods,* (this book).

[6] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems,* North-Holland, 1978.

[7] C. Dawson, S. Sun, and M.F. Wheeler, *Compatible algorithms for coupled flow and transport*, Comput. Methods Appl. Mech. Engrg. **193** (2004), pp. 2565–2580.

[8] B. Rivière, M.F. Wheeler, and V. Girault, *A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems,*SIAM J. Numer. Anal. **39** (2001), pp. 902–931.

[9] S. Sun and M.F. Wheeler, *Symmetric and nonsymmetric discontinuous Galerkin methods for reactive transport in porous media,* SIAM J. Numer. Anal. **43** (2005), pp. 195–219.

[10] M.F. Wheeler, *An elliptic collocation-finite element method with interior penalties,* SIAM J. Numer. Anal. **15** (1978), pp. 152–161.

A DG APPROACH TO HIGHER ORDER ALE FORMULATIONS IN TIME

ANDREA BONITO*, IRENE KYZA†, AND RICARDO H. NOCHETTO‡

Abstract. We review recent results (Bonito et al., SIAM J. Numer. Anal., to appear; Bonito et al., Numer. Math., to appear; Bonito et al., in preparation) on time-discrete discontinuous Galerkin (dG) methods for advection-diffusion model problems defined on deformable domains and written on the arbitrary Lagrangian Eulerian (ALE) framework. ALE formulations deal with PDEs on deformable domains upon extending the domain velocity from the boundary into the bulk with the purpose of keeping mesh regularity. We describe the construction of higher order in time numerical schemes enjoying stability properties independent of the arbitrary extension chosen. Our approach is based on the validity of Reynolds' identity for dG methods which generalize to higher order schemes the geometric conservation law (GCL) condition. Stability, a priori and a posteriori error analyses are briefly discussed and illustrated by insightful numerical experiments.

Key words. ALE formulations, Moving domains, Domain velocity, Material derivative, Discrete Reynolds' identities, dG-methods in time, Stability, Geometric conservation law

AMS(MOS) subject classifications. 65M12, 65M15, 65M50, 65M60.

1. Introduction.

Problems governed by partial differential equations (PDEs) on deformable domains $\Omega_t \subset \mathbb{R}^d$, which change in time $0 \le t \le T < \infty$, are of fundamental importance in science and engineering, especially for space dimensions $d \ge 2$. A typical example is fluid structure interaction problems. They are of particular relevance in the design of many engineering systems, (e.g., aircrafts and bridges) as well as to the analysis of several biological phenomena (e.g., blood flow in arteries).

However, the mathematical understanding of such methods is still precarious even when the deformation of the boundary $\partial\Omega_t$ of Ω_t is prescribed a priori and thus known, instead of more realistic free boundary problems. One obstacle encountered when dealing numerically with such problems is the possibility of excessive mesh distortion. The *arbitrary Lagrangian Eulerian (ALE)* approach, introduced in [16, 26, 27], is a way to overcome

*Department of Mathematics, Texas A&M University, College Station, TX 77843-3368, USA, bonito@math.tamu.edu

†Division of Mathematics, University of Dundee, Dundee, DD1 4HN, Scotland, UK, ikyza@maths.dundee.ac.uk
and
Institute of Applied and Computational Mathematics–FORTH, Nikolaou Plastira 100, Vassilika Vouton, Heraklion-Crete, Greece

‡Department of Mathematics and Institute of Physical Science and Technology, University of Maryland, College Park, MD 20742-4015, USA, rhn@math.umd.edu

X. Feng et al. (eds.), *Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations*, The IMA Volumes in Mathematics and its Applications 157, DOI 10.1007/978-3-319-01818-8_10,

this difficulty. Its main idea is that the mesh boundary is deformed according to the prescribed boundary velocity $\mathbf{w}$, but an arbitrary, yet adequate extension is used to perform the bulk deformation. The extension of $\mathbf{w}$ from $\partial\Omega_t$ to Ω_t can be performed using various techniques such as solving for a suitable boundary value problem with Dirichlet boundary condition $\mathbf{w}$; see [19, 22, 32, 35] and the references therein. This extension induces a map $\mathcal{A}_t : \Omega_0 \to \Omega_t$, the so-called *ALE map*, with the key property that

$$\mathbf{w}(\mathbf{x}, t) = \frac{d}{dt}\mathcal{A}_t(\mathbf{y}), \qquad \mathbf{x} = \mathcal{A}_t(\mathbf{y}).$$

The ALE velocity $\mathbf{w}$ is unrelated to the advection coefficient $\mathbf{b}$ inherent to the underlying system and is mostly dictated by the geometric principle of preserving mesh regularity. In contrast, the pure Lagrangian approach consists of mesh deformation velocities given by $\mathbf{w} = \mathbf{b}$, whereas $\mathbf{w} = \mathbf{0}$ corresponds to the pure Eulerian approach. In the latter case, $\Omega_t = \Omega_0$ for all $t \in [0, T]$, and thus the domain does not change in time. Hence, the ALE is a generalization of both the Lagrangian and the Eulerian approaches.

This paper is a review of our recent results [8–10] on the design, stability, and error control of higher order in time ALE formulations for a linear advection-diffusion model problem defined on time-dependent domains based on the discontinuous Galerkin (dG) approach. In particular, we discuss higher order in time, unconditionally stable numerical methods within the ALE framework, which seems to be lacking in the current literature. In the current paper, we intend to:

- Introduce the major difficulties caused by ALE formulations on deformable domains.
- Emphasize the key ideas leading to unconditionally stable higher order in time ALE formulations and point out the importance of such schemes for efficient error control.
- Compare our results with the existing literature and highlight the novel aspects of our analysis.

As already pointed out by other authors (see, e.g., [5, 20–22, 31]), time-discretization is the main obstruction for the design of unconditionally stable higher order ALE-schemes. This is why we examine in [8–10], and review here, the critical issue of time-discrete schemes without space discretization. This prevents additional technicalities in dealing with new tools developed to handle the domain motion. In addition, it guarantees that no unnecessary CFL type restrictions are required by our techniques.

However, we note that understanding the effect of the finite element discretization in space is an important problem. Extending our analysis to fully discrete schemes is not straightforward, yet plausible as the required regularity on the ALE map in space in our time-discrete approach is compatible with a C^0 finite element framework.

1.1. The ALE Formulation. As in [5, 7, 20–22, 31], we consider the following time-dependent diffusion-advection model problem defined on moving domains:

$$\begin{cases} \partial_t u + \nabla_{\mathbf{x}} \cdot (\mathbf{b}u) - \mu \Delta_{\mathbf{x}} u = f & \mathbf{x} \in \Omega_t,\ t \in [0,T] \\ u(\mathbf{x},t) = 0 & \mathbf{x} \in \partial\Omega_t,\ t \in [0,T] \\ u(\mathbf{x},0) = u_0(\mathbf{x}) & \mathbf{x} \in \Omega_0, \end{cases} \tag{1.1}$$

where $\mu > 0$ is a constant diffusion parameter, $\mathbf{b}$ is a (divergence-free) convective velocity, f is a forcing term, and u_0 is the initial condition.

In order to rewrite (1.1) in the ALE framework, we consider for $t = 0$, Ω_0 as the reference domain assumed to have Lipschitz boundary $\partial\Omega_0$, and let $\Omega_t \subset \mathbb{R}^d$ be the corresponding moving domain at time $t \in (0,T]$. Let $\{\mathcal{A}_t\}_{t\in[0,T]}$ be a family of maps with $\mathcal{A}_0 = \mathrm{I}_d$ the identity map, such that $\Omega_t = \mathcal{A}_t(\Omega_0)$, $t \in [0,T]$. In other words, for $t \in [0,T]$, each $\mathbf{y}$ from the reference domain Ω_0 is mapped through $\mathcal{A}_t$ to the corresponding $\mathbf{x} \in \Omega_t$, i.e., the map $\mathcal{A}_t$ is given by:

$$\mathcal{A}_t : \Omega_0 \subseteq \mathbb{R}^d \to \Omega_t \subseteq \mathbb{R}^d, \quad \mathbf{x}(\mathbf{y},t) = \mathcal{A}_t(\mathbf{y}).$$

We frequently regard $\mathcal{A}_t$ as a space-time function $\mathcal{A}(\mathbf{y},t) := \mathcal{A}_t(\mathbf{y})$, and we refer to $\mathbf{y} \in \Omega_0$ as the ALE coordinate and $\mathbf{x} = \mathbf{x}(\mathbf{y},t)$ as the spatial or Eulerian coordinate. Hereafter, we say that $\{\mathcal{A}_t\}_{t\in[0,T]}$ is a *family of ALE maps* if the following two conditions are satisfied [10]:

- Regularity: $\mathcal{A}(\cdot,\cdot) \in \mathbf{W}^1_\infty\big((0,T);\mathbf{W}^1_\infty(\Omega_0)\big)$;
- Injectivity: there exists a constant $\lambda > 0$ such that for all $t \in [0,T]$,

$$\|\mathcal{A}_t(\mathbf{y}_1) - \mathcal{A}_t(\mathbf{y}_2)\| \geq \lambda \|\mathbf{y}_1 - \mathbf{y}_2\|, \qquad \forall\, \mathbf{y}_1,\ \mathbf{y}_2 \in \Omega_0, \tag{1.2}$$

for some norm $\|\cdot\|$ in $\mathbb{R}^d$. The regularity assumption implies that $\mathcal{A}_t$ is Lipschitz continuous, whereas the combination with injectivity assumption gives that $\mathcal{A}_t : \Omega_0 \to \Omega_t$ is invertible with Lipschitz inverse, i.e., $\mathcal{A}_t$ is bi-Lipschitz and thus a homeomorphism. This implies that $v := \hat{v} \circ \mathcal{A}_t^{-1} \in H^1_0(\Omega_t)$ if and only if $\hat{v} \in H^1_0(\Omega_0)$, [21, Proposition 1].

Using these notations, problem (1.1) is defined in the space-time domain:

$$\mathcal{Q}_T := \big\{(\mathbf{x},t) \in \mathbb{R}^d \times \mathbb{R} : t \in [0,T],\ \mathbf{x} = \mathcal{A}_t(\mathbf{y}),\ \mathbf{y} \in \Omega_0\big\}.$$

The *ALE velocity* $\hat{\mathbf{w}} : \Omega_0 \times [0,T] \to \mathbb{R}^d$ in the ALE frame is given by

$$\hat{\mathbf{w}}(\mathbf{y},t) := \partial_t \mathbf{x}(\mathbf{y},t),$$

and we indicate by $\mathbf{w} : \mathcal{Q}_T \to \mathbb{R}^d$ the corresponding function on the Eulerian frame. We use ∂_t to denote the usual weak partial derivative in time holding the space variable $\mathbf{x}$ constant. Given a function $g : \mathcal{Q}_T \to \mathbb{R}$, we denote by $D_t g$ the *ALE time-derivative*, namely the time-derivative keeping the ALE coordinate $\mathbf{y}$ fixed:

$$(D_t g)(\mathbf{x},t) := (\partial_t g)(\mathcal{A}_t(\mathbf{y}),t).$$

The derivation of the ALE formulation of (1.1) is based on the next lemma, proved in [10].

LEMMA 1.1 (Leibnitz formula in $W_1^1(\mathcal{Q}_T)$). *Let $g \in W_1^1(\mathcal{Q}_T)$ and $\{\mathcal{A}_t\}_{t\in[0,T]}$ be a family of ALE maps. Then, $D_t g \in L^1(\mathcal{Q}_T)$ and*

$$D_t g = \partial_t g + \mathbf{w} \cdot \nabla_{\mathbf{x}} g. \tag{1.3}$$

Leibnitz formula (1.3) relates the usual time-derivative to the corresponding ALE time-derivative through the ALE velocity $\mathbf{w}$ and it is a justification of the chain rule for weak ALE time-derivatives. Using (1.3), (1.1) is equivalently written in the ALE framework as follows:

$$\begin{cases} D_t u + (\mathbf{b} - \mathbf{w}) \cdot \nabla_{\mathbf{x}} u - \mu \Delta_{\mathbf{x}} u = f & \text{in } \mathcal{Q}_T, \\ u = 0 & \text{on } \partial\mathcal{Q}_T, \\ u(\cdot, 0) = u_0 & \text{in } \Omega_0. \end{cases} \tag{1.4}$$

Before setting problem (1.4) in its variational form, we introduce some further notation. For any domain D of $\mathbb{R}^m$, $m = d$ or $d+1$, we denote by $W_r^\ell(D)$ the standard Sobolev spaces with integrability $1 \le r \le \infty$ and differentiability $0 \le \ell < \infty$. We use the notation $L^r(D)$ when $\ell = 0$ and $H^\ell(D)$ when $r = 2$ and $\ell \ge 1$. With $H_0^1(D)$ we denote the subspace of $H^1(D)$ consisting of functions with vanishing trace and equipped with the norm $\|\nabla_{\mathbf{x}} v\|_{L^2(D)}$; we denote its dual by $H^{-1}(D)$. We indicate with $\langle \cdot, \cdot \rangle_D$ both the $H_0^1 - H^{-1}$ duality pairing and the L^2-inner product in D, depending on the context. Spaces of vector-valued functions are written in boldface. For $Y = W_r^\ell$, $\ell \ge 0, 1 \le r \le \infty$, H_0^1, or H^{-1}, we define the spaces

$$L^2(Y; \mathcal{Q}_T) := \Big\{ v : \mathcal{Q}_T \to \mathbb{R} : \int_0^T \|v(t)\|_{Y(\Omega_t)}^2 \, dt < \infty \Big\}.$$

We define accordingly the spaces $C(Y; \mathcal{Q}_T)$ of continuous functions with values in Y, and set

$$L^\infty(\mathrm{div}; \mathcal{Q}_T) := \{ \mathbf{c} : \mathcal{Q}_T \to \mathbb{R}^d : \\ \operatorname{ess\,sup}_{t\in(0,T)} \big(\|\mathbf{c}(t)\|_{\mathbf{L}^\infty(\Omega_t)} + \|\nabla_{\mathbf{x}} \cdot \mathbf{c}(t)\|_{L^\infty(\Omega_t)} \big) < \infty \}.$$

To simplify the notation we omit writing the dependency in $\mathcal{Q}_T$ when there is no confusion.

A *non-conservative* weak ALE formulation for problem (1.4) reads as follows: seek $u \in L^2(H_0^1;\ \mathcal{Q}_T) \cap H^1(L^2; \mathcal{Q}_T)$ satisfying $u(\cdot, 0) = u_0$ and such that for all $v \in L^2(H_0^1)$ and $\tau, t \in [0, T]$ with $\tau < t$,

$$\int_\tau^t \langle D_t u, v \rangle_{\Omega_s} \, ds + \int_\tau^t \langle (\mathbf{b} - \mathbf{w}) \cdot \nabla_{\mathbf{x}} u, v \rangle_{\Omega_s} \, ds \\ + \int_\tau^t \langle (\nabla_{\mathbf{x}} \cdot \mathbf{b}) u, v \rangle_{\Omega_s} \, ds + \mu \int_\tau^t \langle \nabla_{\mathbf{x}} u, \nabla_{\mathbf{x}} v \rangle_{\Omega_s} \, ds = \int_\tau^t \langle f, v \rangle_{\Omega_s} \, ds. \tag{1.5}$$

It is known that (1.5) admits a unique solution, provided that $u_0 \in H_0^1(\Omega_0)$, $f \in L^2(\mathcal{Q}_T)$ and $\mathbf{b} \in L^\infty(\mathrm{div}; \mathcal{Q}_T)$, [10]. This regularity on the data of the problem guarantees that $u \in H^1(\mathcal{Q}_T) \subset C(L^2; \mathcal{Q}_T)$, which implies the further regularity $\Delta_{\mathbf{x}} u$, $D_t u \in L^2(\mathcal{Q}_T)$; cf. [10].

A conservative weak formulation for (1.4) can be obtained using Reynolds' identities, which can be regarded as weak versions of Reynolds' Transport Theorem [10].

LEMMA 1.2 (Reynolds' identities). *Let $\{\mathcal{A}_t\}_{t\in[0,T]}$ be a family of ALE maps. For any $v \in W_1^1(\mathcal{Q}_T)$ there holds*

$$\frac{d}{dt}\int_{\Omega_t} v\, d\mathbf{x} = \int_{\Omega_t} (D_t v + v\nabla_{\mathbf{x}} \cdot \mathbf{w})\, d\mathbf{x}. \tag{1.6}$$

In particular, for w, $v \in H^1(\mathcal{Q}_T)$ we have

$$\frac{d}{dt}\int_{\Omega_t} vw\, d\mathbf{x} = \int_{\Omega_t} w(D_t v + v\nabla_{\mathbf{x}} \cdot \mathbf{w})\, d\mathbf{x} + \int_{\Omega_t} v D_t w\, d\mathbf{x}. \tag{1.7}$$

As we shall see in the sequel, and is thoroughly discussed in [9, 10], Reynolds' identities (1.6), (1.7) play a significant role in the stability and error analysis of dG schemes for (1.4).

Inserting (1.7) into (1.5) gives the *conservative* weak formulation for problem (1.1):

$$\begin{aligned}&\langle u(t), v(t)\rangle_{\Omega_t} + \int_\tau^t \langle \nabla_{\mathbf{x}} \cdot \big[(\mathbf{b}-\mathbf{w})u\big], v\rangle_{\Omega_s}\, ds + \mu \int_\tau^t \langle \nabla_{\mathbf{x}} u, \nabla_{\mathbf{x}} v\rangle_{\Omega_s}\, ds \\ &\quad - \int_\tau^t \langle u, D_t v\rangle_{\Omega_s}\, ds = \langle u(\tau), v(\tau)\rangle_{\Omega_\tau} + \int_\tau^t \langle f, v\rangle_{\Omega_s}\, ds, \quad \forall v \in H_0^1(\mathcal{Q}_T).\end{aligned} \tag{1.8}$$

It is clear that non-conservative and conservative weak formulations (1.5) and (1.8) are equivalent at the continuous level. Moreover, setting $v = u$ in either (1.5) or (1.8), it is possible to prove the following stability result for the PDE (1.1) [10, 21]:

$$\begin{aligned}&\|u(t)\|^2_{L^2(\Omega_t)} + \mu \int_\tau^t \|\nabla_{\mathbf{x}} u(s)\|^2_{L^2(\Omega_s)}\, ds \\ &\qquad \le \|u(\tau)\|^2_{L^2(\Omega_\tau)} + \frac{1}{\mu}\int_\tau^t \|f(s)\|^2_{H^{-1}(\Omega_s)}\, ds,\end{aligned} \tag{1.9}$$

for $0 \le \tau \le t \le T$. Note that in particular, estimate (1.9) does not involve any constants depending on the particular choice of the ALE map $\mathcal{A}_t$ and exhibits monotone behavior of the norm $\|u(t)\|_{L^2(\Omega_t)}$ provided $f \equiv 0$. This is expected, as the original problem (1.1) is independent of the family of the ALE maps. However, this property is not guaranteed anymore after time discretization. In fact, at the discrete level, the arbitrary extension of the ALE map, not only may influence and pollute the stability of the

corresponding discrete scheme, but it may also lead to schemes where conservative and non-conservative formulations are no longer equivalent. We discuss this further in the sequel.

We say that a numerical method for problem (1.4) is *ALE-free stable* with respect to the energy norm if it reproduces (1.9); otherwise, if (1.9) is valid with a constant multiplying the right-hand side that depends on $\mathcal{A}_t$, we say that the method is *ALE stable*. In both cases, we say that the method is *stable*. ALE-free stable schemes are most desirable for problem (1.4) because they enjoy the same stability properties as the continuous problem.

1.2. Existing Literature. Second order ALE methods for advection-dominated diffusion problems on moving domains are discussed in the literature for finite difference and finite volume schemes [14, 17–19]. A numerical scheme is said to satisfy the GCL if it is able to reproduce exactly a constant solution. GCL was introduced in [17, 24, 36] for finite volume schemes as a minimum criterion for unconditional stability. However, the *GCL is not a necessary condition* for numerical schemes to be ALE-free stable. For example, Geuzaine et al. [23] propose second order finite volume ALE schemes which enjoy the same stability properties as the continuous problem without satisfying the GCL.

On the other hand, the only provable ALE-free stable scheme for (1.4) based on finite element (FE) discretization in space, and without time-step constraints (unconditional stability), is the backward Euler method [7, 20–22, 31]. In particular, Formaggia and Nobile [21] consider a conservative backward Euler FE scheme for (1.4) which satisfies the GCL and prove that this scheme is ALE-free stable. Moreover, they study a non-conservative backward Euler FE scheme for (1.4) that fails to satisfy the GCL [21]. This method is not equivalent to the corresponding conservative scheme and it turns out that its stability estimate requires a time-step restriction depending on the ALE map. In addition, the derivation of the stability estimate relies on a Gronwall-type argument which entails constants depending exponentially on the L^∞-norm of the domain velocity $\mathbf{w}$. This does not reflect the stability properties of the continuous problem (1.4). A priori error analysis for the backward Euler FE methods is provided by Gastaldi [22], Boffi and Gastaldi [7], Nobile [31], and Badia and Codina [5].

Besides first order schemes, there are also second order FE schemes based on the Crank–Nicolson method and the backward differentiation formula (BDF) method of second order; see Formaggia and Nobile [20], Boffi and Gastaldi [7], and Badia and Codina [5]. One consequence of these studies is that the *GCL condition is not sufficient* to guarantee an ALE-free stable scheme, and similar stability issues as for the non-conservative backward Euler FE scheme are observed. In fact, numerical simulations show that the monotonicity of $\|u(t)\|_{L^2(\Omega_t)}$ does not hold at the discrete level [20]. At a theoretical level, this effect appears when the ALE velocity $\mathbf{w}$ is

treated as an extra advection for the method. Hence, the stability bound requires a Gronwall-type argument which yields time-step constraints and stability constants depending on the ALE map. This is despite the fact that (1.9) is insensitive to $\mathbf{w}$.

The generalization of the GCL condition proposed in our recent work [9, 10] is a *sufficient* condition to guarantee ALE-free stable schemes (independent of the accuracy of the method). In fact, we propose a class of ALE-free schemes of any desired accuracy based on dG methods in time. These seem to be the first schemes having this property.

At this point, we also mention the work by Mackenzie and Mekwi [29], who propose an adaptive θ-method time integrator and prove it is ALE-free stable. However, the statement about asymptotic second order accuracy hinges on heuristics. This method chooses the parameter θ in each time-step, depending on the domain velocity $\mathbf{w}$, so as to satisfy the GCL.

Moreover, it is worth mentioning that this discussion is about a priori error analysis. We are unaware of any a posteriori error estimates for ALE methods.

1.3. A dG Approach. As already alluded to in the previous subsection, the existing FE methods and their analysis share the following common features (1.4):

- All weak formulations have test functions with vanishing material derivative, which seems not adequate when the space and time are tangled together.
- All time discretizations are defined pointwise, which makes it difficult, if not impossible, to reproduce at the discrete level the subtle cancellations leading to the stability estimate (1.9).
- For second order schemes, the discrete stability is obtained via a discrete Gronwall-type argument, which does not reflect the stability properties of the continuous problem.
- Despite the fact that the role of the GCL for stability and accuracy is not clear, for FE schemes the GCL is closely related to quadrature in time; in fact, for test functions with vanishing material derivative, the GCL is equivalent to a discrete version of Reynolds' identity (1.7).

In view of these observations, we realize that the main obstruction in the design of ALE-free stable schemes is the discretization in time. Hence, we consider in [8–10] time-discrete schemes based on dG methods of any order. The reasons for this specific choice are:

- The backward Euler method, the only known ALE-free stable method, is a dG method.
- dG methods couple time and space in a natural variational way and allow for test functions with nonvanishing material derivative. As we shall see in the sequel, the variational structure of the methods suggests a class of ALE-free stable schemes of any order in time [10].

- dG methods are suitable for time and space adaptivity. In this context, adaptivity is an essential tool for capturing disparate space and time scales efficiently and to cope with the nonlinear interaction between the approximation of the moving domain and the approximation of the solution to (1.1) (defined on the approximate domain).

In [10], we propose dG methods of any order for problems defined on time-dependent domains and the ALE framework, and we study their stability properties. We also derive practical algorithms by enforcing appropriate quadrature in time. The first family of these algorithms are the so-called Reynolds' methods and correspond to quadratures that keep valid the Reynolds' identity at the discrete level. It can be shown that, for piecewise polynomial in time ALE maps, Reynolds' methods are ALE-free stable and lead to optimal order error bounds (both a priori and a posteriori), without any time-step restrictions. The second family of methods is the well-known Runge–Kutta–Radau (RKR) methods that result from dG methods of order $q+1$ by approximating integrals in time by the Radau quadrature with $q+1$ nodes. These methods are also ALE-free stable, but subject to an ALE-time-step constraint (conditional stability). We perform an a priori error analysis for these methods in [9] and an a posteriori error analysis in [8]. Our work extends the analysis of dG methods of any order for non-moving domains [37, Chap. 12] to time-dependent domains within the ALE framework. We also refer to [25] for the implementation of first-order dG methods in the context of fluid–structure interactions. dG methods have been considered earlier by Chrysafinos and Walkington in [15] within a pure Lagrangian approach for advection-dominated diffusion problems but with a purpose distinct from ours. More precisely, in our approach, the ALE velocity $\mathbf{w}$ does not play the role of an advective velocity, while in the approach of Chrysafinos and Walkington, the ALE velocity $\mathbf{w}$ is designed to compensate for large advections $\mathbf{b}$, and thus is chosen to satisfy $\mathbf{w} \approx \mathbf{b}$. For more details on a comparison between [15] and our work, we refer to [10].

In this paper we review the results of [8–10] about stability and error analysis of dG methods for ALE formulations. The main contributions of [10] regarding stability are as follows:

- Propositions of dG methods of any order in time with exact integration for problems defined on time-dependent domains and the ALE framework. These methods lead to ALE-free stability at the nodes $t = t_n$ (nodal stability) and ALE stability for all $t \in [0, T]$ (global stability), both without time-step constraints (unconditional stability).
- Generalization of GCL to higher order in time ALE schemes. The variational structure of the dG methods allow us to impose appropriate quadrature in time that lead to a discrete Reynolds' identity for piecewise polynomial ALE maps in time. The chosen quadrature leads to the practical Reynolds' methods for which we show ALE-free nodal stability and ALE global stability, both without time-step constraints.

- Stability study of dG methods with Radau quadrature, which is the natural quadrature for problems defined on time-independent domains. For these methods, we prove ALE-free nodal stability and ALE global stability, both with an ALE time-constraint, but for any ALE map with piecewise W^2_∞ regularity in time and global W^1_∞ in space ALE maps.

The main contributions of [8, 9] related to error control are the following:

- Introduction of a novel ALE projection and study of its properties. The ALE projection extends the usual dG projection to time-dependent domains and the ALE framework. It is one of the main ingredients in the a priori error analysis.
- Introduction of a novel dG reconstruction in the ALE framework and study of its properties. This reconstruction is a generalization of the dG reconstruction in [30], proposed by Makridakis and Nochetto for time-independent domains. It is one of the main ingredients leading to optimal order a posteriori error bounds.
- Proof of optimal order a priori error estimates for dG of any degree with exact integration and Reynolds' quadrature, the latter provided that the ALE map is a continuous piecewise polynomial in time. These estimates are valid without time-step restrictions. In addition, we prove optimal order a priori error estimates for dG with Radau quadrature, but under a mild time-step restriction depending on μ and $\mathcal{A}_t$. The first a priori error bounds for dG-type methods applied to (1.1) were derived by Jamet [28], but without imposing any quadrature in the involved integrals, and thus leading to non-practical algorithms.
- Proof of optimal order a posteriori error estimates for dG of any order with exact integration and for Reynolds' methods, the latter provided that $\mathcal{A}_t$ is a continuous piecewise polynomial in time. These estimates are valid without any time-step restrictions and are obtained using the same stability PDE techniques as for the continuous problem through the dG reconstruction on the ALE framework. As already mentioned, there seems to be no a-posteriori error estimates available in the current literature for problems defined on moving domains.
- Piecewise polynomial approximation of the ALE map. Given the domain velocity $\mathbf{w}$ at the boundary $\partial\Omega_t$, we approximate $\mathbf{w}$ on the boundary by a piecewise polynomial of degree q (to match with the dG accuracy) in the case of Reynolds' methods. We reconstruct a perturbed ALE map $\tilde{\mathcal{A}}_t$ by time integration and suitable extension inside the domain. Then, we obtain optimal order error bounds by invoking a PDE perturbation argument in which we evaluate the additional geometrical error committed by the approximation of the ALE map.

As discussed in [10, Sect. 6], our results can be extended to problems (1.1) with advections with nonvanishing divergence. This is a prototype problem for the practically more interesting and theoretically more challenging Navier–Stokes equations on time-dependent domains. Some

preliminary estimates for these equations and the Crank–Nicolson FE method that satisfies the GCL have been derived by Nobile in [31]. In [34], Quarteroni and Formaggia use the same method in simulations of their model of the cardiovascular system, based on Navier–Stokes equations on moving domains.

1.4. Organization of the Paper. The paper is organized as follows. In Sect. 2 we review results from [8–10] regarding the stability and the error analysis for dG methods in time of any order $q \geq 0$ for problem (1.4) and assuming exact integration in time. In Sect. 3 we consider practical algorithms that are obtained from the dG methods by applying Reynolds' quadrature, and we discuss the relation between Reynolds' quadrature and the GCL. Finally, Sect. 4 is devoted to RKR methods. We present, without proofs, the main results of [9, 10] on stability and a priori error analysis, and we explain the importance of RKR methods.

2. Discontinuous Galerkin Method in Time: Exact Integration. In this section, we review some recent results from [9, 10] related to the stability and a priori error analysis for dG methods for problem (1.5) [or (1.8)] defined on time-dependent domains. We also report without proofs the main results of the a posteriori error analysis developed in [8]. We discuss exact integration in time and we emphasize the main ideas and the key ingredients leading to efficient (in terms of stability) practical algorithms.

2.1. The dG Method and Stability. The time discretization of (1.5) or (1.8) starts with a partition $0 =: t_0 < t_1 < \cdots < t_N =: T$ of $[0,T]$. For $0 \leq n \leq N-1$, we let $I_n := (t_n, t_{n+1}]$ and $k_n := t_{n+1} - t_n$ be the subintervals and variable time-steps, respectively. We also let $k := \max_{0\leq n\leq N-1} k_n$ and

$$\mathcal{Q}_n := \{(\mathbf{x},t) \in \mathcal{Q}_T : t \in I_n\}.$$

For $q \geq 0$, the discrete space $\mathcal{V}_q$ associated with the dG method in time of order $q+1$ is [10]

$$\mathcal{V}_q := \{V : \mathcal{Q}_T \to \mathbb{R} : V|_{I_n} = \sum_{j=0}^{q} \varphi_j t^j \text{ where } \varphi_j \in L^2(H_0^1) \text{ with } D_t\varphi_j = 0,\ 0 \leq j \leq q\}. \tag{2.1}$$

Note that the definition of the discrete space (2.1) is a natural generalization of the corresponding dG space for problems defined on moving domains and written on the ALE framework. Indeed, when the domain does not undergo deformations, i.e., $\mathcal{A}_t = \mathrm{I}_d$ is the identity map for all $t \in [0,T]$, $\mathcal{V}_q$ is reduced to the standard dG space [1, 37]. Moreover, the definition of $\mathcal{V}_q$ ensures that whenever $V \in \mathcal{V}_q$ is considered on the reference domain,

$\hat{V}(\mathbf{y},t) := V\big(\mathcal{A}_t(\mathbf{y}),t\big)$, $\hat{V}$ is piecewise polynomial of degree at most q with coefficients in H_0^1. In the dG analysis it is customary to consider the space

$$\mathcal{V}_q(I_n) := \{V : \mathcal{Q}_n \to \mathbb{R} : V = W|_{\mathcal{Q}_n},\, W \in \mathcal{V}_q\}, \quad 0 \le n \le N-1,$$

consisting of restrictions to $\mathcal{Q}_n$ of functions in $\mathcal{V}_q$.

The dG approximation U to u via the non-conservative ALE formulation is defined as follows [10]: seek $U \in \mathcal{V}_q$ such that

$$U(\cdot,0) = u_0 \ \text{ in } \ \Omega_0, \tag{2.2}$$

and for $0 \le n \le N-1$,

$$\begin{aligned}
&\int_{I_n} \langle D_t U, V\rangle_{\Omega_t}\, dt + \langle U(t_n^+) - U(t_n), V(t_n^+)\rangle_{\Omega_{t_n}} \\
&\quad + \int_{I_n} \langle (\mathbf{b}-\mathbf{w})\cdot\nabla_{\mathbf{x}} U, V\rangle_{\Omega_t}\, dt + \mu \int_{I_n} \langle \nabla_{\mathbf{x}} U, \nabla_{\mathbf{x}} V\rangle_{\Omega_t}\, dt \\
&= \int_{I_n} \langle f, V\rangle_{\Omega_t}\, dt, \quad \forall V \in \mathcal{V}_q(I_n).
\end{aligned} \tag{2.3}$$

Similarly, the conservative dG formulation based on (1.8) reads as [10]

$$\begin{aligned}
&\langle U(t_{n+1}), V(t_{n+1})\rangle_{\Omega_{t_{n+1}}} - \langle U(t_n), V(t_n^+)\rangle_{\Omega_{t_n}} \\
&\quad + \int_{I_n} \langle \nabla_{\mathbf{x}}\cdot\big((\mathbf{b}-\mathbf{w})U\big), V\rangle_{\Omega_t}\, dt + \mu \int_{I_n} \langle \nabla_{\mathbf{x}} U, \nabla_{\mathbf{x}} V\rangle_{\Omega_t}\, dt \\
&\quad - \int_{I_n} \langle U, D_t V\rangle_{\Omega_t}\, dt = \int_{I_n} \langle f, V\rangle_{\Omega_t}\, dt, \quad \forall V \in \mathcal{V}_q(I_n).
\end{aligned} \tag{2.4}$$

We emphasize that the above choice of discrete space $\mathcal{V}_q$, containing (time) discrete functions with nonvanishing material derivative, implies that the non-conservative formulation (2.3) and the conservative formulation (2.4) remain equivalent at the discrete level as well, and any $q \ge 0$ [10]. In contrast to most pointwise methods, we also note that the dG method produces approximations defined in the deformable domain Ω_t for all times $t \in [0,T]$. The latter consistency on the domain in which the dG approximation is defined is critical for the stability analysis, as first observed by Pironneau et al. [33]. In particular, defining the dG method as in (2.3) or (2.4) provides the same coupling between time and space present at the continuous level, thereby giving rise to desirable stability properties for the time-discrete methods, [10]. Finally, as in the definition of the discrete space, we have that both (2.3) and (2.4) reduce to the standard dG method in time when the domain does not change.

The well posedness of the approximation $U \in \mathcal{V}_q$ is not straightforward as in the cases of time-independent domains. The coupling of time and space in the definition of the dG method on deformable domains forces

us to consider (2.3) [or (2.4)] as a time-space problem. On the contrary, for the corresponding case of non-moving domains, proving the existence of the time discrete dG approximation is equivalent to the existence of the solution of an elliptic problem with homogeneous Dirichlet boundary conditions. We refer to [10, Proposition 3.1] for details on the existence of $U \in \mathcal{V}_q$ satisfying (2.2) and (2.3) [or (2.2)–(2.4)]. We would like to mention though that the key relation for the proof is the discrete Reynolds' identity

$$\begin{aligned}\int_{I_n} \Big(\langle D_t V, V \rangle_{\Omega_t} - \langle \mathbf{w} \cdot \nabla_{\mathbf{x}} V, V \rangle_{\Omega_t} \Big) dt \\ = \frac{1}{2} \|V(t_{n+1})\|^2_{L^2(\Omega_{t_{n+1}})} - \frac{1}{2} \|V(t_n^+)\|^2_{L^2(\Omega_{t_n})},\end{aligned} \tag{2.5}$$

valid for every $V \in \mathcal{V}_q(I_n)$ [10, Lemma 3.1]. This estimate is instrumental to prove the following:

THEOREM 2.1 (Nodal stability with exact integration [10]). *The dG solution $U \in \mathcal{V}_q$ satisfies for $0 \le m < n \le N$:*

$$\begin{aligned}\|U(t_n)\|^2_{L^2(\Omega_{t_n})} + \sum_{j=m}^{n-1} \|U(t_j^+) - U(t_j)\|^2_{L^2(\Omega_{t_j})} + \mu \int_{t_m}^{t_n} \|\nabla_{\mathbf{x}} U(t)\|^2_{L^2(\Omega_t)}\, dt \\ \le \|U(t_m)\|^2_{L^2(\Omega_{t_m})} + \frac{1}{\mu} \int_{t_m}^{t_n} \|f(t)\|^2_{H^{-1}(\Omega_t)}\, dt.\end{aligned} \tag{2.6}$$

The stability estimate (2.6) is the discrete version of (1.9). In particular, it is free from any constants depending on the ALE map, and for $f \equiv 0$ and $m = n - 1$, it implies the discrete monotonicity property of the L^2-norm:

$$\|U(t_n)\|_{L^2(\Omega_{t_n})} \le \|U(t_{n-1})\|_{L^2(\Omega_{t_{n-1}})}, \qquad 1 \le n \le N.$$

This property, also valid for the continuous problem, was not observed in [5, 7, 20].

In the remaining stability analysis of [10] and the error analyses of [8, 9] with exact integration, there appear constants depending explicitly on $\nabla_{\mathbf{y}} \mathcal{A}_t$, the space differential of the ALE map. These constants are of the form

$$A_n \sim \|\nabla_{\mathbf{y}} \mathcal{A}_{t_n \to t}\|_{\mathbf{L}^\infty(I_n;\mathbf{L}^\infty(\Omega_{t_n}))} \|(\nabla_{\mathbf{y}} \mathcal{A}_{t_n \to t})^{-1}\|_{\mathbf{L}^\infty(I_n;\mathbf{L}^\infty(\Omega_{t_n}))}$$

and

$$B_n \sim \|\nabla_{\mathbf{y}} \mathcal{A}_{t_n \to t}\|_{\mathbf{W}^1_\infty(I_n;\mathbf{L}^\infty(\Omega_{t_n}))},$$

$0 \le n \le N$. Here, we do not give precise definitions of A_n, B_n, but rather point out the required regularity on the ALE map, when A_n or B_n appear in the estimates. The regularity assumptions on the family of the ALE map

ensures that the determinant $\det \mathbf{J}_{\mathcal{A}_t}$ of the Jacobian matrix $\mathbf{J}_{\mathcal{A}_t} := \nabla_{\mathbf{y}} \mathcal{A}_t$ is positive and bounded away from 0 and ∞, uniformly for $t \in [0,T]$, [10]. Thus, A_n, B_n are well defined and bounded. It is to be emphasized that $A_n, B_n = \mathcal{O}(1)$ are local and do not involve exponentials of either geometric quantities or T; refer to [10] for details. In the rest of the paper we use the notation $\lesssim$ to indicate absolute constants depending only on the polynomial degree q, the space dimension d as well as constants arising due to the application of the Poincaré inequality.

Next, we state the stability result for $\|U(t)\|_{L^2(\Omega_t)}$ for every $t \in [0,T]$.

THEOREM 2.2 (Global stability with exact integration [10]). *Let $f \in L^2(\mathcal{Q}_T)$ and $\{\mathcal{A}_t\}_{t\in[0,T]}$ be a family of ALE maps. Then, the dG approximation $U \in \mathcal{V}_q$ satisfies for $0 \le n \le N$:*

$$\begin{aligned}\sup_{t\in[0,t_n]} \|U(t)\|^2_{L^2(\Omega_t)} \lesssim & \max_{0\le j\le n-1} \left\{A_j(1+F_j k_j)\right\} \\ & \times \left(\|U(0)\|^2_{L^2(\Omega_0)} + \frac{1}{\mu}\int_0^{t_n} \|f(t)\|^2_{H^{-1}(\Omega_t)}\,dt\right) \quad (2.7)\\ & + \max_{0\le j\le n-1} A_j k_j \int_{I_j} \|f(t)\|^2_{L^2(\Omega_t)}\,dt,\end{aligned}$$

with

$$F_j := B_j + \frac{\|\mathbf{b}-\mathbf{w}\|^2_{L^\infty(\mathcal{Q}_j)}}{\mu} \quad 0 \le j \le n. \tag{2.8}$$

As in the case of non-moving domains [37], to prove an estimate of the form (2.7), we first need to derive a relation between the discrete approximation U and its material derivative $D_t U$. This is possible for the dG approximation in deformable domains as well, because $U \in \mathcal{V}_q$ is a piecewise polynomial of degree at most q when viewed on the reference domain. Therefore, finite-dimensional arguments, such as inverse inequalities, [13, Chap. 4, Lemma 4.5.3], can be applied to relate U with $D_t U$. For details on the proof of Theorem 2.2, we refer to [10]. The stability estimate (2.7) is valid for any choice of the time-steps k_n. However, in contrast to the nodal stability estimate (2.6), estimate (2.7) involves ALE constants. In fact, the global estimate (2.7) suggests that the monotonicity property of $\|U(t)\|_{L^2(\Omega_t)}$ does not hold for all $t \in [0,T]$, but only at the breakpoints t_n. This fact is also observed numerically and is reported in Fig. 1.

2.2. The ALE Projection. Since u does not belong in general to $\mathcal{V}_q$, the derivation of optimal a priori error estimates is achieved by introducing an adequate projection $Pu \in \mathcal{V}_q$ of u. Then, the error $e := u - U$ is decomposed as $e := \rho + \Theta$ with $\rho := u - Pu$ and $\Theta := Pu - U \in \mathcal{V}_q$. Selecting Pu so that ρ has optimal decay in targeted norms (see Proposition 2.2), the a priori error analysis boils down to proving optimal order a priori error estimates for $\Theta \in \mathcal{V}_q$. This is achieved using the stability results of the previous section. We now define Pu.

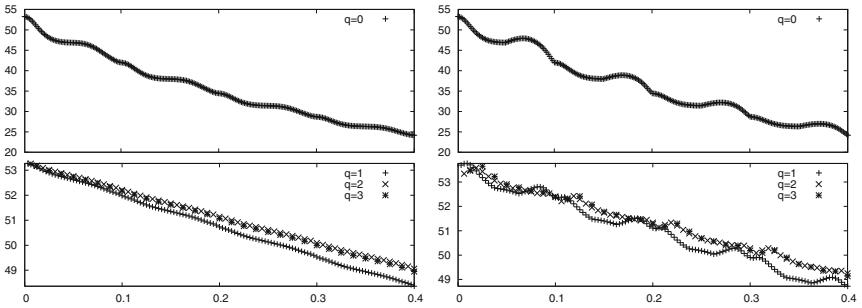


FIG. 1. Evolution of $\|U(t_n)\|_{L^2(\Omega_{t_n})}$ (*left*) and $\max_{t\in I_n}\|U(t)\|_{L^2(\Omega_t)}$ (*right*) for $q=0$ with 2^8 uniform time-steps (*top*) and $q=1,2,3$ with, respectively, 2^7, 2^6, 2^5 uniform time-steps (*bottom*). The space discretization is fine enough not to influence the time discretization. The reference domain is $\Omega_0:=(0,1)\times(0,1)$, the time interval is $[0,0.4]$, the diffusivity is $\mu=0.01$, the domain velocity $\mathbf{w}$ is the L^2-projection over piecewise polynomials of degree q of the time-derivative of the map $(\mathbf{y},t)\mapsto \mathbf{y}(2-\cos(20\pi\, t))$, with $\mathbf{y}\in\Omega_0$, $t\in(0,0.4)$, and the forcing is $f=0$ [21]. The ALE map $\mathcal{A}_t$ is obtained by integrating $\mathbf{w}$ in each time interval I_n, thereby enforcing continuity at the nodes. All schemes display monotone $\|U(t)\|_{L^2(\Omega_t)}$ when restricted to the breakpoints $t=t_n$, as predicted by Theorems 2.1 and 3.1 below, the backward Euler scheme ($q=0$) being much more dissipative than the others ($q>0$). Oscillations of the ALE map destroy this monotonicity property over the whole time interval, thereby corroborating Theorems 2.2 and 3.2.

DEFINITION 2.1 (ALE projection [9]). *For $q>0$, the ALE projection $Pu\in\mathcal{V}_q$ of $u\in C(H^1_0;\mathcal{Q}_T)$ is defined as follows:*

$$Pu(\cdot,0)=u(0),\quad \text{in}\quad \Omega_0, \tag{2.9}$$

and for $0\le n\le N-1$,

$$Pu(\cdot,t_{n+1})=u(\cdot,t_{n+1}),\quad \text{in}\quad \Omega_{t_{n+1}} \tag{2.10}$$

and

$$\int_{I_n}\langle Pu-u,V\rangle_{\Omega_t}\,dt=0,\quad \forall\, V\in\mathcal{V}_{q-1}(I_n), \tag{2.11}$$

while for $q=0$, the latter condition is void.

We refer to Pu as the "ALE projection" of u since, according to (2.11), it is an L^2-type projection over the space-time strip Q_n and therefore defined through the ALE map $\mathcal{A}_t$. More precisely, considering $\Omega_{t_{n+1}}$, $0\le n\le N-1$ as the reference domain and associating with every function $g:\mathcal{Q}_T\to\mathbb{R}$ the function $\hat g:\Omega_{t_{n+1}}\times[0,T]\to\mathbb{R}$ defined by $\hat g(\mathbf{y},t):=g\big(\mathcal{A}_{t_{n+1}\to t}(\mathbf{y}),t\big)$, (2.11) is equivalent to

$$\int_{I_n}\langle(\widehat{Pu}-\hat u)\det\mathbf{J}_{\mathcal{A}_{t_{n+1}\to t}},\hat V\rangle_{\Omega_{t_{n+1}}}=0,\quad \forall\hat V\in\hat{\mathcal{V}}_{q-1}(I_n), \tag{2.12}$$

with

$$\hat{\mathcal{V}}_q(I_n) := \big\{\hat{V} : \ V \in \mathcal{V}_q(I_n)\big\}. \tag{2.13}$$

Equality (2.12) reveals that the operator P is not a pure time operator, as for fixed domains, but it rather depends on the particular family of the ALE maps. Alternatively, we could replace Ω_t by Ω_{t_n} (or $\Omega_{t_{n+1}}$) in (2.11):

$$\int_{I_n} \langle (\widehat{Pu} - \hat{u}), \hat{V} \rangle_{\Omega_{t_n}} = 0, \quad \forall \hat{V} \in \hat{\mathcal{V}}_{q-1}(I_n).$$

In this case, P would be a projection purely in time and the interaction between space and time be avoided. However, the choice of (2.9)–(2.11) is the only one that leads to the following property.

LEMMA 2.1 (Key property of P [9]). *For $0 \le n \le N$, the error $\rho := u - Pu$ satisfies*

$$\begin{aligned} &\int_{I_n} \langle D_t \rho, V \rangle_{\Omega_t} \, dt + \langle \rho(t_n^+) - \rho(t_n), V(t_n^+) \rangle_{\Omega_{t_n}} \\ &\quad + \int_{I_n} \langle \rho \nabla_{\mathbf{x}} \cdot \mathbf{w}, V \rangle_{\Omega_t} \, dt = 0, \ \forall V \in \mathcal{V}_q(I_n). \end{aligned} \tag{2.14}$$

The proof of (2.14) is based on Reynolds' identity (1.7), as well as the definition (2.9)–(2.11) of P. At this point, we emphasize that the orthogonality property (2.14) is essential for the a priori error analysis, as it allows the cancelation of the term $D_t\rho$, which is not expected to be of optimal order of accuracy, as well as the cancelation of the discontinuities $\rho(t_n^+) - \rho(t_n)$ for which an optimal order a priori error bounds are not evident. An orthogonality property similar to (2.14) is used in cases of time-independent domains and the corresponding dG projection, [1, 37]. Actually, having at hand (2.14), the a priori error bounds for the numerical scheme (2.3) can be obtained following the same steps as for non-moving domains. In that respect, the definition of the ALE projection P is a natural generalization of the corresponding dG projection in time-independent domains; cf. e.g., [1, 37].

We next prove the *well posedness* of the ALE projection. The proof of uniqueness is rather easy. Indeed, if $V_1, V_2 \in \mathcal{V}_q$ satisfy (2.10), then $V_1 - V_2 = (t_{n+1} - t)V$ with $V \in \mathcal{V}_{q-1}(I_n)$; this in conjunction with (2.11) leads to $V \equiv 0$, or, $V_1 \equiv V_2$. However, the proof of the existence is technical and requires additional regularity on the ALE map. In contrast to time-independent domains, the condition $u \in C(H_0^1; \mathcal{Q}_n)$ does not imply in general that $(Pu)(t) \in H_0^1(\Omega_t)$ for all $t \in I_n$. As we have shown in [9, Remark 3.1], if the spatial regularity of $\mathcal{A}_t$ is only $\mathbf{W}_\infty^1$, then the H_0^1-norm of $(Pu)(t)$ may blow-up for some $t \in I_n$. The space-time tangling is responsible once again for this difficulty and is a natural consequence of the movement of the domain in time. The ALE projection is not a pure time projection and inherits such an entanglement.

A sufficient condition on the ALE map that guarantees the existence of a $Pu \in \mathcal{V}_q$, whenever $u \in C(H_0^1; \mathcal{Q}_T)$, is the following:

$$\mathcal{A}_{t_{n+1}\to t} \in \mathbf{L}^\infty\big(I_n; \mathbf{W}^2_\infty(\Omega_{t_{n+1}})\big), \qquad 0 \le n \le N-1. \tag{2.15}$$

Let $\mathcal{W}_q$ and $\mathcal{W}_q(I_n)$ be defined as $\mathcal{V}_q$ and $\mathcal{V}_q(I_n)$, respectively, with the difference that H_0^1 is replaced by L^2. Then, the existence of the ALE projection can be established using the next auxiliary lemma:

LEMMA 2.2. *Let $p_n : \mathbb{R} \to \mathbb{R}$ be a nonzero and nonnegative polynomial over I_n of degree s, $s \le q$. Then, for all $w_n \in C(L^2; \mathcal{Q}_n)$ there exists a unique $W_n \in \mathcal{W}_{q-s}(I_n)$ such that*

$$\int_{I_n} p_n(t)\langle W_n(t), V(t)\rangle_{\Omega_t}\, dt = \int_{I_n} \langle w_n(t), V(t)\rangle_{\Omega_t}\, dt, \tag{2.16}$$

for all $V \in \mathcal{W}_{q-s}(I_n)$. If, in addition, the family of ALE maps satisfies (2.15) and $w_n \in C(H_0^1; \mathcal{Q}_n)$, then $W_n \in \mathcal{V}_{q-s}(I_n)$.

Proof. The proof follows the lines of Proposition 3.1 in [9]. More precisely, for the proof of the first assertion, we take $\Omega_{t_{n+1}}$, $0 \le n \le N-1$, as the reference domain, and we let $\hat{\mathcal{W}}_q(I_n) := \{\hat{W} : W \in \mathcal{W}_q(I_n)\}$. Then, since p_n is a nonnegative polynomial of degree $s \le q$, $\hat{\mathcal{W}}_{q-s}(I_n)$ is a Hilbert space, with respect to the inner product $(\cdot,\cdot)_n : [\hat{\mathcal{W}}_{q-s}(I_n)]^2 \to \mathbb{R}$, defined as:

$$(\hat{W}, \hat{V})_n := \int_{I_n} p_n(t)\langle \hat{W} \det \mathbf{J}_{\mathcal{A}_{t_{n+1}\to t}}, \hat{V}\rangle_{\Omega_{t_{n+1}}}, \quad \forall \hat{W}, \hat{V} \in \hat{\mathcal{W}}_{q-s}(I_n).$$

The assumption $w_n \in C(L^2; \mathcal{Q}_n)$ implies that $\hat{w}_n \in C\big(I_n; L^2(\Omega_{t_{n+1}})\big)$. Hence, by Riesz representation Theorem, there exists a unique $\hat{W}_n \in \hat{\mathcal{W}}_{q-s}(I_n)$ such that

$$(\hat{W}_n, V)_n = \int_{I_n} \langle \hat{w}_n \det \mathbf{J}_{\mathcal{A}_{t_{n+1}\to t}}, \hat{V}\rangle_{\Omega_{t_{n+1}}}, \quad \forall \hat{V} \in \hat{\mathcal{W}}_{q-s}(I_n). \tag{2.17}$$

Since (2.17) is equivalent to (2.16), we deduce that there exists a unique $W_n \in \mathcal{W}_{q-s}(I_n)$ satisfying (2.16).

For the proof of the second claim, we write $\hat{W}_n = \sum_{j=0}^{q-s} \hat{W}_{n,j}(t_{n+1}-t)^j$ with $\hat{W}_{n,j} \in L^2(\Omega_{t_{n+1}})$, $0 \le j \le q-s$, because $\hat{W}_n \in \hat{\mathcal{W}}_{q-s}$. Next, we rewrite (2.17) as

$$\int_{\Omega_{t_{n+1}}} \upsilon \int_{I_n} \big(p_n(t)\hat{W}_n - \hat{w}_n\big)(t_{n+1}-t)^i \det \mathbf{J}_{\mathcal{A}_{t_{n+1}\to t}} dt\, d\mathbf{y} = 0,$$

for all $\upsilon \in L^2(\Omega_{t_{n+1}})$ and $0 \le i \le q-s$. For a.e. $\mathbf{y} \in \Omega_{t_{n+1}}$, the above equation is equivalent to the algebraic system for $\hat{\mathbf{W}}_n := (\hat{W}_{n,j})_{j=0}^{q-s}$:

$$\mathbf{A}_n(\mathbf{y})\hat{\mathbf{W}}_n(\mathbf{y}) = \hat{\mathbf{w}}_n(\mathbf{y}),$$

with matrix

$$\mathbf{A}_n(\mathbf{y})_{i,j} := \int_{I_n} p_n(t)(t_{n+1}-t)^{(i-1)+(j-1)} \det \mathbf{J}_{\mathcal{A}_{t_{n+1}\to t}}(\mathbf{y},t)\,dt$$

and right-hand side

$$\hat{\mathbf{w}}_n(\mathbf{y})_i := \int_{I_n} \hat{w}_n(\mathbf{y},t)(t_{n+1}-t)^{i-1} \det \mathbf{J}_{\mathcal{A}_{t_{n+1}\to t}}(\mathbf{y},t)\,dt,$$

for $1 \le i,j \le q-s+1$. The additional regularity (2.15) of $\mathcal{A}_t$ yields that $\mathbf{A}_n$ is Lipschitz continuous and invertible, and that $\mathbf{A}_n^{-1}$ is also Lipschitz (see Step 4 in the proof of [9, Proposition 3.1]), Therefore, using that $\hat{\mathbf{w}}_n \in \mathbf{H}_0^1(\Omega_{t_{n+1}})$, we conclude that $\hat{\mathbf{W}}_n = \mathbf{A}_n^{-1}\hat{\mathbf{w}}_n \in \mathbf{H}_0^1(\Omega_{t_{n+1}})$. This, completes the proof of the second claim. □

Using Lemma 2.2 we deduce:

Proposition 2.1 (Existence of the ALE projection [9]). *Let $\mathcal{A}_t$ satisfy (2.15). Then, for every $u \in C(H_0^1;\mathcal{Q}_T)$ there exists a unique $Pu \in \mathcal{V}_q$ satisfying (2.9)–(2.11).*

Proof. Finding $Pu \in \mathcal{V}_q$ satisfying (2.9)–(2.11) is equivalent to finding $W_n \in \mathcal{V}_{q-1}(I_n)$, $0 \le n \le N-1$, such that

$$\int_{I_n} (t_{n+1}-t)\langle W_n(t), V(t)\rangle_{\Omega_t}\,dt = \int_{I_n} \langle u(t) - u(\mathcal{A}_{t\to t_{n+1}}(\cdot),t_{n+1}), V(t)\rangle_{\Omega_t}\,dt, \quad \forall V \in V_{q-1}(I_n).$$

The asserted claim of the proposition follows immediately from Lemma 2.2 with $p_n(t) := t_{n+1}-t$, $s := 1$ and $w_n := u - u(\mathcal{A}_{t\to t_{n+1}}(\cdot),t_{n+1}) \in C(H_0^1;\mathcal{Q}_n)$. □

We finish the subsection by stating, without proof, the approximation and stability properties of Pu. The proof of the next proposition is based on comparing the ALE projection with the standard dG projection [1, 37]; details can be found in [9].

Proposition 2.2 (Approximation properties and stability of the ALE projection [9]). *Let Pu be the ALE projection defined in (2.9)–(2.11). If the family of ALE maps satisfies (2.15), then, for $t \in I_n$, we have*

$$\|(u-Pu)(t)\|^2_{L^2(\Omega_t)} \le C_n k_n^{2j+1} \int_{I_n} \|D_t^{j+1}u(t)\|^2_{L^2(\Omega_t)}\,dt, \tag{2.18}$$

$$\|\nabla_{\mathbf{x}}(u-Pu)(t)\|^2_{\mathbf{L}^2(\Omega_t)} \le D_n k_n^{2j+1} \times \int_{I_n} \left(\|D_t^{j+1}u(t)\|^2_{L^2(\Omega_t)} + \|\nabla_{\mathbf{x}} D_t^{j+1}u(t)\|^2_{\mathbf{L}^2(\Omega_t)}\right) dt, \tag{2.19}$$

for $0 \le n \le N-1$ and $0 \le j \le q$, where C_n depends on A_n and D_n depends on A_n and $M_n := \|\mathcal{A}_{t_n\to t}\|_{\mathbf{L}^\infty\left(I_n;\mathbf{W}^2_\infty(\Omega_{t_n})\right)}$. In addition, the following stability bounds are valid for P:

$$\int_{I_n} \|D_t^j Pu(t)\|^2_{L^2(\Omega_t)}\,dt \lesssim C_n \int_{I_n} \|D_t^j u(t)\|^2_{L^2(\Omega_t)}\,dt. \tag{2.20}$$

2.3. A Priori Error Analysis. We briefly present now, without proofs, the main results of [9] for the numerical method (2.3) assuming exact integration in time.

As already mentioned in the previous subsection, to obtain an optimal order a priori error bound, we split the error $e = u - U = \rho + \Theta$. The interpolation error $\rho = u - Pu$ is estimated a priori through the approximation properties (2.18) and (2.19). On the other hand, because of the key property (2.14), $\Theta = Pu - U \in \mathcal{V}_q$ satisfies:

$$\begin{aligned}
&\int_{I_n} \langle D_t\Theta, V\rangle_{\Omega_t}\,dt + \langle \Theta(t_n^+) - \Theta(t_n), V(t_n^+)\rangle_{\Omega_{t_n}} \\
&\qquad + \int_{I_n} \langle (\mathbf{b}-\mathbf{w})\cdot\nabla_{\mathbf{x}}\Theta, V\rangle_{\Omega_t}\,dt + \mu\int_{I_n}\langle \nabla_{\mathbf{x}}\Theta, \nabla_{\mathbf{x}}V\rangle_{\Omega_t}\,dt \\
&\quad = \int_{I_n} \langle \rho(\mathbf{b}-\mathbf{w}) - \mu\nabla_{\mathbf{x}}\rho, \nabla_{\mathbf{x}}V\rangle_{\Omega_t}\,dt, \quad \forall V \in \mathcal{V}_q(I_n),
\end{aligned} \tag{2.21}$$

and an optimal order a priori error bound can be established for Θ by applying the stability result (2.6) to (2.21).

THEOREM 2.3 (A priori error estimate for dG in time-dependent domains [9]). *If the family of the ALE maps satisfies* (2.15)*, then the following estimate holds:*

$$\begin{aligned}
&\max_{0\le n\le N} \|(u-U)(t_n)\|^2_{L^2(\Omega_{t_n})} + \mu\int_0^T \|\nabla_{\mathbf{x}}(u-U)(t)\|^2_{\mathbf{L}^2(\Omega_t)}\,dt \\
&\quad \le \frac{1}{\mu}\sum_{n=0}^{N-1} C_n k_n^{2q+2} \sup_{t\in I_n}\|(\mathbf{b}-\mathbf{w})(t)\|^2_{\mathbf{L}^\infty(\Omega_t)} \int_{I_n} \|D_t^{q+1}u(t)\|^2_{L^2(\Omega_t)}\,dt \\
&\qquad + \mu\sum_{n=0}^{N-1} D_n k_n^{2q+2}\int_{I_n}\left(\|D_t^{q+1}u(t)\|^2_{L^2(\Omega_t)} + \|\nabla_{\mathbf{x}}D_t^{q+1}u(t)\|^2_{\mathbf{L}^2(\Omega_t)}\right)dt,
\end{aligned}$$

with u the solution of (1.4) *and U the dG solution of* (2.3)*, and where C_n, D_n, $0 \le n \le N-1$, are constants proportional to those in* (2.18) *and* (2.19)*.*

2.4. The ALE Reconstruction. For the a posteriori error analysis, we follow the reconstruction technique proposed by Akrivis et al. in [2–4, 30] for time discrete schemes on time-independent domains. The main idea is to introduce a continuous approximation $U_{\mathrm{R}} \in \mathcal{V}_{q+1}$ of the dG approximation U of (2.3), called the reconstruction of U, with the following properties:

- $U_{\mathrm{R}}(t_n) = U(t_n), \quad 0 \le n \le N.$
- U_{R} satisfies a perturbation of the original problem (1.4).
- $U_{\mathrm{R}} - U$ is of optimal order of accuracy.

Then, the error $e = u - U$ splits into $u - U_{\mathrm{R}}$ and $U_{\mathrm{R}} - U$ and the final a posteriori error bound is obtained using the stability properties of the continuous equation (1.4).

Following [30], one of the key points to derive a posteriori error estimations is the definition of an appropriate reconstruction of the discrete solution U. Extending the work presented in [30] to moving domains relies mainly on the principle that integration by parts is replaced by Reynolds' identity (1.7). In particular, the reconstruction $U_{\mathrm{R}} \in \mathcal{V}_{q+1}$ of U is defined as follows: For $0 \le n \le N-1$,

$$U_{\mathrm{R}}(t_n^+) = U(t_n) \quad \text{in } \Omega_{t_n} \tag{2.22}$$

and

$$\begin{aligned} &\int_{I_n} \langle D_t U_{\mathrm{R}}, V\rangle_{\Omega_t}\, dt + \int_{I_n} \langle U_{\mathrm{R}} \nabla_{\mathbf{x}} \cdot \mathbf{w}, V\rangle_{\Omega_t}\, dt \\ &\quad = \int_{I_n} \langle D_t U, V\rangle_{\Omega_t}\, dt + \int_{I_n} \langle U \nabla_{\mathbf{x}} \cdot \mathbf{w}, V\rangle_{\Omega_t}\, dt \\ &\qquad + \langle U(t_n^+) - U(t_n), V(t_n^+)\rangle_{\Omega_{t_n}}, \quad \forall V \in \mathcal{V}_q(I_n). \end{aligned} \tag{2.23}$$

Using Lemma 2.2, it is possible to prove that the reconstruction U_{R}, defined through (2.22)–(2.23) is well defined, provided that the family of the ALE maps $\{\mathcal{A}_t\}_{t\in[0,T]}$ satisfies (2.15). The detailed proof of the well posedness of U_{R}, as well as its properties, is discussed in [8].

2.5. A Posteriori Error Analysis. The function U_{R} is instrumental to derive the following bound.

THEOREM 2.4 (A posteriori error estimate for dG in time-dependent domains [8]). *Let U_{R} denote the reconstruction of U defined in (2.22) and (2.23). If the family of ALE maps satisfies (2.15), then the following a posteriori error estimate holds true:*

$$\begin{aligned} &\max_{0\le t\le T} \left\{ \|(u-U_R)(t)\|^2_{L^2(\Omega_t)} + \mu \int_0^t \Big[\|\nabla_{\mathbf{x}}(u-U)(s)\|^2_{L^2(\Omega_s)} \right. \\ &\qquad \left. + \frac{1}{2}\|\nabla_{\mathbf{x}}(u-U_{\mathrm{R}})(s)\|^2_{L^2(\Omega_s)}\Big]\, ds \right\} \\ &\quad \le \mu \int_0^T \|\nabla_{\mathbf{x}}(U-U_{\mathrm{R}})(s)\|^2_{L^2(\Omega_s)}\, ds \\ &\qquad + \frac{4}{\mu}\int_0^t \|\mathcal{E}(s)\|^2_{H^{-1}(\Omega_s)}\, ds + \frac{4}{\mu}\int_0^T \|(f-\Pi_q f)(s)\|^2_{H^{-1}(\Omega_s)}\, ds \end{aligned} \tag{2.24}$$

with

$$\begin{aligned} \mathcal{E} :=& \nabla_{\mathbf{x}} \cdot \big[(\mathbf{b}-\mathbf{w})(U-U_{\mathrm{R}})\big] + \big[U_{\mathrm{R}}\nabla_{\mathbf{x}}\cdot\mathbf{w} - \Pi_q(U_{\mathrm{R}}\nabla_{\mathbf{x}}\cdot\mathbf{w})\big] \\ &+ \Big[\nabla_{\mathbf{x}}\cdot\big[(\mathbf{b}-\mathbf{w})U\big] - \Pi_q\big(\nabla_{\mathbf{x}}\cdot\big[(\mathbf{b}-\mathbf{w})U\big]\big)\Big] + \mu\big[\Pi_q(\Delta_{\mathbf{x}}U) - \Delta_{\mathbf{x}}U\big], \end{aligned}$$

and $\Pi_q : L^2(L^2;\mathcal{Q}_T) \to \mathcal{W}_q$ an L^2-type projection.

We omit giving the precise definition of Π_q, as well as the proof of Theorem 2.4, and we refer to [8] for details. We emphasize though that estimate (2.24) is of the same form as the corresponding one in time-independent domains. The additional terms, appearing in the error indicator $\mathcal{E}$, reflect the geometry of the problem. Note also that because time and space are tangled together, the term $\Pi_q(\Delta_{\mathbf{x}} U) - \Delta_{\mathbf{x}} U$ does not vanish for moving domains, which creates additional difficulties from computational point of view; this issue is discussed in [8]. In addition, in contrast to time-independent domains, for $0 \le n \le N-1$, the difference $U_{\mathrm{R}} - U$ in $\mathcal{Q}_n$, cannot be expressed in terms of the jump estimators

$$J_n(\mathbf{x}, t) := U\big(\mathcal{A}_{t_n} \circ \mathcal{A}_t^{-1}(\mathbf{x}), t_n^+\big) - U\big(\mathcal{A}_{t_n} \circ \mathcal{A}_t^{-1}(\mathbf{x}), t_n\big), \quad (\mathbf{x}, t) \in \mathcal{Q}_n,$$

for $q > 0$. More precisely, for time-independent domains, it can be proven that [30]

$$(U_{\mathrm{R}} - U)(\mathbf{x}, t) := \frac{t - t_n}{k_n} J_n(\mathbf{x}, t), \quad (\mathbf{x}, t) \in \mathcal{Q}_n. \tag{2.25}$$

This is not true for deformable domains due to the tangling of space and time, except for $q = 0$. It can be proven though, [8], that the difference $U_{\mathrm{R}} - U$ in the $L^2(L^2; \mathcal{Q}_n)$ and $L^2(H_0^1; \mathcal{Q}_n)$ can be bounded above by the $L^2(L^2; \mathcal{Q}_n)$ and $L^2(H_0^1; \mathcal{Q}_n)$ norm of the jump estimator J_n multiplied by local constants depending on the ALE map. This is something also observed numerically, as depicted in Fig. 2. Finally, we mention that the a priori error estimate of Theorem 2.3 implies the optimal decay of the left-hand side of (2.24).

2.6. Numerical Experiment. We now check the optimal decay of the proposed a posteriori error estimator. The initial domain is the unit square, $\Omega_0 := (-1, 1) \times (-1, 1)$, and it is deformed according to the ALE map $\mathcal{A}_t(\mathbf{y}) := \mathbf{y}(1 + \frac{1}{2} T_{11}(t))$ for $t \in (0, 99)$, where $T_n(t) := \cos(n \arccos(t))$ is the nth Chebychev polynomial of first kind. We set $\mu = 1.0$ and the exact solution is manufactured to be $u(\mathbf{x}, t) := \exp(x_1\ t) \sin(x_2\ t)$ with $\mathbf{x} := (x_1, x_2) \in \Omega_t$. We consider a naive time adaptivity consisting in the following steps (Dörfler strategy):

- We start with a subdivision of 8 uniform intervals of the time interval $(0, 0.99)$ and compute the space time dG solution.
- We compute the error estimator provided in Theorem 2.4 with the exception of the term involving $\Pi_q(\Delta_{\mathbf{x}} U) - \Delta_{\mathbf{x}} U$, as it is not directly implementable with continuous finite element in space (see above discussion). For later reference we call the resulting quantity the total estimator.
- Using this error estimate, we select a smaller subset of time intervals responsible for 10 % of the total error estimation.
- We bisect these intervals into two equal parts and repeat the process with this new subdivision of $(0, 0.99)$.

Figure 2 displays the errors in the $\ell^\infty(L^2)$ and the $L^2(H^1)$-norms together with the computed total estimator and the total jump estimator $\left(\sum_{i=0}^{N-1}\int_{I_n}(\|J_n(t)\|^2_{L^2(\Omega_t)}+\|\nabla_{\mathbf{x}}J_n(t)\|^2_{L^2(\Omega_t)})\,dt\right)^{1/2}$, all against the number of time-steps used for the computation for different schemes ($q=0,1,2,3$). The space discretization is chosen not to influence the error. The computational rate of convergence is roughly $q+1$ in each case, hence optimal, as depicted in Fig. 2.

We also point out that the adaptive algorithm refines systematically at the end of the interval where the motion is more oscillatory. Figure 3 depicts the Chebyshev polynomial used for the domain evolution together with the subdivision chosen by the algorithm at different refinement cycles for $q=2$.

2.7. Comparison of a Priori and a Posteriori Analyses. We now explain that the a priori and a posteriori analyses are dual versions of one another:

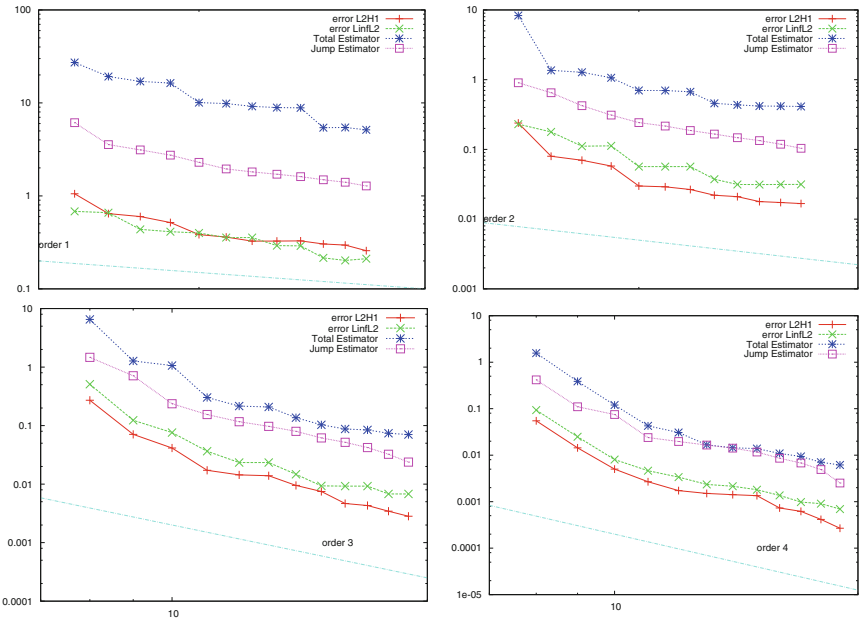


FIG. 2. Errors in the dG approximation together with the total and jump a posteriori estimators are depicted for $q=0$ (*top left*), $q=1$ (*top right*), $q=2$ (*bottom left*) and $q=3$ (*bottom right*). The sequence of time-steps is determined using the adaptive strategy described above. The reference domain is $\Omega_0:=(-1,1)\times(-1,1)$ and it is deformed according to the ALE map $\mathcal{A}_t(\mathbf{y}):=\mathbf{y}(1+\frac{1}{2}\,T_{11}(t))$ for $t\in(0,0.99)$, where $T_n(t):=\cos(n\arccos(t))$ is the nth Chebychev polynomial of first kind. The discretization in space is chosen sufficiently fine, not to influence the time discretization error. All quantities exhibit an approximate decay of $O(N^{-(q+1)})$.

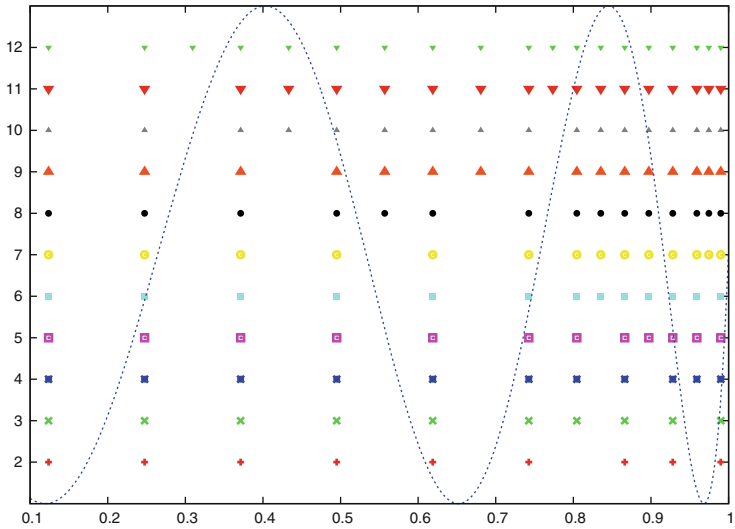


FIG. 3. Subdivision of the time interval during the adaptive procedure (*from bottom to top*) for $q = 2$. Initially eight intervals are considered and the algorithm select for refinement the smallest amount of intervals contributing for 10 % of the total estimator. The Chebychev polynomial used for the domain deformation is plotted in *bold line* for comparison.

- *A priori error analysis*
  - Find $Pu \in \mathcal{V}_q$, the ALE projection of u, such that $u - Pu$ is of optimal order of accuracy in $L^\infty(L^2)$ and $L^\infty(H^1_0)$.
  - Use the stability of the numerical method (2.3) to bound $Pu - U$ a priori, because $Pu - U \in \mathcal{V}_q$ satisfies an error equation at the discrete level.
  - Estimate the error $u - U$ via $u - Pu$ and $Pu - U$.

- *A posteriori error analysis*
  - Find $U_{\mathrm{R}} \in \mathcal{V}_{q+1}$, the reconstruction of U, such that $U_{\mathrm{R}} - U$ is of optimal order of accuracy in $L^\infty(L^2)$ and $L^\infty(H^1_0)$.
  - Use the stability of the continuous PDE (1.4) to bound $u - U_{\mathrm{R}}$ a posteriori, because $u - U_{\mathrm{R}} \in C(H^1_0)$ satisfies an error equation at the continuous level.
  - Estimate the error $u - U$ via $U_{\mathrm{R}} - U$ and $u - U_{\mathrm{R}}$.

A priori analysis	A posteriori analysis
$\mathcal{V}_q$	$C(H^1_0)$
Pu	U_{R}
$Pu - U$	$u - U_{\mathrm{R}}$

3. Practical Algorithms: Reynolds' Methods. In the previous section, we reviewed the results of [8–10] related to dG methods of any order in time within the ALE framework. These methods enjoy the same stability properties as the continuous problem (1.4) and lead to optimal order a priori and a posteriori error bounds. However, the proposed methods are not practical since to implement them we need to employ appropriate quadrature in time. This raises the following questions:

Does there exist a quadrature in time that when applied to the numerical scheme (2.3) *(or equivalently to* (2.4)*) leads to similar stability properties and error bounds as those of the previous section? If so, how to construct such a quadrature?*

The key observation for quadrature is to preserve the Reynolds' identity (2.5) [10]. This is possible, provided the ALE map is a continuous piecewise polynomial in time. The associated quadrature is then called *Reynolds' quadrature*.

In the sequel, we briefly present stability results and a priori error estimates of [9, 10] for Reynolds' methods and polynomial in time ALE maps. At the end of the section, we describe how we handle cases of non-polynomial ALE maps.

3.1. Reynolds' Quadrature and Stability. We assume that the ALE map is a continuous piecewise polynomial of degree $\leq q'$ in time. Since the key ingredient for the validity of the stability estimate (2.6) is Reynolds' identity (2.5), we use quadratures in time of sufficiently high order to keep (2.5) valid. We refer to such quadratures as Reynolds' quadratures chosen so that for $t \in I_n$, $0 \leq n \leq N-1$, the integrals

$$\int_{\Omega_t} D_t V W \, d\mathbf{x}, \qquad \int_{\Omega_t} (\nabla_{\mathbf{x}} \cdot \mathbf{w}) V W \, d\mathbf{x}, \tag{3.1}$$

appearing in (2.5) are computed exactly for $V, W \in \mathcal{V}_q(I_n)$. As shown in [10], the integrals in (3.1) are polynomials of degree

$$p := 2q + dq' - 1 \tag{3.2}$$

in time if $q' \geq 1$. If $q' = 0$, the second term in (3.1) vanishes and the first term is a polynomial degree p in time, provided $q \geq 1$, and vanishes otherwise. Taking into account these observations, we propose the following definition of Reynolds' quadratures:

Definition 3.1 (Reynolds' quadrature [10]). *We say that a quadrature Q on $(0,1]$ with positive weights ω_j and nodes τ_j, $j = 0, 1, \ldots, r$, is a* Reynolds' quadrature *if it is exact for polynomials of degree p defined in* (3.2). *The corresponding quadrature in $I_n = (t_n, t_{n+1}]$, $0 \leq n \leq N-1$, is denoted by Q_n and the corresponding weights $\{\omega_{n,j}\}_{j=0}^r$ and quadrature points $\{t_{n,j}\}_{j=0}^r$ in $I_n = (t_n, t_{n+1}]$ are given by*

$$\omega_{n,j} = k_n \omega_j, \quad t_{n,j} = t_n + k_n \tau_j, \quad 0 \leq j \leq r.$$

Applying a Reynolds' quadrature to (2.5) we obtain the discrete Reynolds' identity [10, Lemma 4.2]

$$\begin{aligned}\frac{1}{2}\|V(t_{n+1})\|^2_{L^2(\Omega_{t_{n+1}})} - \frac{1}{2}\|V(t_n^+)\|^2_{L^2(\Omega_{t_n})} \\ = Q_n\big(\langle D_t V - \mathbf{w}\cdot\nabla_{\mathbf{x}} V, V\rangle_{\Omega_t}\big).\end{aligned} \tag{3.3}$$

For $q = 0$, (3.3) is the geometric conservation law (GCL) appearing in [7, 20–22, 31]. In that respect, (3.3) may be regarded as a generalization of the GCL to higher order dG methods $q > 0$ and test functions with nonvanishing material derivative.

In addition, applying a Reynolds' quadrature to the non-conservative dG formulation (2.3), we get

$$\begin{aligned}&Q_n\big(\langle D_t U, V\rangle_{\Omega_t}\big) + \langle U(t_n^+) - U(t_n), V(t_n^+)\rangle_{\Omega_{t_n}} \\ &\quad + Q_n\big(\langle(\mathbf{b}-\mathbf{w})\cdot\nabla_{\mathbf{x}} U, V\rangle_{\Omega_t}\big) + \mu Q_n\big(\langle\nabla_{\mathbf{x}} U, \nabla_{\mathbf{x}} V\rangle_{\Omega_t}\big) \\ &\quad = Q_n\big(\langle f, V\rangle_{\Omega_t}\big), \quad \forall V \in \mathcal{V}_q(I_n),\end{aligned} \tag{3.4}$$

whereas the conservative dG formulation (2.4) can be written as

$$\begin{aligned}&\langle U(t_{n+1}), V(t_{n+1})\rangle_{\Omega_{t_{n+1}}} - \langle U(t_n), V(t_n^+)\rangle_{\Omega_{t_n}} \\ &\quad + Q_n\big(\langle\nabla_{\mathbf{x}}\cdot\big((\mathbf{b}-\mathbf{w})U\big), V\rangle_{\Omega_t}\big) + \mu Q_n(\langle\nabla_{\mathbf{x}} U, \nabla_{\mathbf{x}} V\rangle_{\Omega_t}) \\ &\quad - Q_n(\langle U, D_t V\rangle_{\Omega_t}) = Q_n(\langle f, V\rangle_{\Omega_t}), \quad \forall V \in \mathcal{V}_q(I_n).\end{aligned} \tag{3.5}$$

If Q_n is the Reynolds' quadrature, then we can prove that the non-conservative and conservative formulations (3.4) and (3.5) are equivalent. Moreover, for $q = 0$ and the mid-point integration rule, (3.5) reduces to the unconditionally stable backward Euler method proposed by Formaggia and Nobile in [20]. We refer to [10] for a detail discussion of these topics, as well as the proof of the next stability result [compare estimate (3.6) below with estimate (2.6)].

THEOREM 3.1 (Nodal stability with Reynolds' quadrature [10]). *Let $f \in C(H^{-1}; \mathcal{Q}_T) \cap L^2(\mathcal{Q}_T)$ and the ALE map $\mathcal{A}_t$ be a continuous piecewise polynomial in time of degree q'. Let $U \in \mathcal{V}_q$ be the solution of problem (3.4) or (3.5), together with (2.2), using a Reynolds' quadrature Q_n over I_n. If $0 \le m < n \le N$, then*

$$\begin{aligned}&\|U(t_n)\|^2_{L^2(\Omega_{t_n})} + \sum_{j=m}^{n-1}\|U(t_j^+) - U(t_j)\|^2_{L^2(\Omega_{t_j})} \\ &\quad + \mu\sum_{j=m}^{n-1} Q_j(\|\nabla_{\mathbf{x}} U(t)\|^2_{L^2(\Omega_t)}) \\ &\quad \le \|U(t_m)\|^2_{L^2(\Omega_{t_m})} + \frac{1}{\mu}\sum_{j=m}^{n-1} Q_j(\|f(t)\|^2_{H^{-1}(\Omega_t)}).\end{aligned} \tag{3.6}$$

Theorem 3.1 provides an unconditional stability estimate of the discrete $L^2(H^1)$-norm. A similar estimate for the continuous $L^2(H^1)$-norm can be obtained using the equivalence of the discrete and continuous norms in conjunction with (3.6) [10, Lemma 4.3, Theorem 4.2]. The stability estimate in the continuous energy norm is needed for the derivation of optimal order a priori error bounds for the numerical scheme (3.4) or (3.5).

Finally, following similar arguments as for the proof of Theorem 2.2 and accounting for the quadrature error for the terms that are not integrated exactly in (3.4) or (3.5), it is possible to derive a stability estimate in the whole time interval I_n [10, Theorem 4.3].

THEOREM 3.2 (Global stability with Reynolds' quadrature [10]). *Let $f \in C(L^2)$ and the ALE map $\mathcal{A}_t$ be a continuous piecewise polynomial of degree q'. Then the solution $U \in \mathcal{V}_q$ of either (3.4) or (3.5), together with (2.2), satisfies for $1 \le n \le N$, the following stability result*

$$\begin{aligned}\sup_{t\in[0,t_n]} \|U(t)\|^2_{L^2(\Omega_t)} \lesssim & \max_{0\le j\le n-1}\{A_j(1+k_jF_j)\} \\ & \times \Big(\|U(0)\|^2_{L^2(\Omega_0)} + \frac{1}{\mu}\sum_{j=0}^{n-1} Q_j(\|f(t)\|^2_{H^{-1}(\Omega_t)})\Big) \\ & + \max_{0\le j\le n-1} k_jA_jQ_j\Big(\|f(t)\|^2_{L^2(\Omega_t)}\Big),\end{aligned} \tag{3.7}$$

where F_j is defined in (2.8).

Note that estimate (3.7) is the discrete analogue (in terms of quadrature) of estimate (2.7).

3.2. Error Analysis for Polynomial ALE Maps. Since the stability estimate (3.6) is valid for a Reynolds' quadrature, we expect to be able to prove optimal order a priori error estimates with the aid of the ALE projection Pu (see Definition 2.1), provided that the ALE map is a continuous piecewise polynomial of degree q' in time.

Indeed, in this case, the error $\Theta = Pu - U$ satisfies the equation

$$\begin{aligned}& Q_n(\langle D_t\Theta, V\rangle_{\Omega_t}) + \langle \Theta(t_n^+) - \Theta(t_n), V(t_n^+)\rangle_{\Omega_{t_n}} \\ & \quad + Q_n(\langle(\mathbf{b}-\mathbf{w})\cdot\nabla_{\mathbf{x}}\Theta, V\rangle_{\Omega_t}) + \mu Q_n(\langle\nabla_{\mathbf{x}}\Theta, \nabla_{\mathbf{x}}V\rangle_{\Omega_t}) \\ & = Q_n(\langle(\mathbf{b}-\mathbf{w})\rho, \nabla_{\mathbf{x}}V\rangle_{\Omega_t}) - \mu Q_n(\langle\nabla_{\mathbf{x}}\rho, \nabla_{\mathbf{x}}V\rangle_{\Omega_t}) \\ & \quad + E_n(\langle(\mathbf{b}-\mathbf{w})u, \nabla_{\mathbf{x}}V\rangle_{\Omega_t}) - \mu E_n(\langle\nabla_{\mathbf{x}}u, \nabla_{\mathbf{x}}V\rangle_{\Omega_t}) + E_n(\langle f, V\rangle_{\Omega_t}),\end{aligned} \tag{3.8}$$

for all $V \in \mathcal{V}_q(I_n)$ and where $E_n(\cdot) := \int_{I_n}\cdot - Q_n(\cdot)$ denotes the quadrature error over I_n. Equation (3.8) is similar to the error equation (2.21), except that $\int_{I_n}$ is replaced by Q_n and the last three terms on the right-hand side reflect the effect of quadrature.

The quadrature error that appears on the right-hand side of (3.8) has to be of order $q+1$ (the order of the dG method). To achieve this, we need Reynolds' quadrature satisfying

$$q \le r \le p - q. \tag{3.9}$$

We point out that Reynolds' quadrature satisfying (3.9) do exist. For example, we could use $r+1$ Radau or Gauss quadrature points with $r = q + [\frac{dq'}{2}]$, where $[\cdot]$ denotes the integer part [9, Sect. 2].

We next state a result analogous to Theorem 2.3, corresponding to Reynolds' quadratures and piecewise polynomial ALE maps.

THEOREM 3.3 (A priori error estimate with Reynolds' quadrature [9]). *Let the ALE map $\mathcal{A}_t$ be a piecewise polynomial in time of degree q' and satisfy (2.15). Let U be the solution of (2.2)–(3.4) with Q_n a Reynolds' quadrature satisfying (3.9). Then the following a priori error estimate holds*

$$\begin{aligned} &\max_{0\le n\le N} \|(u-U)(t_n)\|^2_{L^2(\Omega_{t_n})} \\ &\quad + \frac{\mu}{2}\sum_{n=0}^{N-1} Q_n(\|\nabla_{\mathbf{x}}(u-U)(t)\|^2_{\mathbf{L}^2(\Omega_t)}) \le \sum_{i=1}^{5} \mathcal{E}_i^2, \end{aligned} \tag{3.10}$$

with

$$\mathcal{E}_1^2 := \frac{1}{\mu}\sum_{n=0}^{N-1} C_n k_n^{2q+2} \sup_{t\in I_n} \|(\mathbf{b}-\mathbf{w})(t)\|^2_{\mathbf{L}^\infty(\Omega_t)} \int_{I_n} \|D_t^{q+1}u(t)\|^2_{L^2(\Omega_t)}\,dt$$

$$\mathcal{E}_2^2 := \frac{\mu}{2}\sum_{n=0}^{N-1} D_n k_n^{2q+2} \int_{I_n} \left(\|D_t^{q+1}u(t)\|^2_{L^2(\Omega_t)} + \|\nabla_{\mathbf{x}} D_t^{q+1}u(t)\|^2_{\mathbf{L}^2(\Omega_t)}\right) dt$$

$$\mathcal{E}_3^2 := \frac{1}{\mu}\sum_{n=0}^{N-1} G_{n,q+1} k_n^{2q+2} \sum_{i=0}^{q+1} \int_{I_n} \|D_t^q\Big((\mathbf{b}-\mathbf{w})u\Big)(t)\|^2_{\mathbf{L}^2(\Omega_t)}\,dt$$

$$\mathcal{E}_4^2 := \mu \sum_{n=0}^{N-1} G_{n,j+1} k_n^{2j+2} \sum_{i=0}^{j+1} \int_{I_n} \|\nabla_{\mathbf{x}} D_t^i u(t)\|^2_{\mathbf{L}^2(\Omega_t)}\,dt$$

$$\mathcal{E}_5^2 := \frac{1}{\mu}\sum_{n=0}^{N-1} G_{n,q+1} k_n^{2q+2} \sum_{i=0}^{q+1} \int_{I_n} \|D_t^i f(t)\|^2_{H^{-1}(\Omega_t)}\,dt,$$

and constants C_n, D_n, $0 \le n \le N-1$, proportional to those in (2.18) and (2.19), respectively, and

$$G_{n,q+1} := A_n B_{n,q+1}, \qquad B_{n,q+1} := \|\nabla_{\mathbf{y}} \mathcal{A}_{t_n\to t}\|_{W^{q+1}_\infty(I_n;L^\infty(\Omega_{t_n}))}. \tag{3.11}$$

Note that estimate (3.10) holds for any choice of the time-steps k_n (unconditional a priori error bound). Also, as discussed in the previous subsection, it can be proven that the continuous and discrete energy norms

are equivalent. Thus, an a priori error estimate similar to (3.10) can be derived for the continuous energy norm as well.

The derivation of optimal order *a posteriori error estimates* is a very interesting and nontrivial question. Despite the fact that the definition (2.22) and (2.23) of the reconstruction remains the same for Reynolds' methods, the analysis is not a direct generalization of the dG methods with exact integration in time. Since the main error analysis is performed at the continuous level instead of the discrete, several projections must be used to handle the terms that are not integrated exactly in (3.4). It turns out that they are not pure time-projections, as it happens for the ALE projection and the reconstruction, which adds additional technical difficulties and makes the analysis tedious. It is, however, possible to prove rigorously optimal order a posteriori error bounds for dG methods with Reynolds' quadrature.

3.3. Non-polynomial ALE Maps. The previous subsection was devoted to stability and error analysis for practical Reynolds' algorithms for problem (1.4) written on the ALE framework, but under the assumption that the ALE map is continuous piecewise polynomial in time. Since this is not always realistic, the following question arises:

How to apply the previous analysis to non-polynomial in time ALE maps?

The answer to this question is briefly discussed below and details are given in [9, Sect. 5].

We approximate the domain velocity $\hat{\mathbf{w}}$ in the ALE frame by $\hat{\mathbf{W}}$ using a piecewise polynomial in time of order q, i.e., $\hat{\mathbf{W}} \in \hat{\mathcal{V}}_q(I_n)$. This creates a new family $\{\tilde{\mathcal{A}}_t\}_{t\in[0,T]}$ of ALE maps that corresponds to $\hat{\mathbf{W}}$ and defined by $\tilde{\mathcal{A}}_0 = I_d$ and for $0 \le n \le N-1$,

$$\tilde{\mathcal{A}}_t(\mathbf{y}) = \tilde{\mathcal{A}}_{t_n}(\mathbf{y}) + \int_{t_n}^{t} \hat{\mathbf{W}}(\mathbf{y}, s)\, ds, \qquad \tilde{\mathbf{x}}(\mathbf{y}, t) := \tilde{\mathcal{A}}_t(\mathbf{y}). \tag{3.12}$$

For every $t \in [0, T]$, $\hat{\mathbf{W}}$ also creates a perturbed domain $\tilde{\Omega}_t = \tilde{\mathcal{A}}_t(\Omega_0)$. Since $\hat{\mathbf{W}}$ is a piecewise polynomial in time of degree q, the definition (3.12) of $\tilde{\mathcal{A}}_t$ implies that $\tilde{\mathcal{A}}_t$ is a continuous piecewise polynomial in time of degree $q+1$.

The idea is to define the discrete dG space and the dG method with Reynolds' quadrature with solution $\tilde{U}$ with respect to the perturbed domain $\tilde{\Omega}_t$ (and the perturbed ALE map $\tilde{\mathcal{A}}_t$). Since the solution u of (1.4) is defined in Ω_t and $\tilde{U}$ in $\tilde{\Omega}_t$, which are different but close domains, the key point of the analysis is to write u with respect to $\tilde{\mathcal{A}}_t$ and denote it $\tilde{u}$. Then, it is possible to prove, using a perturbation argument, that $\tilde{u}$ satisfies an equation of the form (1.4) with a defect of optimal order of accuracy. The defect includes geometric quantities, due to the approximation of Ω_t by $\tilde{\Omega}_t$. Enforcing the geometrical defect to be of optimal order of accuracy in time,

entails approximating $\hat{\mathbf{w}}$ by a piecewise polynomial of degree q. Finally, proceeding as in the previous subsection, an a priori error estimate similar to (3.10) is derived in [9]. As expected, the upper bound contains additional error terms due to the domain approximation. The computational rates of convergence depicted in Fig. 4 corroborate theory.

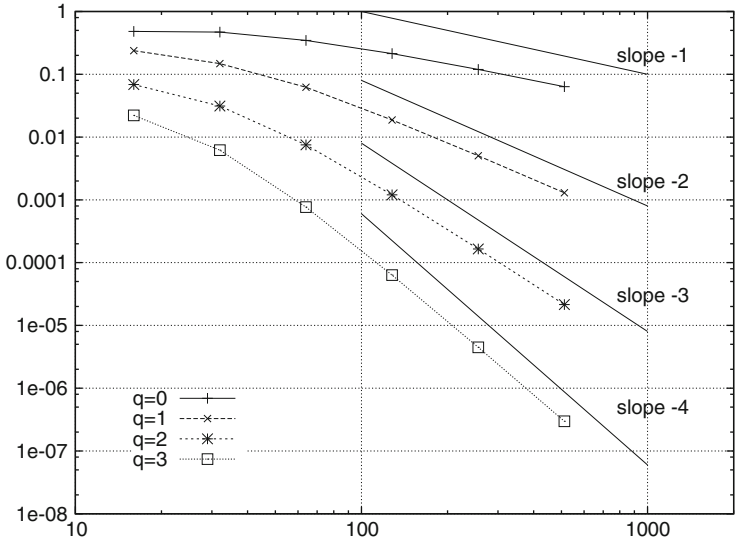


FIG. 4. Error in the $\ell^\infty(L^2)$-norm against the number N of uniform time-steps is depicted for $q = 0, 1, 2, 3$. The reference domain is $\Omega_0 := (-1, 1) \times (-1, 1)$ and it is deformed according to the ALE map $\mathcal{A}_t(\mathbf{y}) := \mathbf{y}(1 + \frac{1}{2}\, T_{11}(t))$ for $t \in (0, 0.99)$, where $T_n(t) := \cos(n \arccos(t))$ is the nth Chebychev polynomial of first kind. We set $\mu = 1.0$ and the exact solution is manufactured to be $u(\mathbf{x}, t) := \exp(x_1\, t) \sin(x_2\, t)$ with $\mathbf{x} := (x_1, x_2) \in \Omega_t$. We take $\hat{\mathbf{W}}$ to be the L^2-projection of the ALE velocity $\hat{\mathbf{w}}$ onto $\hat{\mathcal{V}}_q$ and compute $\tilde{\mathcal{A}}_t$ according to (3.12). Over each interval I_n, a Reynolds' quadrature based on $q+1+[d(q+1))/2]$ Radau points is used, cf. (3.9). The discretization in space is chosen sufficiently fine not to influence the time discretization error. All schemes exhibit the optimal $\mathcal{O}(N^{-(q+1)})$ order of convergence.

4. Practical Algorithms: Runge–Kutta–Radau Methods. In the previous section, we used Reynolds' quadrature to approximate the integrals in time, appearing in dG method (2.3) [or (2.4)]. Using such quadratures leads to unconditional stability and a priori error bounds. However, Reynolds' quadratures are dimensional dependent and become computationally more intensive for higher dimensions. Indeed, a Reynolds' quadrature integrates exactly polynomials of degree $p = 2q + dq' - 1$, where q' is the degree of the polynomial ALE map [see (3.2)].

4.1. Runge–Kutta–Radau Methods. We now review the results of [9, 10] related to RKR methods, which in the ALE framework can be obtained from (2.3) by applying the Radau quadrature with $q + 1$ nodes.

Such a quadrature is enough for dG methods to be unconditionally stable when the domain is not moving and give optimal order a priori error estimates [37, Chap. 12]. Notice that $q+1$ Radau nodes integrate exactly polynomials of degree $\leq 2q$, which compares favorably with p in (3.2) for $q' \geq 1$. Moreover, if the domain does not move ($q' = 0$), then Reynolds' quadrature is exact for polynomials of degree $p = 2q-1$ which can be realized with $q+1$ Radau nodes.

As discussed in [10], the use of the Radau quadrature with $q+1$ nodes leads to stable practical numerical methods, subject to a *mild constraint on the time-steps depending on the ALE map (conditional stability)*. In [9], we were also able to prove that RKR methods on the ALE framework lead to optimal order a priori error bounds, but under the same time-step restriction as for the stability. The reason for this time-step constraint is the violation of Reynolds' identity when using $q+1$ Radau quadrature points for the approximation of the integrals in (2.5). However, it is to be emphasized that the time-step restriction is *not a CFL condition*, as the considered numerical schemes are only discrete in time, i.e., the space is continuous. Despite the fact that RKR methods in the ALE framework lead to conditional stability, these methods have some advantages in comparison with Reynolds' methods. Inevitably, a natural question arises:

> *Which family of methods is more appropriate for problems defined on time-dependent domains and the ALE framework: Reynolds' or RKR methods?'*

Reynolds' methods are more appropriate when dealing with highly oscillatory ALE maps, because they lead to ALE-free stable schemes. On the contrary, the minimal complexity of RKR methods makes them more appropriate in cases of non-oscillatory maps, when the time-step requirement is less restrictive and in practice unnoticeable.

We continue now with a brief description, without proofs, of the main results of [9, 10] regarding the analysis of RKR methods in the ALE framework. Our analysis is, in some sense, related to the one by Badia and Codina [5], who proposed first and second order accurate BDF schemes in the ALE framework for time-dependent domains. Their schemes do not satisfy the GCL and are stable and optimally accurate under a time-step constraint similar to ours.

Let ω_j and τ_j, $0 \leq j \leq q$, be the weights and nodes, respectively, for the Radau quadrature rule Q^q in $(0,1]$ and let $\{\omega_{n,j}\}_{j=0}^q$ and $\{\tau_{n,j}\}_{j=0}^q$ be those for I_n, $0 \leq n \leq N-1$. Using such a quadrature in (2.3), say Q_n^q, the RKR method in the non-conservative ALE framework reads as:

$$\begin{aligned} &Q_n^q(\langle D_t U, V\rangle_{\Omega_t}) + \langle U(t_n^+) - U(t_n), V(t_n^+)\rangle_{\Omega_{t_n}} \\ &\quad + Q_n^q(\langle(\mathbf{b}-\mathbf{w})\cdot\nabla_{\mathbf{x}} U, V\rangle_{\Omega_t}) + \mu Q_n^q(\langle\nabla_{\mathbf{x}} U, \nabla_{\mathbf{x}} V\rangle_{\Omega_t}) \\ &= Q_n^q(\langle f, V\rangle_{\Omega_t}), \quad \forall V \in \mathcal{V}_q(I_n), \end{aligned} \tag{4.1}$$

with $U(\cdot,0) = u_0$ in Ω_0. We point out that for RKR methods, conservative and non-conservative formulations are no longer equivalent. This is because, in contrast to Reynolds' quadrature Q_n, Radau quadrature Q_n^q does not integrate exactly the terms appearing in Reynolds' identity (2.5). Nevertheless, similar stability results and a priori error bounds are valid for both RKR methods.

To compensate for the extra variational crime, introduced due to the violation of Reynolds' identity, we need to impose an extra local-time regularity condition on the family of the ALE maps $\{\mathcal{A}_t\}_{t\in[0,T]}$:

$$B_{n,2} := \|D\mathcal{A}_{t_n\to t}\|_{\mathbf{W}^2_\infty(I_n;\mathbf{L}^\infty(\Omega_{t_n}))} < \infty. \tag{4.2}$$

Using similar arguments as in the proof of Theorem 3.1 and the Bramble–Hilbert Theorem to bound the quadrature error of the terms appearing in Reynolds' identity, it is possible to prove the following theorem:

THEOREM 4.1 (Conditional nodal stability for RKR [10]). *Let $f \in C(H^{-1};\mathcal{Q}_T)\cap L^2(\mathcal{Q}_T)$ and (4.2) be valid. If*

$$A_n(1+B_{n,2})k_n \lesssim \mu, \quad \forall\, 0\le n<N, \tag{4.3}$$

then the solution $U\in\mathcal{V}_q$ of problem (2.2) and (4.1) satisfies, for $0\le m<n\le N$,

$$\begin{aligned}\|U(t_n)\|^2_{L^2(\Omega_{t_n})} &+ \sum_{j=m}^{n-1}\|U(t_j^+)-U(t_j)\|^2_{L^2(\Omega_{t_j})}\\ &+\mu\sum_{j=m}^{n-1} Q_j(\|\nabla_{\mathbf{x}}U(t)\|^2_{\Omega_{t_j}})\\ &\le \|U(t_m)\|^2_{L^2(\Omega_{t_m})}+\frac{2}{\mu}\sum_{j=m}^{n-1}Q_j(\|f(t)\|^2_{H^{-1}(\Omega_t)}).\end{aligned} \tag{4.4}$$

Figure 5 documents the behavior of $\|U(t_n)\|_{L^2(\Omega_{t_n})}$ for RKR methods of order $0\le q\le 3$ and the same oscillatory case of [21], already displayed in Fig. 1.

The numerical experiments depicted in Fig. 5 illustrate that for $f\equiv 0$, the monotonicity of $\|U(t_n)\|_{L^2(\Omega_{t_n})}$ is retained for $q=0$ provided that the time-step constraint (4.3) is enforced. However, in this particular example, it seems that the case $q>0$ does not require any time-step constraints to exhibit the monotone behavior.

Global stability in the whole time-interval I_n is also valid for RKR methods, as it happens for dG methods with exact integration and Reynolds' quadrature, provided that the time-steps are chosen so that (4.3) is satisfied.

Regarding the a priori error analysis, the error $\Theta = Pu - U$ satisfies an equation similar to (3.8) with the additional quadrature error terms

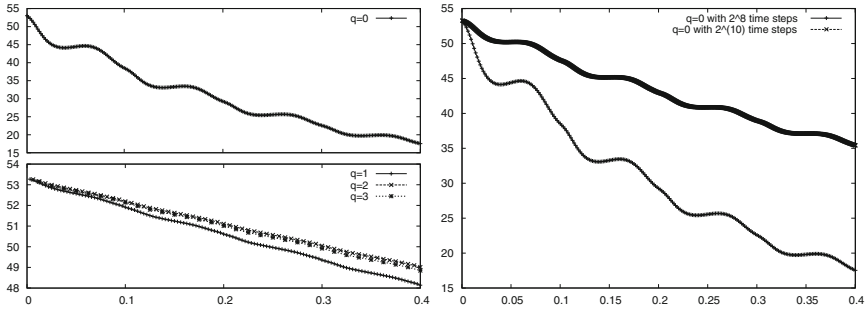


FIG. 5. Evolution of $\|U(t_n)\|_{L^2(\Omega_{t_n})}$ for $q = 0$ with 2^8 uniform time-steps (*top-left*), for $q = 1, 2, 3$ with 2^7, 2^6, 2^5 uniform time-steps, respectively (*bottom-left*), and for $q = 0$ with 2^8 (*upper curve*) and 2^{10} (*lower curve*) uniform time-steps (*right*). The space discretization is fine enough not to influence the time discretization. The reference domain is $\Omega_0 := (0,1) \times (0,1)$, the time interval is $[0, 0.4]$, the diffusivity is $\mu = 0.01$, the domain velocity $\mathbf{w}$ is the L^2-projection over piecewise polynomials of degree q of the time-derivative of the map $(\mathbf{y}, t) \mapsto \mathbf{y}(2 - \cos(20\pi\, t))$, with $\mathbf{y} \in \Omega_0$, $t \in (0, 0.4)$, and the forcing is $f = 0$. The ALE map $\mathcal{A}_t$ is obtained by integrating $\mathbf{w}$ in each time interval I_n, enforcing continuity at the nodes. Monotonicity of $\|U(t_n)\|_{L^2(\Omega_{t_n})}$ for $q = 0$ is sensitive to the time-step size (conditional stability), a property of RKR methods proved in Theorem 4.1 for all $q \geq 0$ (see *right*). Stability of higher order RKR methods ($q > 0$) is less sensitive to the time-steps (*bottom-left*).

$E_n(\langle D_t Pu, V\rangle_{\Omega_t})$ and $E_n(\nabla_{\mathbf{x}} \cdot \mathbf{w} Pu, V\rangle_{\Omega_t})$. These terms arise because RKR methods violate Reynolds' identity. Estimating these quadrature errors leads to derivatives $D_t^i Pu$ for $0 \leq i \leq q+1$, which are handled via the stability bounds in (2.20) for the ALE projection Pu. Thus, proceeding as in the proof of Theorem 3.3, and using the stability estimate (4.4), we managed in [9] to prove the following:

THEOREM 4.2 (A priori error estimate for RKR methods [9]). *Let $U \in \mathcal{V}_q$ be the solution of (4.1) with $q+1$ Radau quadrature points and let $\mathcal{A}_t$ satisfy (4.2). If the time-steps satisfy (4.3), then the following error estimate for $u - U$ is valid:*

$$\begin{aligned}
&\max_{0\leq n\leq N} \|(u-U)(t_n)\|^2_{L^2(\Omega_{t_n})} + \frac{\mu}{2}\sum_{n=0}^{N-1} Q_n\big(\|\nabla_{\mathbf{x}}(u-U)(t)\|^2_{\mathbf{L}^2(\Omega_t)}\big) \\
&\quad \leq \mathcal{E}(u, f, \mathcal{A}_t, \mathbf{b}) + \frac{1}{\mu}\sum_{n=0}^{N-1} C_n G_{n,q+1} k_n^{2q+2} \\
&\qquad \times \sum_{i=0}^{q+1} \int_{I_n} \Big(1 + \|D_t^{q+1-i}\nabla_{\mathbf{x}}\mathbf{w}(t)\|^2_{\mathbf{L}^\infty(\Omega_t)}\Big)\|D_t^i u(t)\|^2_{L^2(\Omega_t)}\, dt,
\end{aligned} \tag{4.5}$$

where $\mathcal{E}(u, f, \mathcal{A}_t, \mathbf{b})$ denotes the right-hand side of (3.10) (with proportionality constants), Pu is the ALE projection defined in (2.10)–(2.11) and $C_n, G_{n,q+1}$ are as in (2.18) and (3.11), respectively.

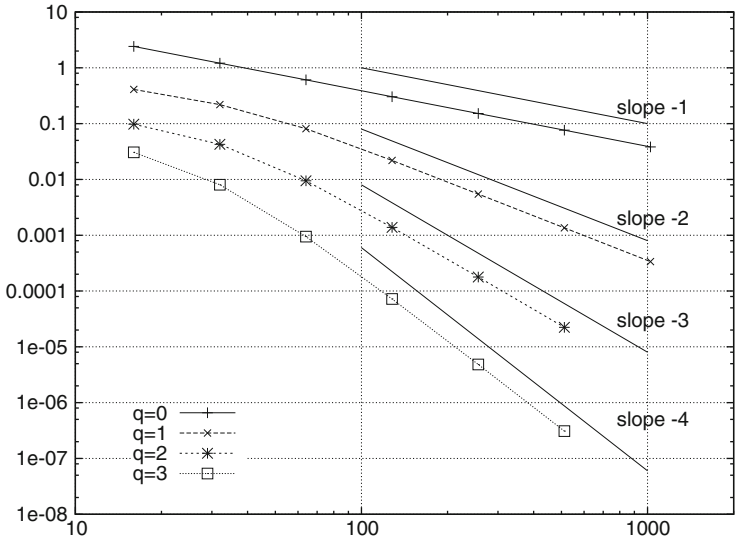


FIG. 6. The error $\max_{0\le n\le N}\|(u-U)(t_n)\|^2_{L^2(\Omega_{t_n})}$ versus the number of uniform time-steps N is depicted for $q=0,1,2,3$. The reference domain is $\Omega_0:=(-1,1)\times(-1,1)$ and it is deformed according to the ALE map $\mathcal{A}_t(\mathbf{y}):=\mathbf{y}(1+\frac{1}{2}\,T_{11}(t))$ for $t\in(0,0.99)$, where $T_n(t):=\cos(n\arccos(t))$ is the nth Chebychev polynomial of first kind. We set $\mu=1.0$ and the exact solution is manufactured to be $u(\mathbf{x},t):=\exp(x_1\ t)\sin(x_2\ t)$ with $\mathbf{x}:=(x_1,x_2)\in\Omega_t$. We take $\hat{\mathbf{W}}$ to be the L^2-projection of the ALE velocity $\hat{\mathbf{w}}$ onto $\hat{\mathcal{V}}_q$ and compute $\tilde{\mathcal{A}}_t$ according to (3.12). Over each interval I_n, a quadrature based on $q+1$ Radau points is used. The discretization in space is chosen not to influence the time discretization error. As predicted by estimate (4.5), all schemes exhibit the optimal $\mathcal{O}(N^{-(q+1)})$ order of convergence.

Figure 6 displays an optimal rate of convergence $q+1$ for RKR methods with $0\le q\le 3$ and the same experiment as in Fig. 4, for Reynolds' quadrature.

4.2. Implicit-Explicit Runge–Kutta (IERK) Method. Finally, we mention that similar stability and a priori error estimates can be established for the IERK method of first order:

$$\begin{aligned}&\langle(U_{n+1}-U_n)+k_n\big((\mathbf{b}-\mathbf{w})(t_n)\cdot\nabla_{\mathbf{x}}U_{n+1}\big),V\rangle_{\Omega_{t_n}}\\&\quad+\mu k_n\langle\nabla_{\mathbf{x}}U_{n+1},\nabla_{\mathbf{x}}V\rangle_{\Omega_{t_n}}=k_n\langle f(t_n),V\rangle_{\Omega_{t_n}},\quad\forall\,V\in\mathcal{V}_0(I_n),\end{aligned}\tag{4.6}$$

where $U_{n+1}:=U(\mathcal{A}_{t_n\to t_{n+1}}(\cdot),t_{n+1})$ for $U\in\mathcal{V}_0(I_n)$. Method (4.6) can be obtained by approximation of the integrals in (2.3) with the left-side rectangle quadrature. Despite the fact that (4.6) is not an RKR method, the error analysis for RKR methods in the ALE framework is applicable to (4.6) as well [10]. This method is natural for free-boundary problems, [6, 11, 12], because it is implicit with respect to the approximation U, but explicit with respect to the moving domain. The latter is beneficial

whenever we do not know in advance $\Omega_{t_{n+1}}$ at step n, while the implicit nature of the method in U helps avoiding any CFL condition. Rigorous error analysis for the IERK method (4.6) appears for the first time in [10].

We conclude with an application of IERK to *fluid–membrane interaction* due to Bonito et al. [12]. The deformable domain Ω_t contains an incompressible fluid governed by the Navier–Stokes equation

$$\begin{aligned} \rho D_t \mathbf{v} - \operatorname{div} \Sigma(\mathbf{v}, p) = 0 \qquad & \text{in } \Omega_t, \\ \operatorname{div} \mathbf{v} = 0 \qquad & \text{in } \Omega_t, \end{aligned} \tag{4.7}$$

where $\mathbf{v}$ is the fluid velocity, p is its pressure, $\Sigma(\mathbf{v}, p) = -pI + \mu D(\mathbf{v})$ is the Cauchy stress tensor, $D(\mathbf{v}) = \frac{1}{2}(\nabla \mathbf{v} + \nabla \mathbf{v}^T)$ is the symmetric part of the gradient, and μ is the viscosity. The membrane $\partial\Omega_t$ is governed by the *Canham–Helfrich energy*

$$J(\partial\Omega_t) = \frac{1}{2} \int_{\partial\Omega_t} H^2 + \lambda \left(\int_{\partial\Omega_t} 1 - \int_{\partial\Omega_0} 1 \right), \tag{4.8}$$

where H stands for the *mean curvature* of $\partial\Omega_t$ and λ is the Lagrange multiplier enforcing area conservation. The fluid and membrane interact through the boundary condition, which represents a balance of forces at the interface $\partial\Omega_t$:

$$\Sigma \boldsymbol{\nu} = \kappa \, \delta J(\partial\Omega_t), \tag{4.9}$$

where κ is the membrane bending rigidity coefficient and $\delta J(\partial\Omega_t)$ is the variational (or shape) derivative. The latter obeys the expression

$$\delta J(\partial\Omega_t) = \left(\Delta_{\partial\Omega_t} H + \frac{1}{2} H^3 - 2KH + \lambda H \boldsymbol{\nu} \right) \boldsymbol{\nu}, \tag{4.10}$$

where $\Delta_{\partial\Omega_t}$ is the Laplace–Beltrami operator on $\partial\Omega_t$ and K is the Gaussian curvature of $\partial\Omega_t$. It is important to notice that $\delta J(\partial\Omega_t)$ is a vector field perpendicular to $\partial\Omega_t$ because $\boldsymbol{\nu}$ is the unit normal to $\partial\Omega_t$. The discretization of (4.7)–(4.10) consists of Taylor–Hood finite elements coupled with IERK. Figure 7 displays the complex behavior of the fluid membrane and quite noticeable inertial effects.

Acknowledgements. The work of the Andrea Bonito author was supported in part by NSF Grant DMS-0914977. The work of the Irene Kyza author was supported in part by the European Social Fund (ESF)-European Union (EU) and National Resources of the Greek State within the framework of the Action "Supporting Postdoctoral Researchers" of the Operational Programme "Education and Lifelong Learning (EdLL)". The work of the Ricardo H. Nochetto author was supported in part by NSF Grants DMS-0807811 and DMS-1109325.

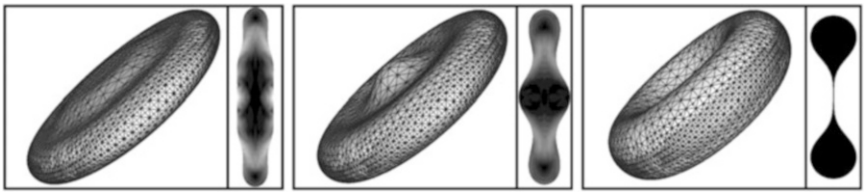


FIG. 7. Evolution of a fluid membrane with initial axisymmetric ellipsoidal shape of aspect ratio $5 \times 5 \times 1$ and final shape similar to a red blood cell. Each frame shows the membrane mesh and a symmetry cut along a big axis. The fluid flow is quite complex, creating first a bump in the middle and next moving towards the circumference and producing a depression in the center with flat pinching profile. The inertial effects are due to unrealistic physical parameters.

REFERENCES

[1] G. Akrivis and Ch. Makridakis. Galerkin time-stepping methods for nonlinear parabolic equations. *M2AN Math. Model. Numer. Anal.*, 38(2):261–289, 2004.

[2] G. Akrivis, Ch. Makridakis, and R.H. Nochetto. A posteriori error estimates for the Crank-Nicolson method for parabolic equations. *Math. Comp.*, 75(254):511–531, 2006.

[3] G. Akrivis, Ch. Makridakis, and R.H. Nochetto. Optimal order a posteriori error estimates for a class of Runge-Kutta and Galerkin methods. *Numer. Math.*, 114(1):133–160, 2009.

[4] G. Akrivis, Ch. Makridakis, and R.H. Nochetto. Galerkin and Runge-Kutta methods: unified formulation, a posteriori error estimates and nodal superconvergence. *Numer. Math.*, 118(3):429–456, 2011.

[5] S. Badia and R. Codina. Analysis of a stabilized finite element approximation of the transient convection-diffusion equation using an ALE framework. *SIAM J. Numer. Anal.*, 44(5):2159–2197 (electronic), 2006.

[6] E. Bänsch. Finite element discretization of the Navier-Stokes equations with a free capillary surface. *Numer. Math.*, 88(2):203–235, 2001.

[7] D. Boffi and L. Gastaldi. Stability and geometric conservation laws for ALE formulations. *Comput. Methods Appl. Mech. Engrg.*, 193(42–44):4717–4739, 2004.

[8] A. Bonito, I. Kyza, and R.H. Nochetto. Time-discrete higher order ALE formulations: A posteriori error analysis. In preparation.

[9] A. Bonito, I. Kyza, and R.H. Nochetto. Time-discrete higher order ALE formulations: A priori error analysis. *Numer. Math.*, 125(2):225–257, 2013.

[10] A. Bonito, I. Kyza, and R.H. Nochetto. Time-discrete higher order ALE formulations: Stability. *SIAM J. Numer. Anal.*, 51(1):577–604, 2013.

[11] A. Bonito, R.H. Nochetto, and M.S. Pauletti. Parametric FEM for geometric biomembranes. *J. Comput. Phys.*, 229(9):3171–3188, 2010.

[12] A. Bonito, R.H. Nochetto, and M.S. Pauletti. Dynamics of biomembranes: Effect of the bulk fluid. *Math. Model. Nat. Phenom.*, 6(5):25–43, 2011.

[13] S.C. Brenner and L.R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.

[14] S. Brogniez, A. Rajasekharan, and Ch. Farhat. Provably stable and time-accurate extensions of Runge-Kutta schemes for CFD computations on moving grids. *Internat. J. Numer. Methods Fluids*, 69(7):1249–1270, 2012.

[15] K. Chrysafinos and N.J. Walkington. Lagrangian and moving mesh methods for the convection diffusion equation. *M2AN Math. Model. Numer. Anal.*, 42(1):25–55, 2008.

[16] J. Donéa, S. Giuliani, and J.P. Halleux. An arbitrary Lagrangian-Eulerian finite element method for transient dynamic fluid-structure interactions. *Comput. Methods Appl. Mech. Engrg.*, 33(1–3):689–723, 1982.

[17] Ch. Farhat, Ph. Geuzaine, and C. Grandmont. The discrete geometric conservation law and the nonlinear stability of ALE schemes for the solution of flow problems on moving grids. *J. Comput. Phys.*, 174(2):669–694, 2001.

[18] Ch. Farhat and Ph. Geuzaine. Design and analysis of robust ALE time-integrators for the solution of unsteady flow problems on moving grids. *Internat. J. Numer. Methods Fluids*, 193(39–41):4073–4095, 2004.

[19] Ch. Farhat, M. Lesoinne, and N. Maman. Mixed explicit/implicit time integration of coupled aeroelastic problems: three-field formulation, geometric conservation and distributed solution. *Internat. J. Numer. Methods Fluids*, 21(10): 807–835, 1995. Finite element methods in large-scale computational fluid dynamics (Tokyo, 1994).

[20] L. Formaggia and F. Nobile. A stability analysis for the arbitrary Lagrangian Eulerian formulation with finite elements. *East-West J. Numer. Math.*, 7(2): 105–131, 1999.

[21] L. Formaggia and F. Nobile. Stability analysis of second-order time accurate schemes for ALE-FEM. *Comput. Methods Appl. Mech. Engrg.*, 193(39–41): 4097–4116, 2004.

[22] L. Gastaldi. A priori error estimates for the arbitrary Lagrangian Eulerian formulation with finite elements. *East-West J. Numer. Math.*, 9(2):123–156, 2001.

[23] Ph. Geuzaine, C. Grandmont, and Ch. Farhat. Design and analysis of ale schemes with provable second-order time-accuracy for inviscid and viscous flow simulations. *Journal of Computational Physics*, 191(1):206–227, 2003.

[24] H. Guillard and Ch. Farhat. On the significance of the geometric conservation law for flow computations on moving meshes. *Comput. Methods Appl. Mech. Engrg.*, 190(11–12):1467–1482, 2000.

[25] P. Hansbo, J. Hermansson, and T. Svedberg. Nitsche's method combined with space-time finite elements for ALE fluid-structure interaction problems. *Comput. Methods Appl. Mech. Engrg.*, 193(39–41):4195–4206, 2004.

[26] C.W. Hirt, A.A. Amsden, and J.L. Cook. An arbitrary Lagrangian-Eulerian computing method for all flow speeds [J. Comput. Phys. **14** (1974), no. 3, 227–253]. *J. Comput. Phys.*, 135(2):198–216, 1997. With an introduction by L. G. Margolin, Commemoration of the 30th anniversary {of J. Comput. Phys.}.

[27] T.J.R. Hughes, W.K. Liu, and T.K. Zimmermann. Lagrangian-Eulerian finite element formulation for incompressible viscous flows. *Comput. Methods Appl. Mech. Engrg.*, 29(3):329–349, 1981.

[28] P. Jamet. Galerkin-type approximations which are discontinuous in time for parabolic equations in a variable domain. *SIAM J. Numer. Anal.*, 15(5): 912–928, 1978.

[29] J. A. Mackenzie and W. R. Mekwi. An unconditionally stable second-order accurate ALE-FEM scheme for two-dimensional convection-diffusion problems. *IMA J. Numer. Anal.*, 32(3):888–905, 2012.

[30] Ch. Makridakis and R.H. Nochetto. A posteriori error analysis for higher order dissipative methods for evolution problems. *Numer. Math.*, 104(4):489–514, 2006.

[31] F. Nobile. *Numerical approximation on fluid-structure interaction problems with application to haemodynamics*. PhD thesis, École Polytechnique Fédérale de Lausanne, 2001.

[32] L.A. Ortega and G. Scovazzi. A geometrically-conservative, synchronized, flux-corrected remap for arbitrary Lagrangian-Eulerian computations with nodal finite elements. *J. Comput. Phys.*, 230(17):6709–6741, 2011.

[33] O. Pironneau, J. Liou, and T. Tezduyar. Characteristic-Galerkin and Galerkin/least-squares space-time formulations for the advection-diffusion equations with time-dependent domains. *Comput. Methods Appl. Mech. Engrg.*, 100(1):117–141, 1992.

[34] A. Quarteroni and L. Formaggia. *Mathematical modelling and numerical simulation of the cardiovascular system*, volume XII of *Handb. Numer. Anal.* North-Holland, Amsterdam, 2004.

[35] A. Quarteroni, M. Tuveri, and A. Veneziani. Computational vascular fluid dynamics: problems, models and methods. *Comput. Visual. Sci.*, 2(4):163–197, 2000.

[36] P.D. Thomas and C.K. Lombard. The geometric conservation law–a link between finite-difference and finite–volume methods for flow computation on moving grids. *AIAA Paper 78–1208*, 1978.

[37] V. Thomée. *Galerkin finite element methods for parabolic problems*, volume 25 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 2006.

DISCONTINUOUS FINITE ELEMENT METHODS FOR COUPLED SURFACE–SUBSURFACE FLOW AND TRANSPORT PROBLEMS

BEATRICE RIVIERE*

Abstract. A numerical method is proposed to solve the coupled flow and transport problems in adjacent surface and subsurface regions. The flow problem is characterized by the Navier–Stokes (or Stokes) equations coupled by Darcy equations. In the subsurface, the diffusion coefficient of the transport equation depends on the velocity field in a nonlinear manner. The interior penalty discontinuous Galerkin method is used for the spatial discretization, and the backward Euler technique for the time integration. Convergence of the scheme is theoretically derived. Numerical examples show the robustness of the method for heterogeneous and fractured porous media.

Key words. Navier–Stokes, Darcy, Transport, Discontinuous Galerkin, Heterogeneous media, Convergence, Multinumerics

AMS(MOS) subject classifications. 65M60, 76S05, 76D05

1. Introduction.

The study of a coupled flow and transport system in adjacent surface and subsurface regions is of interest for the environmental problem of contaminated aquifers through rivers. The flow in the surface region is characterized by the steady-state Navier–Stokes (or Stokes) equations whereas the flow in the subsurface region is characterized by Darcy's law. The concentration of the contaminant satisfies a transport equation coupled to the flow problem in the following sense: the flow velocity appears in both the diffusion and the convection terms of the concentration equation. This type of multiphysics couplings is also of importance in the industrial filtration processes [20].

This paper follows a series of papers on the coupled surface/subsurface flows by the author. In [6–8, 10, 11, 18], the flow problem coupling Navier–Stokes equations with Darcy equations was analyzed theoretically for different interface conditions and discretized by finite element methods and discontinuous Galerkin methods. The usual interface conditions include the Beavers–Joseph–Saffman law [4, 26], the continuity of normal component of velocity, and the balance of forces across the interface.

A weak formulation of the coupling of surface/subsurface flow with transport was analyzed in [9]. The main objective of this paper is to propose a robust numerical scheme for approximating the weak solution. We assume a one-way coupling, i.e. the velocity field obtained from solving the surface/subsurface flow problem becomes input data for the transport problem. The multiphysics problem is approximated by the discon-

*Department of Computational and Applied Mathematics, Rice University, 6100 Main St MS-134, Houston, TX 77005, USA, riviere@rice.edu

X. Feng et al. (eds.), *Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations*, The IMA Volumes in Mathematics and its Applications 157, DOI 10.1007/978-3-319-01818-8_11,

tinuous Galerkin (DG) method. Because of its flexibility and local mass conservation property, the DG method is a well-suited method for modeling flow in heterogeneous porous media. The transport problem is solved by an improved discontinuous Galerkin method that upwinds the numerical fluxes in the subsurface region [24]. In this case, one does not need to use slope limiters.

The coupled surface/subsurface flow problem has recently gained a lot of interest in the scientific community. Most of the published literature covers the coupling of Stokes and Darcy equations (see, for instance, [14–16, 21, 23]). The published literature is very sparse on the coupling of Stokes–Darcy-transport problem. In [28], a mixed method is proposed for the coupled Stokes/Darcy equations and a local discontinuous Galerkin method [13] is used for the transport problem.

The outline of the paper is as follows. The next section introduces the model problem and its weak formulation. In Sect. 3, the numerical scheme is defined and error estimates are obtained. Numerical examples are shown in Sect. 4. Conclusions follow.

2. Model Problem. For simplicity we assume that the surface region is contained in a domain $\Omega_1 \subset \mathbb{R}^2$ and the subsurface region in a domain $\Omega_2 \subset \mathbb{R}^2$. The results in Sect. 3 are valid for three-dimensional domains as well. Let $\boldsymbol{u}_i$ and p_i denote the fluid velocity and pressure in Ω_i, for $i = 1, 2$. Let $\boldsymbol{\tau}_{12}$ and $\boldsymbol{n}_{12}$ be a unit tangential vector and a unit normal vector at the interface $\Gamma_{12} = \partial\Omega_1 \cap \partial\Omega_2$. The vector $\boldsymbol{n}_{12}$ is assumed to be outward of Ω_1. The surface/subsurface flow is characterized by the following Navier–Stokes equations coupled with the Darcy equations, and appropriate interface conditions.

$$-\nabla \cdot (2\mu \boldsymbol{D}(\boldsymbol{u}_1)) + \nabla p_1 + \boldsymbol{u}_1 \cdot \nabla \boldsymbol{u}_1 = \boldsymbol{f}_1, \qquad \nabla \cdot \boldsymbol{u}_1 = 0, \text{ in } \Omega_1, \tag{2.1}$$

$$\boldsymbol{u}_2 = -\frac{\boldsymbol{K}}{\mu}(\nabla p_2 - \rho \boldsymbol{g}), \qquad \nabla \cdot \boldsymbol{u}_2 = f_2, \text{ in } \Omega_2, \tag{2.2}$$

$$\boldsymbol{u}_1 \cdot \boldsymbol{n}_{12} = \boldsymbol{u}_2 \cdot \boldsymbol{n}_{12}, \text{ on } \Gamma_{12}, \tag{2.3}$$

$$G\boldsymbol{K}^{-1/2}\boldsymbol{u}_1 \cdot \boldsymbol{\tau}_{12} = -2\mu \boldsymbol{D}(\boldsymbol{u}_1)\boldsymbol{n}_{12} \cdot \boldsymbol{\tau}_{12}, \text{ on } \Gamma_{12}, \tag{2.4}$$

$$(-2\mu \boldsymbol{D}(\boldsymbol{u}_1)\boldsymbol{n}_{12}) \cdot \boldsymbol{n}_{12} + p_1 = p_2, \text{ on } \Gamma_{12}. \tag{2.5}$$

Let $\boldsymbol{u}$ denote the velocity field over the whole domain, namely $\boldsymbol{u}|_{\Omega_i} = \boldsymbol{u}_i$. The concentration c of one species transported in the domain $\Omega = \Omega_1 \cup \Omega_2$ over the time interval $(0, T)$ satisfies the following equation

$$\frac{\partial}{\partial t}(\varphi c) - \nabla \cdot (\boldsymbol{F}(\boldsymbol{u})\nabla c - c\boldsymbol{u}) = f, \quad \text{in } (0, T) \times \Omega. \tag{2.6}$$

We remark that if the nonlinear term $\boldsymbol{u}_1 \cdot \nabla \boldsymbol{u}_1$ is removed from the momentum equation in (2.1), the resulting problem is a coupled Stokes–Darcy flow with transport, and (2.1) is replaced by:

$$-\nabla \cdot (2\mu \boldsymbol{D}(\boldsymbol{u}_1)) + \nabla p_1 = \boldsymbol{f}_1, \qquad \nabla \cdot \boldsymbol{u}_1 = 0, \quad \text{in } \Omega_1.$$

Throughout the paper, we will point out the simplifications obtained if the Stokes equations are used instead of the Navier–Stokes equations in the free flow region. Define $\Gamma_i = \partial\Omega_i \backslash \Gamma_{12}$ and denote by $\boldsymbol{n}$ the unit outward normal to $\partial\Omega$. The system of equations is completed by boundary conditions and an initial condition for the concentration.

$$\boldsymbol{u}_1 = \boldsymbol{0}, \text{ on } \Gamma_1, \boldsymbol{u}_2 \cdot \boldsymbol{n} = \mathcal{U}, \text{ on } \Gamma_2, \tag{2.7}$$

$$\boldsymbol{F}(\boldsymbol{u})\nabla c \cdot \boldsymbol{n} - c\boldsymbol{u} \cdot \boldsymbol{n} = -\mathcal{C}\boldsymbol{u} \cdot \boldsymbol{n}, \text{ on } (0,T) \times \{x \in \partial\Omega : \mathcal{U}(x) < 0\}, \tag{2.8}$$

$$\boldsymbol{F}(\boldsymbol{u})\nabla c \cdot \boldsymbol{n} = 0, \text{ on } (0,T) \times \{x \in \partial\Omega : \mathcal{U}(x) \geq 0\}, \tag{2.9}$$

$$c = c_0, \text{ in } \{0\} \times \Omega. \tag{2.10}$$

We now describe the coefficients that appear in the equations above.

- The fluid kinematic viscosity μ and fluid density ρ are positive constants. The vector of gravitational acceleration is denoted by $\boldsymbol{g}$.
- The rate of strain matrix is symmetric and defined by $\boldsymbol{D}(\boldsymbol{u}) = 0.5(\nabla\boldsymbol{u} + (\nabla\boldsymbol{u})^T)$.
- The vector function $\boldsymbol{f}_1$ and scalar functions f_2 and f represent the source/sink terms.
- The permeability matrix $\boldsymbol{K}$ is symmetric positive definite and bounded above and below: there exist $\underline{k} > 0, \bar{k} > 0$ such that

$$\forall \xi \in \mathbb{R}^2, \quad \underline{k}\xi \cdot \xi \leq \xi \cdot \boldsymbol{K}\xi \leq \bar{k}\xi \cdot \xi.$$

- The coefficient G that appears in the Beavers–Joseph–Saffman interface condition (2.4) is a positive constant. It is obtained experimentally and depends on the properties of the fluid and the porous medium.
- The coefficient φ is a positive constant bounded by one. Restricted to Ω_2, the value of φ corresponds to the porosity of the subsurface. By convention, the coefficient φ is simply equal to one in Ω_1.
- The coefficient $\boldsymbol{F}(\boldsymbol{u})$ is a diffusion/dispersion matrix. In Ω_1, it is simply equal to $d_m\boldsymbol{I}$, where d_m is a positive constant, and $\boldsymbol{I}$ is the identity matrix. In the porous region Ω_2, the matrix $\boldsymbol{F}(\boldsymbol{u})$ depends on the velocity in the following manner:

$$\boldsymbol{F}(\boldsymbol{u}) = (\alpha_T\|\boldsymbol{u}\| + d_m)\boldsymbol{I} + (\alpha_l - \alpha_t)\frac{\boldsymbol{u}\boldsymbol{u}^T}{\|\boldsymbol{u}\|}.$$

  The coefficient $d_m > 0$ is the molecular diffusivity constant, $\alpha_l \geq 0$ and $\alpha_t \geq 0$ are the longitudinal and transverse dispersivities and $\|\cdot\|$ denotes the Euclidean norm. One can show (see, for instance, [27]) that $\boldsymbol{F}$ is Lipschitz and that there exist $\alpha > 0, M > 0$ such that

$$\boldsymbol{F}(\boldsymbol{w})\psi \cdot \psi \geq \alpha\psi \cdot \psi, \quad \|\boldsymbol{F}(\boldsymbol{w})\| \leq M(1 + \|\boldsymbol{w}\|). \tag{2.11}$$

In addition, we assume that there is $\bar{F} > 0$ such that

$$\|\boldsymbol{F}(\boldsymbol{w})\| \leq \bar{F}. \tag{2.12}$$

- The boundary flux $\mathcal{U}$ belongs to $L^2(\Gamma_2)$. The data f_2 and $\mathcal{U}$ must satisfy the compatibility condition

$$\int_{\Gamma_2} \mathcal{U} = \int_{\Omega_2} f_2.$$

- The function $\mathcal{C} \geq 0$ is the prescribed concentration on the inflow boundary. It is assumed to be bounded. For any function z, we denote $z^+ = \max(0, z)$ and $z^- = \max(0, -z)$. Extending $\mathcal{U}$ to zero on Γ_1, we can rewrite the boundary conditions (2.8) and (2.9) as

$$\boldsymbol{F}(\boldsymbol{u})\nabla c \cdot \boldsymbol{n} + c(\boldsymbol{u} \cdot \boldsymbol{n})^- = \mathcal{C}\mathcal{U}^- \text{ on } (0,T) \times \partial\Omega. \tag{2.13}$$

- The initial concentration c_0 is nonnegative and bounded.

We will solve for the unknowns $(\boldsymbol{u}_1, p_1, p_2, c)$. We note that the Darcy velocity $\boldsymbol{u}_2$ can be obtained from the Darcy pressure p_2 via the first equation in (2.2). For any domain $\mathcal{O}$, the standard notation for $L^k(\mathcal{O})$ spaces and Sobolev spaces $H^k(\mathcal{O})$ is used. The L^2 inner-product of two functions is denoted by $(\cdot,\cdot)_{\mathcal{O}}$. Let $H^1_{0,\Gamma_1}(\Omega_1)$ denote the space of functions in $H^1(\Omega_1)$ whose trace vanishes on Γ_1. The dual space of $H^1(\Omega)$ is denoted by $H^1(\Omega)'$ and the duality pairing is $\langle\cdot,\cdot\rangle_{(H^1(\Omega)',H^1(\Omega))}$.

A weak solution to the problem (2.1)–(2.7) with (2.10) and (2.13) is the quadruple $(\boldsymbol{u}_1, p_1, p_2, c)$ that belongs to $H^1_{0,\Gamma_1}(\Omega_1)^2 \times L^2(\Omega_1) \times H^1(\Omega_2) \times (L^2(0,T;H^1(\Omega)) \cap L^\infty((0,T)\times\Omega))$ satisfying for all $\boldsymbol{v}_1 \in H^1_{0,\Gamma_1}(\Omega_1)^2$, $q_1 \in L^2(\Omega_1)$, $q_2 \in H^1(\Omega_2)$:

$$\begin{aligned} 2\mu(\boldsymbol{D}(\boldsymbol{u}_1), \boldsymbol{D}(\boldsymbol{v}_1))_{\Omega_1} + (\boldsymbol{u}_1 \cdot \nabla \boldsymbol{u}_1, \boldsymbol{v}_1)_{\Omega_1} + \left(\frac{\boldsymbol{K}}{\mu}\nabla p_2, \nabla q_2\right)_{\Omega_2} - (\nabla \cdot \boldsymbol{v}_1, p_1)_{\Omega_1} \\ + (p_2, \boldsymbol{v}_1 \cdot \boldsymbol{n}_{12})_{\Gamma_{12}} + G(\boldsymbol{K}^{-1/2}\boldsymbol{u}_1 \cdot \boldsymbol{\tau}_{12}, \boldsymbol{v}_1 \cdot \boldsymbol{\tau}_{12})_{\Gamma_{12}} - (\boldsymbol{u}_1 \cdot \boldsymbol{n}_{12}, q_2)_{\Gamma_{12}} \\ + (\nabla \cdot \boldsymbol{u}_1, q_1)_{\Omega_1} = (\boldsymbol{f}_1, \boldsymbol{v}_1)_{\Omega_1} + \left(f_2 + \frac{\boldsymbol{K}}{\mu}\rho\boldsymbol{g}, q_2\right)_{\Omega_2} - (\mathcal{U}, q_2)_{\Gamma_2}, \end{aligned} \tag{2.14}$$

and for all $z \in L^2(0,T;H^1(\Omega))$

$$\begin{aligned} \int_0^T \langle\varphi\frac{\partial c}{\partial t}, z\rangle_{(H^1(\Omega)',H^1(\Omega))}dt - \int_0^T (c\boldsymbol{u}, \nabla z)_\Omega dt + \int_0^T (\boldsymbol{F}(\boldsymbol{u})\nabla c, \nabla z)_\Omega dt \\ + \int_0^T ((d\mathcal{U}^+ - \mathcal{C}\mathcal{U}^-), z)_{\partial\Omega} dt = \int_0^T (f, z)_\Omega dt. \end{aligned} \tag{2.15}$$

with $t \to c(t,\cdot) \in \mathcal{C}^0([0,T];(H^1(\Omega))')$, $t \to \frac{\partial c}{\partial t}(t,\cdot) \in L^2(0,T;(H^1(\Omega))')$ and

$$c(0,\cdot) = c_0(\cdot) \text{ a.e. in } \Omega. \tag{2.16}$$

Because only Neumann boundary conditions hold for the flow problem, the additional constraint $\int_{\Omega_2} p_2 = 0$ is imposed. Existence of a weak solution can be derived by combining results from [2, 18, 22]. We state the result below.

THEOREM 2.1. *Assume that* $\boldsymbol{f}_1 \in L^2(\Omega_1)^2$, $f_2 \geq 0$, $f_2 \in L^2(\Omega_2)$ *and* $f \geq 0$, $f \in L^1(0,T;L^\infty(\Omega)) \cap L^2(0,T;L^2(\Omega))$. *There exists a constant* $\tilde{M} > 0$ *such that if*

$$\mu^2 > \tilde{M}(\|\boldsymbol{f}_1\|^2_{L^2(\Omega_1)} + \mu\|f_2\|^2_{L^2(\Omega_2)} + \mu\|\mathcal{U}\|^2_{L^2(\Gamma_2)} + \|\boldsymbol{g}\|^2_{L^2(\Omega_2)}) \tag{2.17}$$

then there exists a weak solution $(\boldsymbol{u}_1, p_1, p_2, c)$ *to the weak problem (2.14)–(2.16).*

REMARK 2.1. *If the Stokes equations are used, existence of the weak solution is unconditional, i.e. there is no need to assume a small data condition like (2.17). The same result holds true if the Navier–Stokes equations are used and the interface condition (2.5) is replaced by*

$$(-2\mu\boldsymbol{D}(\boldsymbol{u}_1)\boldsymbol{n}_{12}) \cdot \boldsymbol{n}_{12} + p_1 + \frac{1}{2}\boldsymbol{u}_1 \cdot \boldsymbol{u}_1 = p_2.$$

In this case, the coupled flow model is numerically discussed in [10]. We also note that existence of a weak solution of a more general coupled flow and transport problem is shown in [9].

In the next section, we define a numerical approximation of the weak problem.

3. Numerical Discretization. Let $\mathcal{E}^h$ be a regular family of triangulations of $\overline{\Omega}$ (see [12]) and let h denote the maximum diameter of the triangles. We assume that the interface Γ_{12} is a finite union of triangle edges. Therefore, the restriction of $\mathcal{E}^h$ to Ω_i is also a regular family of triangulations of $\overline{\Omega}_i$; we denote it by $\mathcal{E}_i^h$ and impose that the two meshes $\mathcal{E}_i^h$ coincide at the interface Γ_{12}. This restriction simplifies the analysis, but it can be relaxed.

For $i = 1, 2$, let Γ_i^h denote the set of edges of $\mathcal{E}_i^h$ interior to Ω_i and let $\Gamma^h = \Gamma_1^h \cup \Gamma_2^h$. To each edge e of $\mathcal{E}^h$ we associate once and for all a unit normal vector $\boldsymbol{n}_e$. For the edges in Γ_i^h, this can be done by ordering the triangles of $\mathcal{E}_i^h$ and orienting the normal in the direction of increasing numbers. For $e \in \Gamma_{12}$, we set $\boldsymbol{n}_e = \boldsymbol{n}_{12}$, i.e. $\boldsymbol{n}_e$ is the exterior normal to Ω_1. For a boundary edge $e \in \Gamma_i$, $\boldsymbol{n}_e$ coincides with the outward normal vector $\boldsymbol{n}$ to $\partial\Omega$. If $\boldsymbol{n}_e$ points from the element E^1 to the element E^2, the jump $[\cdot]$ and average $\{\cdot\}$ of a function ϕ are given by:

$$[\phi] = \phi|_{E^1} - \phi|_{E^2}, \quad \{\phi\} = \frac{1}{2}\phi|_{E^1} + \frac{1}{2}\phi|_{E^2}.$$

By convention, for a boundary edge on Γ_i, the jump and average are defined to be equal to the trace of the function on that edge. The length of an edge e is denoted by $|e|$.

3.1. Numerical Approximation of Flow Problem. Let $\boldsymbol{X}_1^h, Q_1^h$, Q_2^h be finite dimensional subspaces to be defined later. Formally, the discrete weak formulation of (2.1)–(2.5) can be written as: find $\boldsymbol{U}_1^h \in \boldsymbol{X}_1^h, P_1^h \in Q_1^h, P_2^h \in Q_2^h$ such that

$$\begin{gathered}\forall \boldsymbol{v}_1^h \in \boldsymbol{X}_1^h, \forall q_2^h \in Q_2^h, \quad a_{\mathrm{NS}}(\boldsymbol{U}_1^h, \boldsymbol{v}_1^h) + b_{\mathrm{NS}}(\boldsymbol{v}_1^h, P_1^h) + c_{\mathrm{NS}}(\boldsymbol{U}_1^h; \boldsymbol{U}_1^h, \boldsymbol{v}_1^h) \\ + a_{\mathrm{D}}(P_2^h, q_2^h) + \gamma(\boldsymbol{U}_1^h, P_2^h; \boldsymbol{v}_1^h, q_2^h) = \ell(\boldsymbol{v}_1^h, q_2^h), \\ \forall q_1^h \in Q_1^h, \quad b_{\mathrm{NS}}(\boldsymbol{U}_1^h, q_1^h) = 0, \\ \int_{\Omega_2} P_2^h = 0,\end{gathered}$$

where $a_{\mathrm{NS}}, b_{\mathrm{NS}}, c_{\mathrm{NS}}, a_{\mathrm{D}}$ are discretizations of the operators $-\nabla \cdot (2\mu \boldsymbol{D}(\boldsymbol{u}))$, ∇p, $\boldsymbol{u} \cdot \nabla \boldsymbol{u}$, and $-\nabla \cdot ((\boldsymbol{K}/\mu)\nabla p)$, respectively. These forms depend on the choice of the numerical method. Since the discrete problem is steady-state and nonlinear, Picard iterations are computed with an initial zero Navier–Stokes velocity.

Denote by $\boldsymbol{U}^h$ the resulting velocity field of the coupled Navier–Stokes and Darcy equations. The velocity $\boldsymbol{U}^h$ is defined in Ω by:

$$\boldsymbol{U}^h = \begin{cases} \boldsymbol{U}_1^h, & \text{in } \Omega_1 \\ -\frac{\boldsymbol{K}}{\mu}(\nabla P_2^h - \rho \boldsymbol{g}), & \text{in } \Omega_2. \end{cases} \tag{3.1}$$

The form γ couples the two different physical flows through the interface Γ_{12}.

$$\begin{aligned}\gamma(\boldsymbol{U}_1^h, P_2^h; \boldsymbol{v}_1^h, q_2^h) = (P_2^h, \boldsymbol{v}_1^h \cdot \boldsymbol{n}_{12})_{\Gamma_{12}} + G(\boldsymbol{K}^{-1/2}\boldsymbol{U}_1^h \cdot \boldsymbol{\tau}_{12}, \boldsymbol{v}_1^h \cdot \boldsymbol{\tau}_{12})_{\Gamma_{12}} \\ -(\boldsymbol{U}_1^h \cdot \boldsymbol{n}_{12}, q_2^h)_{\Gamma_{12}}.\end{aligned} \tag{3.2}$$

The form ℓ is defined as:

$$\ell(\boldsymbol{v}_1^h, q_2^h) = (\boldsymbol{f}_1, \boldsymbol{v}_1^h)_{\Omega_1} + \left(f_2 + \frac{\boldsymbol{K}}{\mu}\rho \boldsymbol{g}, q_2^h\right)_{\Omega_2} + (\mathcal{U}, q_2^h)_{\Gamma_2}.$$

We remark that if the Stokes equations are used instead of the Navier–Stokes equations, the numerical scheme remains the same with the choice $c_{\mathrm{NS}} = 0$. One can use various discretizations in either subdomains. We choose to present the discontinuous Galerkin (DG) method. In what follows we give a description of the linear forms $a_{\mathrm{NS}}, a_{\mathrm{D}}, b_{\mathrm{NS}}, c_{\mathrm{NS}}$.

3.1.1. DG Scheme. The primal DG method is applied to both the Navier–Stokes equations and the Darcy equations. The penalty parameter is denoted by $\sigma > 0$ and the symmetrizing parameter by ϵ. The parameter ϵ takes the values -1 or $+1$, which corresponds to either the symmetric interior penalty Galerkin (SIPG) method or the non-symmetric interior penalty Galerkin (NIPG) method [3, 25, 29]. We can allow for different values of σ for each edge, and for different values of ϵ for the forms a_{NS}

and a_{D}. To simplify the text, we assume that σ and ϵ are fixed constants for both forms. Let k_1, k_2 be positive integers, each greater than or equal to one. In that case the finite dimensional spaces are

$$\begin{aligned}
\boldsymbol{X}_1^h &= \{\boldsymbol{v}_h \in L^2(\Omega_1)^2 : \boldsymbol{v}_h|_E \in (\mathbb{P}_{k_1}(E))^2, \forall E \in \mathcal{E}_1^h\},\\
Q_1^h &= \{q_h \in L^2(\Omega_1) : q_h|_E \in \mathbb{P}_{k_1-1}(E), \forall E \in \mathcal{E}_1^h\},\\
Q_2^h &= \{q_h \in L^2(\Omega_2) : q_h|_E \in \mathbb{P}_{k_2}(E), \forall E \in \mathcal{E}_2^h\},
\end{aligned}$$

where $\mathbb{P}_k$ is the space of polynomials of total degree less than or equal to k and the forms are:

$$\begin{aligned}
a_{\mathrm{NS}}(\boldsymbol{w}_h, \boldsymbol{v}_h) = {} & 2\mu \sum_{E\in\mathcal{E}_1^h} (\boldsymbol{D}(\boldsymbol{w}_h), \boldsymbol{D}(\boldsymbol{v}_h))_E - 2\mu \sum_{e\in\Gamma_1^h\cup\Gamma_1} (\{\boldsymbol{D}(\boldsymbol{w}_h)\boldsymbol{n}_e\}, [\boldsymbol{v}_h])_e\\
& + 2\epsilon\mu \sum_{e\in\Gamma_1^h\cup\Gamma_1} (\{\boldsymbol{D}(\boldsymbol{v}_h)\boldsymbol{n}_e\}, [\boldsymbol{w}_h])_e + \mu \sum_{e\in\Gamma_1^h\cup\Gamma_1} \frac{\sigma}{|e|}([\boldsymbol{w}], [\boldsymbol{v}])_e,
\end{aligned} \tag{3.3}$$

$$b_{\mathrm{NS}}(\boldsymbol{v}_h, q_1^h) = -\sum_{E\in\mathcal{E}_1^h} (q_1^h, \nabla\cdot\boldsymbol{v}_h)_E + \sum_{e\in\Gamma_1^h\cup\Gamma_1} (\{q_1^h\}, [\boldsymbol{v}_h]\cdot\boldsymbol{n}_e)_e, \tag{3.4}$$

$$\begin{aligned}
a_{\mathrm{D}}(z_2^h, q_2^h) = {} & \sum_{E\in\mathcal{E}_2^h} \left(\frac{\boldsymbol{K}}{\mu}\nabla z_2^h, \nabla q_2^h\right)_E - \sum_{e\in\Gamma_2^h} \left(\left\{\frac{\boldsymbol{K}}{\mu}\nabla z_2^h\cdot\boldsymbol{n}_e\right\}, [q_2^h]\right)_e\\
& + \epsilon \sum_{e\in\Gamma_2^h} \left(\left\{\frac{\boldsymbol{K}}{\mu}\nabla q_2^h\cdot\boldsymbol{n}_e\right\}, [z_2^h]\right)_e + \sum_{e\in\Gamma_2^h} \frac{\sigma}{|e|}([z_2^h], [q_2^h])_e.
\end{aligned} \tag{3.5}$$

The DG discretization of the nonlinear term $\boldsymbol{u}\cdot\nabla\boldsymbol{u}$ has been studied extensively, for instance, in [19]. For this, we introduce additional notation. For an element $E \in \mathcal{E}^h$, let $\mathcal{N}(E)$ denote the neighboring element sharing part of ∂E. When the side of E belongs to $\partial\Omega$, then $\mathcal{N}(E)$ is not defined, and by convention we set $\boldsymbol{v}_h|_{\mathcal{N}(E)} = \boldsymbol{0}$ for any function $\boldsymbol{v}_h \in \boldsymbol{X}_1^h$. We also denote by $\boldsymbol{n}_E$ the unit outward normal to E. The inflow boundary of E with respect to a function $\boldsymbol{z}_h \in \boldsymbol{X}_1^h$ is defined by

$$\partial E_-(\boldsymbol{z}_h) = \{\boldsymbol{x} \in \partial E : \{\boldsymbol{z}_h(\boldsymbol{x})\}\cdot\boldsymbol{n}_E < 0\}.$$

We are now ready to define the form c_{NS}.

$$\begin{aligned}
c_{\mathrm{NS}}(\boldsymbol{z}_h; \boldsymbol{v}_h, \boldsymbol{w}_h) = {} & \sum_{E\in\mathcal{E}_1^h} (\boldsymbol{z}_h\cdot\nabla\boldsymbol{v}_h, \boldsymbol{w}_h)_E + \frac{1}{2}\sum_{E\in\mathcal{E}_1^h} (\nabla\cdot\boldsymbol{z}_h, \boldsymbol{v}_h\cdot\boldsymbol{w}_h)_E\\
& - \frac{1}{2}\sum_{e\in\Gamma_1^h\cup\Gamma_1} ([\boldsymbol{z}_h]\cdot\boldsymbol{n}_e, \{\boldsymbol{v}_h\cdot\boldsymbol{w}_h\})_e\\
& + \sum_{E\in\mathcal{E}_1^h} (\{\boldsymbol{z}_h\}\cdot\boldsymbol{n}_E(\boldsymbol{v}_h|_E - \boldsymbol{v}_h|_{\mathcal{N}(E)}), \boldsymbol{w}_h|_E)_{\partial E_-(\boldsymbol{z}_h)\setminus\Gamma_{12}}.
\end{aligned} \tag{3.6}$$

The norms associated with the discrete spaces are:

$$\|\boldsymbol{v}\|_{\boldsymbol{X}_1^h} = \left(\sum_{E\in\mathcal{E}_1^h} \|\boldsymbol{D}(\boldsymbol{v})\|^2_{L^2(E)} + \sum_{e\in\Gamma_1^h\cup\Gamma_1} |e|^{-1}\|[\boldsymbol{v}]\|^2_{L^2(e)} \right)^{1/2},$$

$$\|q\|_{Q_1^h} = \|q\|_{L^2(\Omega_1)},$$

$$\|q\|_{Q_2^h} = \left(\sum_{E\in\mathcal{E}_2^h} \left\| \frac{\boldsymbol{K}^{1/2}}{\mu^{1/2}}\nabla q \right\|^2_{L^2(E)} + \sum_{e\in\Gamma_2^h} |e|^{-1}\|[q]\|^2_{L^2(e)} \right)^{1/2}.$$

3.1.2. Error Analysis. The DG method was analyzed in [18] for different boundary conditions for the Darcy pressure. It is a simple technicality to redo the analysis for the case of Neumann boundary condition. Existence and uniqueness of the numerical solution $(\boldsymbol{U}_1^h, P_1^h, P_2^h)$ are obtained under small data condition similar to (2.17). Convergence rates are optimal. More precisely, there is a constant M independent of h such that

$$\|\boldsymbol{u}_1 - \boldsymbol{U}_1^h\|_{\boldsymbol{X}_1^h} + \|p_1 - P_1^h\|_{Q_1^h} + \|p_2 - P_2^h\|_{Q_2^h} \le M(h^{k_1} + h^{k_2}). \tag{3.7}$$

Using (3.1) and the fact that $\|\cdot\|_{L^2(\Omega_1)} \le M\|\cdot\|_{\boldsymbol{X}_1^h}$ (see [19]), we obtain an error bound of the velocity field in the L^2-norm.

$$\|\boldsymbol{u} - \boldsymbol{U}^h\|_{L^2(\Omega)} \le M(h^{k_1} + h^{k_2}). \tag{3.8}$$

As a consequence, using a trace theorem, an inverse inequality, and the Lagrange interpolant of $\boldsymbol{u}$, we have

$$\forall e \in \Gamma^h, \quad \|\boldsymbol{u} - \boldsymbol{U}^h\|_{L^2(e)} \le M(h^{k_1-1/2} + h^{k_2-1/2}). \tag{3.9}$$

One can also show that the velocity $\boldsymbol{U}^h$ is bounded in the L^2 norm by the data: there is a constant $\overline{M} > 0$ independent of h, but dependent on the data $\mu, \|\boldsymbol{f}_1\|_{L^2(\Omega_1)}$, $\|f_2\|_{L^2(\Omega_2)}$ and $\|\mathcal{U}\|_{L^2(\partial\Omega)}$, such that

$$\|\boldsymbol{U}^h\|_{L^2(\Omega)} \le \overline{M}. \tag{3.10}$$

Remark 3.1. *If the Stokes equations are used instead of the Navier–Stokes equations, existence and uniqueness of the numerical solution is unconditional. Error bounds such as (3.7) are valid.*

3.2. Numerical Approximation of Transport Problem. Equation (2.6) is discretized by a combined backward Euler and DG method. Let Δt be a positive time step and let $t^j = j\Delta t$ denote the time at the jth step. Let Q_h denote the space of discontinuous piecewise polynomials of degree r. The approximation of the initial concentration is obtained by an L^2 projection:

$$\forall q_h \in Q_h, \quad (C_0^h, q_h)_\Omega = (c_0, q_h)_\Omega.$$

For any $j \geq 0$, the approximation C^h_{j+1} of the concentration c at time t^{j+1} is defined by the following discrete variational problem.

$$\forall q_h \in Q_h, \quad \varphi\left(\frac{C^h_{j+1} - C^h_j}{\Delta t}, q_h\right)_\Omega + a_T(\boldsymbol{U}^h; C^h_{j+1}, q_h) + d_T(\boldsymbol{U}^h; C^h_{j+1}, q_h) = L_T(t^{j+1}; q_h), \tag{3.11}$$

where the bilinear form a_T is a DG discretization of the operator $-\nabla \cdot (\boldsymbol{F}(\boldsymbol{u})\nabla c)$ and the bilinear form d_T is a DG discretization of the operator $\nabla \cdot (\boldsymbol{u}c)$. Before defining these forms, we introduce the upwind value $q_h^\uparrow$ of a function q_h in Q_h with respect to the velocity field $\boldsymbol{U}^h$, defined by (3.1). Let e be an edge shared by the elements E^1 and E^2 and assume the unit normal vector $\boldsymbol{n}_e$ points outward of E^1.

$$q_h^\uparrow = \begin{cases} q_h|_{E^1} & \text{if } \{\boldsymbol{U}^h\} \cdot \boldsymbol{n}_e > 0, \\ q_h|_{E^2} & \text{if } \{\boldsymbol{U}^h\} \cdot \boldsymbol{n}_e \leq 0. \end{cases}$$

The penalty parameter is denoted by σ. The symmetrization parameter is denoted by $\epsilon \in \{-1, 1\}$. The forms a_T, d_T, L_T are given below for any θ_h, q_h in Q_h:

$$\begin{aligned} a_T(\boldsymbol{U}^h; \theta_h, q_h) = & \sum_{E \in \mathcal{E}^h} (\boldsymbol{F}(\boldsymbol{U}^h)\nabla\theta_h, \nabla q_h)_E + \sum_{e \in \Gamma^h} |e|^{-1}(\sigma[\theta_h], [q_h])_e \\ & - \sum_{e \in \Gamma^h} ((\boldsymbol{F}(\boldsymbol{U}^h)\nabla\theta_h \cdot \boldsymbol{n}_e)^\uparrow, [q_h])_e \\ & + \epsilon \sum_{e \in \Gamma^h} ((\boldsymbol{F}(\boldsymbol{U}^h)\nabla q_h \cdot \boldsymbol{n}_e)^\uparrow, [\theta_h])_e + \sum_{e \in \partial\Omega} (\theta_h, \mathcal{U}^+ q_h)_e, \\ d_T(\boldsymbol{U}^h; \theta_h, q_h) = & - \sum_{E \in \mathcal{E}^h} (\theta_h \boldsymbol{U}^h, \nabla q_h)_E + \sum_{e \in \Gamma^h} (\theta_h^\uparrow \{\boldsymbol{U}^h \cdot \boldsymbol{n}_e\}, [q_h])_e, \\ L_T(t^{j+1}; q_h) = & \int_\Omega f(t^{j+1}) q_h + \int_{\partial\Omega} \mathcal{C}(t^{j+1}) \mathcal{U}^- q_h. \end{aligned}$$

This scheme uses an improved DG method in which the diffusive fluxes are upwinded whereas in the standard DG method the diffusive fluxes are averaged. The improved method is more robust and does not require the use of slope limiters even in the case of degenerate diffusion coefficients [24]. The space Q_h is equipped with the following semi-norm:

$$|q_h|_{Q_h} = \left(\sum_{E \in \mathcal{E}^h} \|\nabla q_h\|^2_{L^2(E)} + \sum_{e \in \Gamma^h} |e|^{-1} \|\sigma^{1/2}[q_h]\|^2_{L^2(e)} \right).$$

We now recall the coercivity property of the form a_T: there is a constant $\kappa > 0$ such that

$$\forall q_h \in Q_h, \quad a_T(\boldsymbol{U}^h; q_h, q_h) \geq \kappa |q_h|^2_{Q_h} + \|(\mathcal{U}^+)^{1/2} q_h\|^2_{L^2(\partial\Omega)}. \tag{3.12}$$

This is straightforward for the NIPG method ($\epsilon = 1$) and in that case the constant $\kappa = \min(1, \alpha)$ where α is the lower bound for $\boldsymbol{F}(\boldsymbol{u})$. For the SIPG method ($\epsilon = -1$), we use the fact that the matrix $\boldsymbol{F}(\boldsymbol{U}^h)$ is bounded above and the coercivity is obtained if the penalty parameter is large enough.

We will use the following inverse inequality. There is a constant $M > 0$ independent of h such that

$$\forall q_h \in Q_h, \forall E \in \mathcal{E}^h, \quad \|q_h\|_{L^\infty(E)} \leq Mh^{-1}\|q_h\|_{L^2(E)}. \tag{3.13}$$

3.2.1. Existence and Uniqueness of Concentration. As the system is linear, it suffices to show uniqueness. Clearly the initial concentration is uniquely defined. Fix $j \geq 0$. Let $\theta_h = C_h^{j+1} - \tilde{C}_h^{j+1}$ be the difference of two solutions of (3.11). The function θ_h satisfies

$$\frac{\varphi}{\Delta t}\|\theta_h\|^2_{L^2(\Omega)} + a_T(\boldsymbol{U}^h; \theta_h, \theta_h) + d_T(\boldsymbol{U}^h; \theta_h, \theta_h) = 0.$$

Next, we use the coercivity (3.12) of a_T:

$$\frac{\varphi}{\Delta t}\|\theta_h\|^2_{L^2(\Omega)} + \kappa|\theta_h|^2_{Q_h} \leq |d_T(\boldsymbol{U}^h; \theta_h, \theta_h)|.$$

The first term in $d_T(\boldsymbol{U}^h; \theta_h, \theta_h)$ is bounded using Cauchy–Schwarz's inequality, Young's inequality, the inverse inequality (3.13) and the bound (3.10).

$$\begin{aligned}
\left|\sum_{E\in\mathcal{E}^h} (\theta_h \boldsymbol{U}^h, \nabla\theta_h)_E\right| &\leq \sum_{E\in\mathcal{E}^h} \|\theta_h\|_{L^\infty(E)}\|\boldsymbol{U}^h\|_{L^2(E)}\|\nabla\theta_h\|_{L^2(E)} \\
&\leq Mh^{-1}\sum_{E\in\mathcal{E}^h} \|\theta_h\|_{L^2(E)}\|\boldsymbol{U}^h\|_{L^2(E)}\|\nabla\theta_h\|_{L^2(E)} \\
&\leq M\overline{M}h^{-1}\sum_{E\in\mathcal{E}^h} \|\theta_h\|_{L^2(E)}\|\nabla\theta_h\|_{L^2(E)} \\
&\leq \frac{M^2\overline{M}^2}{\kappa h^2}\|\theta_h\|^2_{L^2(\Omega)} + \frac{\kappa}{4}\sum_{E\in\mathcal{E}^h} \|\nabla\theta_h\|^2_{L^2(E)}.
\end{aligned}$$

The second term in $d_T(\boldsymbol{U}^h; \theta_h, \theta_h)$ is bounded similarly, but here we take advantage of the penalty term:

$$\begin{aligned}
&\left|\sum_{e\in\Gamma^h} (\theta_h^\uparrow\{\boldsymbol{U}^h \cdot \boldsymbol{n}_e\}, [\theta_h])_e\right| \\
&\leq M\sum_{e\in\Gamma^h} |e|^{-1/2}\|\sigma^{1/2}[\theta_h]\|_{L^2(e)}h^{1/2}\|\theta_h^\uparrow\|_{L^\infty(e)}\|\{\boldsymbol{U}^h \cdot \boldsymbol{n}_e\}\|_{L^2(e)} \\
&\leq M\sum_{e\in\Gamma^h} |e|^{-1/2}\|\sigma^{1/2}[\theta_h]\|_{L^2(e)}h^{1/2}h^{-1}\|\theta_h\|_{L^2(E_e^{12})}\|\{\boldsymbol{U}^h \cdot \boldsymbol{n}_e\}\|_{L^2(e)} \\
&\leq M\sum_{e\in\Gamma^h} |e|^{-1/2}\|\sigma^{1/2}[\theta_h]\|_{L^2(e)}h^{1/2}h^{-1}\|\theta_h\|_{L^2(E_e^{12})}h^{-1/2}\|\boldsymbol{U}^h\|_{L^2(E_e^{12})}.
\end{aligned}$$

In the bound above we have used the inverse inequality $\|\boldsymbol{U}^h\|_{L^2(e)} \leq Mh^{-1/2}\|\boldsymbol{U}^h\|_{L^2(E)}$. We also defined the union of the elements who share the edge e by E_e^{12}. Next, we use the bound on the discrete velocity (3.10) and we obtain by Young's inequality:

$$\left|\sum_{e\in\Gamma^h}(\theta_h^{\uparrow}\{\boldsymbol{U}^h\cdot\boldsymbol{n}_e\},[\theta_h])_e\right| \leq \frac{M^2\overline{M}^2}{h^2\kappa}\|\theta_h\|^2_{L^2(\Omega)}+\frac{\kappa}{4}\sum_{e\in\Gamma^h}|e|^{-1}\|\sigma^{1/2}[\theta_h]\|^2_{L^2(e)}.$$

Therefore we have

$$\left(\frac{\varphi}{\Delta t}-\frac{2M^2\overline{M}^2}{\kappa h^2}\right)\|\theta_h\|^2_{L^2(\Omega)}+\frac{3\kappa}{4}|\theta_h|^2_{Q_h}\leq 0.$$

We conclude that $\theta_h = 0$ if the time step satisfies the following condition:

$$\Delta t < \frac{\kappa h^2\varphi}{2M^2\overline{M}^2}.$$

We summarize our result below.

LEMMA 3.1. *There is a constant $M_0 > 0$ such that if $\Delta t < M_0h^2$, there is a unique solution to the scheme (3.11).*

3.2.2. Error Analysis. We decompose the error at each time step into an approximation error η and a numerical error ξ. Let $\tilde{c}\in Q_h\cap\mathcal{C}(\overline{\Omega})$ be an approximation of c in the sense that the following approximation bounds [5] hold:

$$\|c(t^j)-\tilde{c}(t^j)\|_{L^2(\Omega)}\leq Mh^{r+1}\|c(t^j)\|_{H^{r+1}(\Omega)},$$

$$\|\nabla(c(t^j)-\tilde{c}(t^j))\|_{L^2(\Omega)}\leq Mh^{r}\|c(t^j)\|_{H^{r+1}(\Omega)},$$

$$\|c(t^j)-\tilde{c}(t^j)\|_{L^\infty(\Omega)}\leq Mh^{r+1}\|c(t^j)\|_{H^{r+1}(\Omega)},$$

$$\|\nabla(c(t^j)-\tilde{c}(t^j))\|_{L^\infty(\Omega)}\leq Mh^{r}\|c(t^j)\|_{H^{r+1}(\Omega)}.$$

We write

$$C_h^j-c(t^j)=\eta^j-\xi^j,\quad \eta^j=C_h^j-\tilde{c}(t^j),\quad \xi^j=c(t^j)-\tilde{c}(t^j).$$

THEOREM 3.1. *Under the assumption of Lemma 3.1 and the additional regularity assumptions $c\in L^2(0,T;H^{r+1}(\Omega))\cap W^{1,\infty}(\Omega)$, $c_t\in L^2(0,T;H^r(\Omega))$, and $c_0\in H^r(\Omega)$, there is a constant M independent of h and Δt such that for all $m\geq 1$, and for Δt small enough, we have the error bound*

$$\|\eta^m\|^2_\Omega+\kappa\Delta t\sum_{j=1}^m|\eta^j|^2_{Q_h}+\Delta t\sum_{j=1}^m\||\mathcal{U}|^{1/2}\eta^j\|^2_{\partial\Omega}\leq M(h^{2r}+h^{2k_1}+h^{2k_2}+\Delta t^2).$$

Proof. The error equation becomes

$$\forall q_h \in Q_h, \quad \left(\varphi \frac{\eta^{j+1}-\eta^j}{\Delta t}, q_h\right)_\Omega + a_T(\boldsymbol{U}^h; \eta^{j+1}, q_h) + d_T(\boldsymbol{u}; \eta^{j+1}, q_h)$$
$$= \left(\varphi \frac{\partial \xi}{\partial t}(t^{j+1}), q_h\right)_\Omega + \left(\varphi \frac{\partial \tilde{c}}{\partial t}(t^{j+1}) - \varphi \frac{\tilde{c}^{j+1}-\tilde{c}^j}{\Delta t}, q_h\right)_\Omega$$
$$+ d_T(\boldsymbol{u}-\boldsymbol{U}^h; \eta^{j+1}, q_h) + a_T(\boldsymbol{U}^h; \xi^{j+1}, q_h) + d_T(\boldsymbol{U}^h; \xi^{j+1}, q_h)$$
$$+ d_T(\boldsymbol{u}-\boldsymbol{U}^h; c(t^{j+1}), q_h) + a_T(\boldsymbol{u}; c(t^{j+1}), q_h) - a_T(\boldsymbol{U}^h; c(t^{j+1}), q_h).$$

We take $q_h = \eta^{j+1}$ and we use the coercivity (3.12) of a_T:

$$\frac{\varphi}{2\Delta t}(\|\eta^{j+1}\|^2_{L^2(\Omega)} - \|\eta^j\|^2_{L^2(\Omega)}) + \kappa |\eta^{j+1}|^2_{Q_h} + d_T(\boldsymbol{u}; \eta^{j+1}, \eta^{j+1})$$
$$+ \|(\mathcal{U}^+)^{1/2}\eta^{j+1}\|^2_{L^2(\partial\Omega)} \le \left|\left(\frac{\partial \xi}{\partial t}(t^{j+1}), \eta^{j+1}\right)_\Omega\right| + |d_T(\boldsymbol{u}-\boldsymbol{U}^h; \eta^{j+1}, \eta^{j+1})|$$
$$+ \left|\left(\frac{\partial \tilde{c}}{\partial t}(t^{j+1}) - \frac{\tilde{c}^{j+1}-\tilde{c}^j}{\Delta t}, \eta^{j+1}\right)_\Omega\right| + |a_T(\boldsymbol{U}^h; \xi^{j+1}, \eta^{j+1})|$$
$$+ |d_T(\boldsymbol{U}^h; \xi^{j+1}, \eta^{j+1})| + |d_T(\boldsymbol{u}-\boldsymbol{U}^h; c(t^{j+1}), \eta^{j+1})|$$
$$+ |a_T(\boldsymbol{u}; c(t^{j+1}), \eta^{j+1}) - a_T(\boldsymbol{U}^h; c(t^{j+1}), \eta^{j+1})|. \tag{3.14}$$

Since the weak solution satisfies $\nabla \cdot \boldsymbol{u}|_{\Omega_1} = \boldsymbol{0}$ and $\nabla \cdot \boldsymbol{u}|_{\Omega_2} = f_2 \ge 0$, we use integration by parts and obtain:

$$d_T(\boldsymbol{u}; \eta^{j+1}, \eta^{j+1}) + \|(\mathcal{U}^+)^{1/2}\eta^{j+1}\|^2_{L^2(\partial\Omega)} = \frac{1}{2}(\mathcal{U}^+, (\eta^{j+1})^2)_{\partial\Omega}$$
$$+ \frac{1}{2}(\mathcal{U}^-, (\eta^{j+1})^2)_{\partial\Omega}.$$

We now bound the first and second terms in the right-hand side of (3.14), by the approximation properties, under the regularity assumption for the exact solution c.

$$\left|\left(\frac{\partial \xi}{\partial t}(t^{j+1}), \eta^{j+1}\right)_\Omega\right| \le \|\eta^{j+1}\|^2_{L^2(\Omega)} + Mh^{2r}\left\|\frac{\partial c}{\partial t}(t^{j+1})\right\|^2_{H^r(\Omega)},$$

and

$$\left|\left(\frac{\partial \tilde{c}}{\partial t}(t^{j+1}) - \frac{\tilde{c}^{j+1}-\tilde{c}^j}{\Delta t}, \eta^{j+1}\right)_\Omega\right| \le \|\eta^{j+1}\|^2_{L^2(\Omega)} + \frac{\Delta t}{12}\int_{t^j}^{t^{j+1}} \left\|\frac{\partial^2 \tilde{c}}{\partial t^2}\right\|^2_{L^2(\Omega)}.$$

We now bound the d_T terms. Using standard techniques and inequality (3.13), we obtain

$$d_T(\boldsymbol{u}-\boldsymbol{U}^h; \eta^{j+1}, \eta^{j+1}) \le Mh^{-1}\|\eta^{j+1}\|_{L^2(\Omega)}\|\boldsymbol{u}-\boldsymbol{U}^h\|_{L^2(\Omega)}|\eta^{j+1}|_{Q_h}.$$

Using the velocity bound (3.8) and the fact that $k_1 \ge 1, k_2 \ge 1$, we have

$$d_T(\boldsymbol{u}-\boldsymbol{U}^h; \eta^{j+1}, \eta^{j+1}) \le \frac{\kappa}{8}|\eta^{j+1}|^2_{Q_h} + M\|\eta^{j+1}\|^2_{L^2(\Omega)}.$$

Similarly, using (3.10), we have

$$d_T(\boldsymbol{U}^h;\xi^{j+1},\eta^{j+1}) \le M\|\xi^{j+1}\|_{L^\infty(\Omega)}\|\boldsymbol{U}^h\|_{L^2(\Omega)}|\eta^{j+1}|_{Q_h} \le M\|\xi^{j+1}\|_{L^\infty(\Omega)}|\eta^{j+1}|_{Q_h}.$$

and using (3.8), (3.9) and the boundedness of the weak solution, we have

$$d_T(\boldsymbol{u}-\boldsymbol{U}^h;c(t^{j+1}),\eta^{j+1}) \le M\|c(t^{j+1})\|_{L^\infty(\Omega)}|\eta^{j+1}|_{Q_h} \left(\|\boldsymbol{u}-\boldsymbol{U}^h\|_{L^2(\Omega)}+\left(\sum_{e\in\Gamma^h}|e|\|\boldsymbol{u}-\boldsymbol{U}^h\|^2_{L^2(e)}\right)^{1/2}\right) \le \frac{\kappa}{8}|\eta^{j+1}|^2_{Q_h}+M(h^{2k_1}+h^{2k_2}).$$

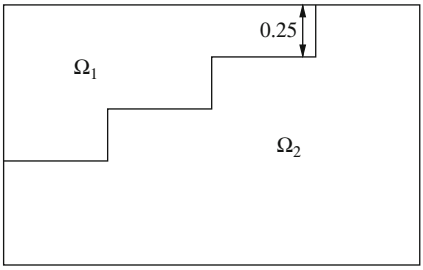


FIG. 1. Domain with step interface

The diffusive term $a_T(\boldsymbol{U}^h;\xi^{j+1},\eta^{j+1})$ is bounded using standard techniques.

$$a_T(\boldsymbol{U}^h;\xi^{j+1},\eta^{j+1}) \le \frac{\kappa}{8}|\eta^{j+1}|^2_{Q_h}+\frac{1}{8}\|(\mathcal{U}^+)^{1/2}\eta^{j+1}\|^2_{L^2(\partial\Omega)} +Mh^{2r}\|c(t^{j+1})\|^2_{H^{r+1}(\Omega)}.$$

To bound the remaining diffusive terms, we use the boundedness of c, the Lipschitz continuity of $\boldsymbol{F}$ and the bounds (3.8) and (3.9).

$$a_T(\boldsymbol{u};c(t^{j+1}),\eta^{j+1})-a_T(\boldsymbol{U}^h;c(t^{j+1}),\eta^{j+1}) \le \frac{\kappa}{8}|\eta^{j+1}|^2_{Q_h}+M(h^{2k_1}+h^{2k_2}).$$

We can now conclude by combining all bounds, summing over the time steps, and using Gronwall's inequality. □

4. Numerical Examples. In this section, we show that our scheme is robust under different physical conditions (faults, discontinuous permeability field). In all the numerical examples, the fluid viscosity is equal to 1, and the Beavers–Joseph–Saffman constant G is equal to 0.1. Meshes are

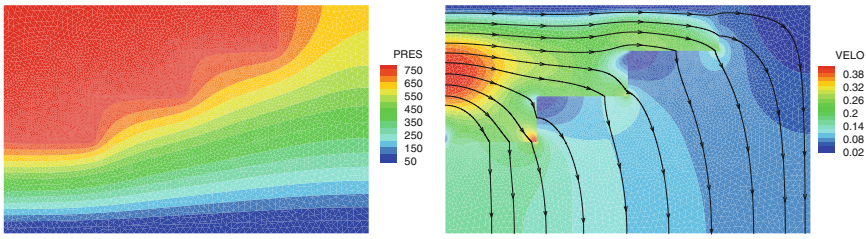


FIG. 2. Step interface problem: pressure contours (*left*) and velocity norm and streamlines (*right*)

generated using Gmsh [17], visualization is done using Tecplot [1] and the simulations are done using software developed by the author. The linear systems are solved by a sparse direct solver. Uniqueness of the pressure is obtained by imposing a Dirichlet boundary condition on part of the subsurface boundary.

4.1. Step Interface. In the first example, the rectangular domain $\Omega = (0,2) \times (0,1.25)$ is partitioned into two subdomains by a polygonal interface with three successive uniform steps (see Fig. 1). For the flow problem, the Stokes equations are solved in Ω_1 and the Darcy equations in Ω_2.

The permeability of Ω_2 is $\boldsymbol{K} = 10^{-4}\boldsymbol{I}$. Zero Dirichlet boundary conditions are imposed on the bottom horizontal side of Ω_2 and zero Neumann boundary conditions on the remainder of $\Omega_2 \setminus \Gamma_{12}$. The Stokes velocity on Γ_1 is set equal to $(-3(y-1.25)(y-0.5), 0)$, which means the velocity profile is parabolic along the vertical side of Γ_1. The pressure contours and Euclidean norm of velocity contours with streamlines are shown in Fig. 2. The DG scheme is used with $\epsilon = 1$, $\sigma = 0.1$ and $k_1 = k_2 = 2$. The mesh contains 5,760 triangles of varying size so that the triangles in the neighborhood of the interface are the smallest. We now describe the characteristics of the transport problem for this example. The coefficients are: $\varphi = 0.2$, $\alpha_l = 1$, $\alpha_t = 0.1$, $c_0 = \mathcal{C} = 0$, $d_m = 10^{-3}$ in Ω_2, $d_m = 5 \times 10^{-3}$ in Ω_1. In this example, we simulate the leakage of a contaminant in the surface by the source function:

$$f(t,x,y) = \begin{cases} 1, & t < 1, \text{ and } ((x-0.2)^2 + (y-0.85)^2)^{1/2} \leq 0.15 \\ 0, & \text{otherwise.} \end{cases}$$

Concentration contours at the time $t = 1$ where the plume reaches its maximum peak are shown in Fig. 3 (left). At this time, the contaminant has just reached the interface. Concentration contours at later times are shown on Fig. 3 (right) and Fig. 4. Once the contaminant penetrates the subsurface it is transported downwards and exits the domain via the bottom

horizontal boundary. The numerical approximation of the concentration is obtained with the DG method with $\epsilon = 1, r = 1, \sigma = 0.1$ and $\Delta t = 2.5 \times 10^{-3}$.

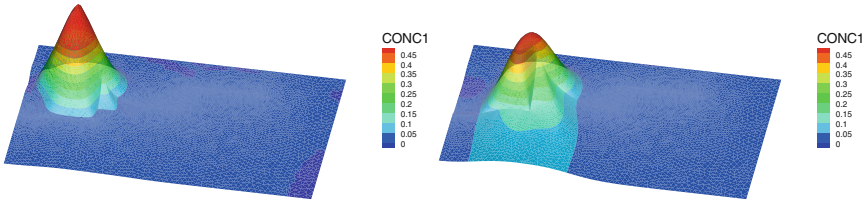


FIG. 3. Concentration contours at $t = 1$ (*left*) and $t = 1.5$ (*right*)

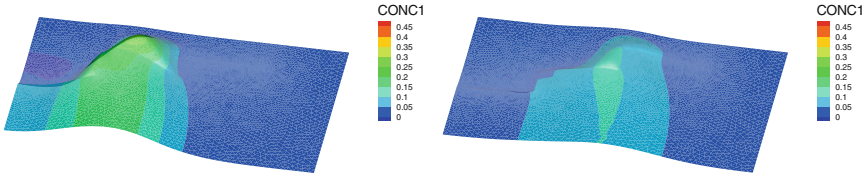


FIG. 4. Concentration contours at $t = 2.5$ (*left*) and $t = 4$ (*right*)

FIG. 5. Nonuniform permeability problem

4.2. Nonuniform Permeability Field. In this second example, the permeability of the subsurface takes random values between $10^{-7}\boldsymbol{I}$ and $3.8 \times 10^{-5}\boldsymbol{I}$. The domain is $\Omega = (0, 12) \times (0, 6)$ and the interface is a horizontal line containing two steps of opposite direction. Figure 5 shows the domain and the permeability distribution.

First for the flow problem, we impose a parabolic velocity profile on the left vertical boundary of Ω_1 and a similar profile on the right vertical boundary of Ω_1 but with a smaller magnitude. Zero Neumann boundary

conditions are imposed on the Darcy pressure for the vertical boundaries of Ω_2 and Dirichlet pressure is prescribed on bottom horizontal boundary. The Dirichlet values are given below:

$$\forall y \geq 4,\ \boldsymbol{u}_1(0,y) = \left(\frac{1}{4}(y-4)(8-y), 0\right),\ \boldsymbol{u}_1(12,y) = \left(\frac{3}{16}(y-4)(8-y), 0\right),$$
$$\forall 0 \leq x \leq 12, \quad \boldsymbol{u}_1(x,6) = (1,0), \quad p_2(x,0) = 10^5.$$

The Navier–Stokes equations are solved in Ω_1 and the Darcy equations in Ω_2. The mesh contains 562 triangles in the surface and 625 triangles in the subsurface. The DG method with parameters $\sigma = 1, \epsilon = 1, k_1 = k_2 = 2$ is used. The Picard iterations for the flow problem converge after nine iterations, with a set tolerance of 10^{-7}. Figure 6 shows the pressure contours and the velocity field. Since the exact solution is unknown, we compute the differences between the solutions obtained on two successive meshes (i.e., of size h and $h/2$). We obtain a rate of $\mathcal{O}(h^{0.4})$ for the H^1 norm of the Navier–Stokes velocity and $\mathcal{O}(h^{0.4})$ for the H^1 norm of the Darcy pressure. These rates confirm convergence of the scheme for solutions with low regularity.

Second for the transport problem, the concentration is prescribed on the inflow boundary ($\mathcal{C} = 1$). The initial concentration is zero. The other parameters defining the problem are: $r = 1$, $\varphi = 0.2$, $\alpha_l = 0.1$, $\alpha_t = 0.01$, $d_m = 10^{-4}$ in Ω_2, $d_m = 10^{-2}$ in Ω_1. Discontinuous piecewise linear approximation of the concentration is computed with the following parameters: $\sigma = \epsilon = r = 1$. Figures 7–9 present the concentration contours at successive times. We observe that the contaminant sweeps the surface region very fast, then percolates down the subsurface at a slower rate. This is expected as the velocity in the subsurface is much smaller than the velocity in the surface. We also note that the contaminant is transported downwards in the subsurface in a nonuniform way. This is explained by the discontinuous distribution of the permeability field.

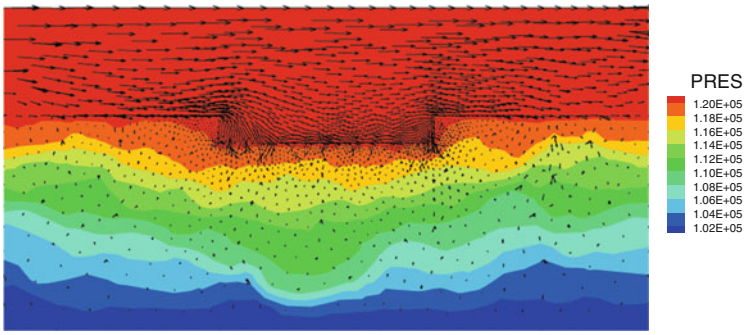


FIG. 6. Nonuniform permeability problem: pressure contours and velocity field

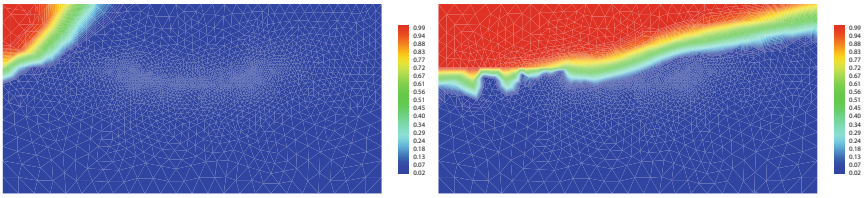


FIG. 7. Concentration at different times: $t_1 = 0.5$ (*left*) and $t_2 = 3$ (*right*)

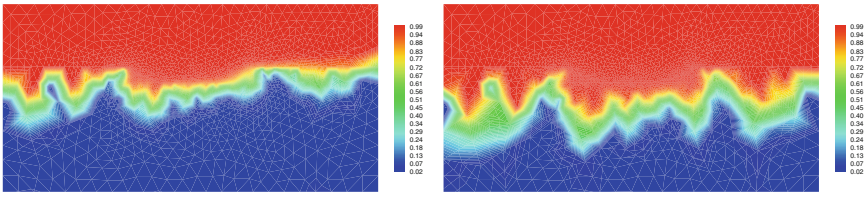


FIG. 8. Concentration at different times: $t_3 = 8$ (*left*) and $t_4 = 13$ (*right*)

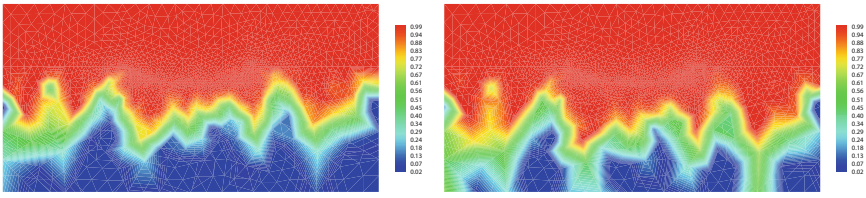


FIG. 9. Concentration at different times: $t_5 = 18$ (*left*) and $t_6 = 20$ (*right*)

4.3. Fractured Subsurface. In this example, the porous medium contains three horizontal layers of varying permeability that are intersected by two slanted faults ($\Omega = (0, 12) \times (0, 6)$). The permeability matrix is equal to $10^{-4}\mathbf{I}$, $10^{-9}\mathbf{I}$, $10^{-5}\mathbf{I}$, $10^{-7}\mathbf{I}$ in the faults, the top layer, the middle layer, and the bottom layer, respectively (see Fig. 10). Boundary conditions for the flow problem are the same as in the previous example (Sect. 4.2). Figure 11 shows the pressure contours and the velocity field obtained with the DG method of first and second order, which yields 8,707 and 17,679 degrees of freedom, respectively. The pressure follows a vertical gradient, and thus the velocity in the middle layer (denoted by B in Fig. 10) remains small.

Next we describe the parameters chosen for the transport problem. The coefficients are: $\varphi = 0.2$, $\alpha_l = 0.1$, $\alpha_t = 0.01$, $\mathcal{C} = 0$, $d_m = 10^{-4}$ in Ω_2, $d_m = 10^{-2}$ in Ω_1. As in the first example, we simulate the leakage of a contaminant in the surface. The initial concentration is equal to one in a localized region in the surface, and zero elsewhere. In addition, there is a temporary source of contaminant (for $t \leq t^*$, with $t^* = 3$) defined by:

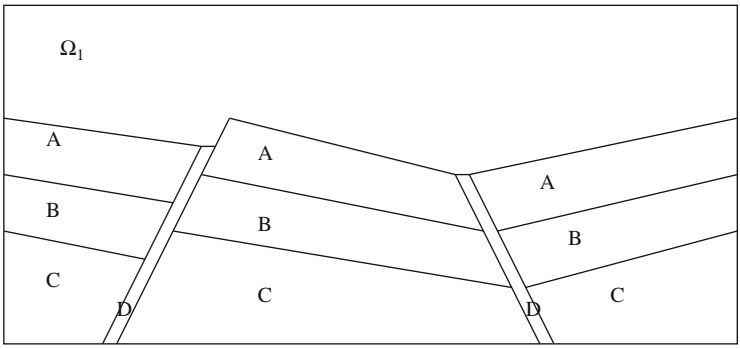


FIG. 10. Domain for surface coupled with fractured subsurface. Permeability value is 10^{-9} in A region, 10^{-5} in B region, 10^{-7} in C region, and 10^{-4} in D region (slanted fractures)

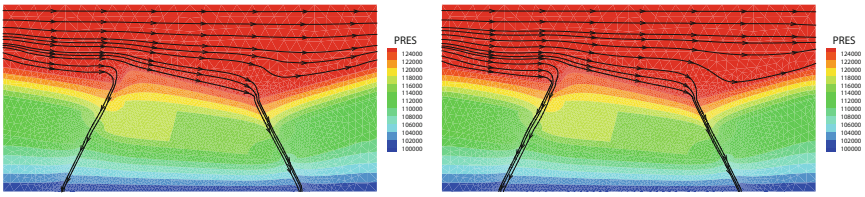


FIG. 11. Fractured subsurface problem: pressure and velocity field obtained with the DG method of order one (*left figure*) and order two (*right figure*)

$$f(t,x,y) = \begin{cases} 0.5, & t < 3, \text{ and } ((x-2.0)^2 + (y-5.1)^2)^{1/2} \leq 0.5 \\ 0, & \text{otherwise.} \end{cases}$$

As in the previous example, we obtain the numerical approximation of the concentration by the DG method with parameters $r = \epsilon = \sigma = 1$. In Figs. 12–14, we show the concentration contours at different times. We note that the mesh used for the transport problem is the same as the one used in Fig. 11. The overall behavior of the solution is as expected: the contaminant is transported faster in the surface region, and some of it penetrates the subsurface via the slanted fractures. Because of the intermediate value of the permeability in the middle layer, some of the contaminant appears in part of region B neighboring the fractures.

5. Conclusions. The coupling of surface/subsurface flow and transport is studied theoretically and numerically by the use of finite element methods and discontinuous Galerkin methods. It is shown that the DG scheme is robust and yields accurate solutions for inhomogeneous or fractured subsurface. It would be of interest to study the effects of projection of the velocity field, if independent meshes are used for the flow and transport problems.

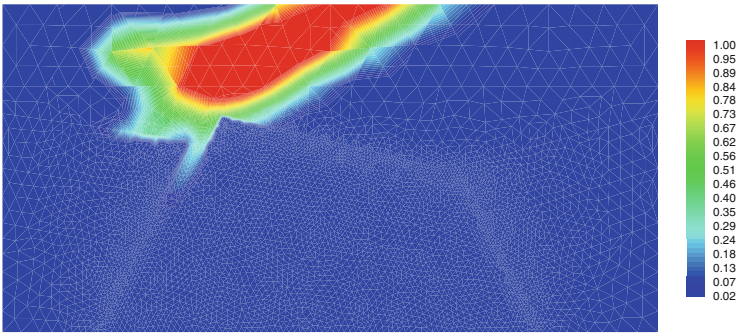


FIG. 12. Concentration contours at time t_1

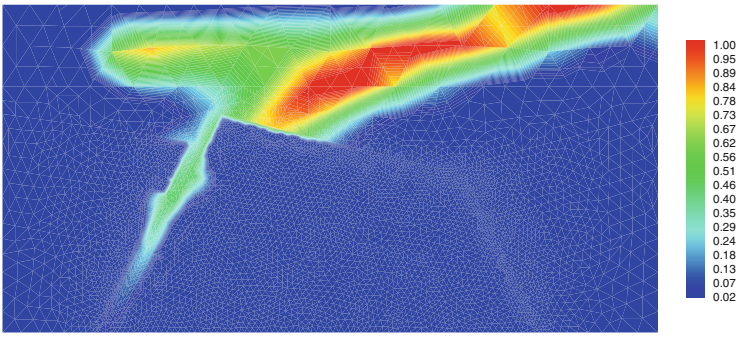


FIG. 13. Concentration contours at time $t_2 = 2t_1$

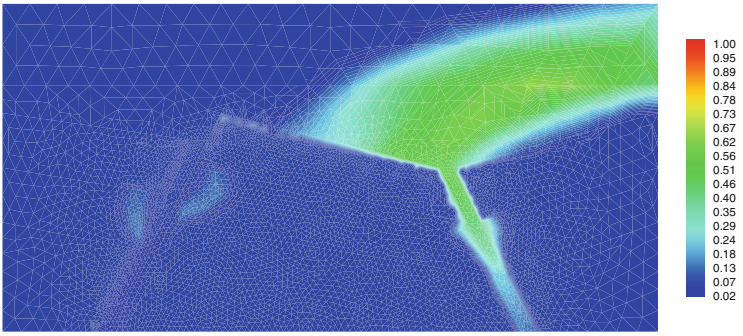


FIG. 14. Concentration contours at time $t_3 = 5t_1$

Acknowledgment. The author acknowledges the support of National Science Foundation through the grant DMS 0810422.

REFERENCES

[1] http://www.tecplot.com.
[2] H. W. Alt and S. Luckhaus. Quasilinear elliptic-parabolic differential equations. *Math. Z.*, 183(3):311–341, 1983.
[3] D.N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19:742–760, 1982.
[4] G.S. Beavers and D.D. Joseph. Boundary conditions at a naturally impermeable wall. *J. Fluid. Mech*, 30:197–207, 1967.
[5] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer-Verlag, New York, 1994.
[6] A. Cesmelioglu, V. Girault, and B. Rivière. Time-dependent coupling of Navier-Stokes and Darcy flows. *ESAIM: Mathematical Modelling and Numerical Analysis*, 47:539–554, 2013.
[7] A. Cesmelioglu and B. Rivière. Analysis of time-dependent Navier-Stokes flow coupled with Darcy flow. *Journal of Numerical Mathematics*, 16:249–280, 2008.
[8] A. Cesmelioglu and B. Rivière. Primal discontinuous Galerkin methods for time-dependent coupled surface and subsurface flow. *Journal of Scientific Computing*, 40:115–140, 2009.
[9] A. Cesmelioglu and B. Rivière. Existence of a weak solution for the fully coupled Navier-Stokes/Darcy-transport problem. *Journal of Differential Equations*, 252:4138–4175, 2012.
[10] P. Chidyagwai and B. Rivière. On the solution of the coupled Navier-Stokes and Darcy equations. *Computer Methods in Applied Mechanics and Engineering*, 198:3806–3820, 2009.
[11] P. Chidyagwai and B. Rivière. Numerical modelling of coupled surface and subsurface flow systems. *Advances in Water Resources*, 33:92–105, 2010.
[12] P. Ciarlet. *The finite element method for elliptic problems*. North-Holland, Amsterdam, 1978.
[13] B. Cockburn and C.-W. Shu. The local discontinuous Galerkin method for time-dependent convection-diffusion systems. *SIAM Journal on Numerical Analysis*, 35:2440–2463, 1998.
[14] M. Discacciati and A. Quarteroni. Analysis of a domain decomposition method for the coupling of Stokes and Darcy equations. In Brezzi et al, editor, *Numerical Analysis and Advanced Applications - ENUMATH 2001*, pages 3–20. Springer, Milan, 2003.
[15] M. Discacciati, A. Quarteroni, and A. Valli. Robin-Robin domain decomposition methods for the Stokes-Darcy coupling. *SIAM J. Numer. Anal.*, 45(3):1246–1268, 2007.
[16] G.N. Gatica, R. Oyarzúa, and F.-J. Sayas. Convergence of a family of Galerkin discretizations for the Stokes-Darcy coupled problem. *Numerical Methods for Partial Differential Equations*, 2009. DOI: 10.1002/num.20548.
[17] C. Geuzaine and J.-F. Remacle. Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities. *International Journal for Numerical Methods in Engineering*, 79:1309–1331, 2009. http://www.geuz.org/gmsh/.
[18] V. Girault and B. Rivière. DG approximation of coupled Navier-Stokes and Darcy equations by Beaver-Joseph-Saffman interface condition. *SIAM Journal on Numerical Analysis*, 47:2052–2089, 2009.

[19] V. Girault, B. Rivière, and M. Wheeler. A discontinuous Galerkin method with non-overlapping domain decomposition for the Stokes and Navier-Stokes problems. *Mathematics of Computation*, 74:53–84, 2004.

[20] N.S. Hanspal, A.N. Waghode, V. Nassehi, and R.J. Wakeman. Numerical analysis of coupled Stokes/Darcy flows in industrial filtrations. *Transport in Porous Media*, 64(1):1573–1634, 2006.

[21] W.J. Layton, F. Schieweck, and I. Yotov. Coupling fluid flow with porous media flow. *SIAM J. Numer. Anal.*, 40(6):2195–2218, 2003.

[22] F. Marpeau and M. Saad. Mathematical analysis of radionuclides displacement in porous media with nonlinear adsorption. *J. Differ. Equations*, 228(2):412–439, 2006.

[23] M. Mu and J. Xu. A two-grid method of a mixed Stokes-Darcy model for coupling fluid flow with porous media flow. *SIAM Journal on Numerical Analysis*, 45:1801–1813, 2007.

[24] J. Proft and B. Rivière. Discontinuous Galerkin methods for convection-diffusion equations with varying and vanishing diffusivity. *International Journal of Numerical Analysis and Modeling*, 6:533–561, 2009.

[25] B. Rivière, M.F. Wheeler, and V. Girault. A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems. *SIAM Journal on Numerical Analysis*, 39(3):902–931, 2001.

[26] P. Saffman. On the boundary condition at the surface of a porous media. *Stud. Appl. Math.*, 50:292–315, 1971.

[27] S. Sun, B. Rivière, and M.F. Wheeler. A combined mixed finite element and discontinuous Galerkin method for miscible displacement problem in porous media. *In Recent Progress in Computational and Applied PDEs*, pages 323–348, 2002. Kluwer/Plenum, New York.

[28] D. Vassilev and I. Yotov. Coupling Stokes-Darcy flow with transport. *SIAM Journal on Scientific Computing*, 31(5):3661–3684, 2009.

[29] M.F. Wheeler. An elliptic collocation-finite element method with interior penalties. *SIAM Journal on Numerical Analysis*, 15(1):152–161, 1978.