Adaptive Finite Element Approximation of Hyperbolic Problems

Endre Süli[1] and Paul Houston[2,⋆]

[1] University of Oxford,
Computing Laboratory,
Wolfson Building, Parks Road,
Oxford OX1 3QD, UK
email: endre@comlab.ox.ac.uk

[2] University of Leicester,
Department of Mathematics and Computer Science,
University Road,
Leicester LE1 7RH, UK
email: Paul.Houston@mcs.le.ac.uk

Abstract. We review some recent developments concerning the *a posteriori* error analysis of $h-$ and $hp-$version finite element approximations to hyperbolic problems. The error bounds stem from an error representation formula which equates the error in an output functional of interest to the inner product of the finite element residual with the solution of a dual (adjoint) problem whose data is the density function of the target functional. *Type I a posteriori* error bounds are derived which, unlike the cruder *Type II* bounds, retain the dual solution in the bound as a local weight-function. The relevance of Type I *a posteriori* bounds is argued by showing that the local size of the error in a hyperbolic problem may be only very weakly correlated to the local size of the residual; consequently, adaptive refinement algorithms based on the size of the local residual alone can be ineffective. The sharpness of Type I *a posteriori* error bounds is demonstrated on both structured and adaptively refined meshes.

1 Introduction

Over the last decade, tremendous progress has been made in the area of *a posteriori* error estimation and adaptive finite element approximation of partial differential equations (PDEs); for a review of some of the main developments in the subject we refer to the recent monographs [4,56,57], and the articles [14,8]. Despite a number of significant advances in the field, much of the research to date has focused on partial differential equations of elliptic type, while for hyperbolic partial differential equations the theory of *a posteriori* error estimation and adaptivity has remained in its infancy.

This state of affairs is perhaps not accidental: the mechanisms of error propagation in finite element approximations of hyperbolic PDEs are more

⋆ Paul Houston acknowledges the financial support of the EPSRC (GR/N24230).

complex and less well understood than in elliptic problems; a further hindrance is that the theory of well-posedness and regularity for hyperbolic PDEs is in a less satisfactory state than the corresponding theoretical framework for elliptic problems due, among other things, to the fact that unlike elliptic differential operators, first-order hyperbolic operators have no inherent smoothing properties in isotropic Sobolev spaces; finally, in engineering applications, finite element methods have been traditionally employed for the numerical approximation of elliptic boundary value problems and their use for the solution of hyperbolic PDEs is more recent and less widespread.

The aim of these notes is to provide a brief introduction to some recent advances in the field of *a posteriori* error estimation and adaptivity for hyperbolic PDEs. The focus of this work and its scope is, necessarily, influenced by our own interests in the subject and we do not attempt to give an exhaustive survey. For further developments in the field we refer to [11,24,29,52,53], and [25–28,30,35,40,42,54] for some more recent contributions. In order to keep the presentation simple, we shall concentrate on two popular finite element methods: the streamline–diffusion finite element method and the discontinuous Galerkin finite element method; nevertheless, the ideas developed here apply more generally, to other partial differential equations and to other finite element discretizations. The *a posteriori* error analyses for the two methods will be pursued within a single, general theoretical framework based on the idea of duality.

The relevance of duality arguments in *a posteriori* error estimation has been highlighted in the review articles [14] and [8] (see, also, [7,24,30,35,42,44–47,49,53,54]); concerning the use of duality arguments in post-processing and design, we refer to [17,18,20,21,48], and references therein. The key ingredient in duality-based error estimation is an auxiliary PDE problem, the dual problem, involving the formal adjoint of the partial differential operator under consideration. The data for the dual problem is the quantity of interest: in engineering applications, this is typically an output functional of the analytical solution such as a mean value, point value, boundary flux, for example, or, in fluid dynamics, it may be the pressure–drop between inflow and outflow, or the drag or lift coefficients of a body immersed into the fluid (see, [7,8,17,18,20,21,27,42,45–48], for example).

The relevance and generality of duality-based error estimation has been powerfully argued in the work of Johnson and his collaborators; see, for example, [14] for an excellent survey. The *a posteriori* error bounds resulting from this analysis involve the finite element residual which is obtained by inserting the computed finite element solution into the partial differential equation; the residual measures the extent to which the finite element approximation to the analytical solution fails to satisfy the underlying PDE. In the framework of [14], the error bounds are arrived at by exploiting Galerkin orthogonality (a fundamental property of all finite element methods expressing the fact that the residual is orthogonal to the finite element space), and strong stability (well-posedness/regularity in isotropic Sobolev norms of positive index) of the dual problem. In view of the remarks at the end of the second paragraph concerning the inherent lack of isotropic smoothing properties in hyperbolic problems, it is this latter step in the process of *a posteriori* error estimation that causes difficulties when dealing with hyperbolic PDEs: the underlying problems will be explained in more detail in the next section.

Stimulated by our findings, in this paper we revisit the problem of *a posteriori* error estimation and adopt a minimalistic variant of duality–based error estimation which avoids explicitly invoking strong stability of the dual problem through retention of the dual solution in the bound; the approach in question is stimulated by the work of Becker and Rannacher [7] on weighted-residual-based *a posteriori* error estimation; see also the recent review article [8] and references therein. As in [14] and [8], the starting point in our error analysis is an *error representation formula* which equates the error in the output functional of interest with the inner product over the computational domain of the finite element residual and the analytical solution to the dual problem. In our approach, the inner product over the computational domain contained in the error representation formula is decomposed as a sum of inner products over elements in the finite element partition, and hence the absolute value of the error in the output functional is bounded by the sum of absolute values of elemental inner products; this process is referred to as *localization*. However, unlike [14] and [8], we shall refrain from applying the Cauchy–Schwarz inequality to these elemental inner products; rather, we shall compute them numerically through replacing the analytical dual solution by a finite element approximation. If necessary, the error committed in the course of this additional approximation can be controlled through an auxiliary *a posteriori* error estimate. It will be shown that the proposed approach leads to sharp *a posteriori* error bounds, in the sense that the *effectivity index* (the ratio of the error bound and the true error) is typically very close to unity.

The paper is structured as follows. We begin, in Section 2, by outlining the fundamental problems associated with residual-based error estimation in the context of hyperbolic problems. We then continue, in Section 3, with a brief excursion into the theory of function spaces to introduce the notational conventions. In Section 4, we discuss some of the key model problems whose numerical approximations we shall consider. In this work, we shall concentrate on two established finite element methods for hyperbolic problems: the streamline–diffusion finite element method (SDFEM) and the discontinuous Galerkin finite element method (DGFEM). Thus, Section 5 is devoted to the formulation of the SDFEM which uses continuous piecewise polynomial functions to approximate the underlying analytical solution. For the sake of simplicity we shall confine ourselves to a scalar linear hyperbolic model problem. We shall be concerned with the derivation of *a posteriori* error bounds for the SDFEM within the framework of h–version finite element methods (i.e., with variable mesh size h, but fixed polynomial degree p, $p \geq 1$); in particular, we shall focus on situations when the quantity of interest is a linear functional of the analytical solution rather than the solution itself. The derivation of the *a posteriori* error bound is based on a duality argument which may involve either the formal adjoint of the partial differential operator or the formal adjoint of a perturbed partial differential operator which stems from the bilinear form of the stabilized finite element method. It will be shown in Section 5 that for stabilized finite element methods, such as the SDFEM, the second approach leads to sharper *a posteriori* error bounds. In Section 6, we shall consider the hp–version of the discontinuous Galerkin finite element method, based on the use of discontinuous piecewise polynomial approximations, which, in addition to local variation of the granularity h of the computational mesh, also allows local adjustment of the degree p of the approximating piecewise polynomial. We shall discuss the *a posteriori* error analysis of the hp–version of the discontinuous Galerkin finite element method and consider the implementation of our *a posteriori* error bounds into an hp–adaptive algorithm. For hp–adaptive finite element methods, a key question is to decide when to h–refine and when to p–refine. The decision concerning which of the two refinement strategies to adopt will be based on monitoring the local smoothness of the solutions to the primal and dual problems and will involve an algorithm for the estimation of their local Sobolev regularity indices. In Section 7, we conclude with a brief summary of the main results.

2 So, what is the problem, then?

In order to describe the fundamental problem in the simplest possible terms, let us suppose that $A : D(A) \subset \mathcal{H} \to \mathcal{H}$ is a linear differential operator defined on its domain $D(A)$ contained in a certain Hilbert space $\mathcal{H}$. Given $f \in \mathcal{H}$, suppose that $u \in D(A)$ solves

$$Au = f\,.$$

In this simple discussion we shall assume that any (homogeneous) boundary condition that u has to satisfy is imposed through the definition of $D(A)$.

Now, let us seek a numerical approximation u_h to u by means of a Galerkin method. Thereby, we consider a sequence of finite-dimensional linear subspaces $\{\mathcal{H}^h\}$ of $D(A)$ parametrized by $h \in (0,1]$, and look for $u_h \in \mathcal{H}^h$ such that

$$P_h A u_h = P_h f\,, \tag{1}$$

where P_h is the orthogonal projector in $\mathcal{H}$ onto $\mathcal{H}^h$. On defining the *residual* $R(u_h) = f - Au_h$, (1) can be restated in the following equivalent form:

$$P_h R(u_h) = P_h(f - Au_h) = 0\,;$$

this identity is usually referred to as the *Galerkin orthogonality property*, as it expresses the fact that $R(u_h)$ is orthogonal to $\mathcal{H}^h$ in $\mathcal{H}$.

Suppose, for example, that Ω, the computational domain, is a bounded open set in $\mathbb{R}^n$, $\mathcal{H} = L_2(\Omega)$ is the set of all real-valued square-integrable functions defined on Ω, and let ζ be a given nonnegative weight function in $L_2(\Omega)$. [A summary of some standard notation and a collection of basic results from the theory of function spaces will be provided in the next section.] Then, with $(\cdot,\cdot)$ denoting the inner product in $L_2(\Omega)$, $v \mapsto J(v) = (v,\zeta)$ is a bounded linear functional on $L_2(\Omega)$, and $J(u)$ represents the weighted mean value of u over Ω.

A typical practical question one might encounter is the following: given a certain tolerance $\mathtt{TOL} > 0$, find $u_h \in \mathcal{H}^h$ such that

$$|J(u) - J(u_h)| \leq \mathtt{TOL}\,. \tag{2}$$

In the field of *a posteriori* error estimation a problem of this kind is approached by deriving a bound on $|J(u) - J(u_h)|$ in terms of the residual $R(u_h)$ of the form

$$|J(u) - J(u_h)| \leq \mathtt{Computable\ expression}(R(u_h))\,, \tag{3}$$

where the right-hand side is a computable expression involving available data and the residual, $R(u_h) = f - Au_h$, which can be calculated *after* u_h has been computed; hence the terminology *a posteriori error bound*. Once such an *a posteriori* error bound is available, the task of satisfying (2) is reduced to finding $\mathcal{H}^h$ and $u_h \in \mathcal{H}^h$ such that

$$\mathtt{Computable\ expression}(R(u_h)) \leq \mathtt{TOL}\,,$$

– an objective that is usually achieved through a feed-back process, by adaptive adjustment of $\mathcal{H}^h$. It is clear from this discussion that the fundamental theoretical problem, therefore, is to derive a bound of the kind (3).

A general approach to the derivation of *a posteriori* error bounds is based on the use of a *duality argument* involving the adjoint A^* of the linear operator A, defined on its domain $D(A^*) \subset \mathcal{H} = L_2(\Omega)$. For this purpose, we consider the *dual problem* of finding z in $D(A^*) \subset \mathcal{H}$ such that

$$A^* z = \zeta\,, \tag{4}$$

whose right-hand side is the given weight-function ζ. Thus the data for the dual problem is the density function (Riesz representer) of the linear functional of interest $J(\cdot)$. Hence,

$$\begin{aligned} J(u) - J(u_h) &= (u - u_h, \zeta) \\ &= (Au - Au_h, z) \\ &= (f - Au_h, z) \\ &\equiv (R(u_h), z)\,. \end{aligned} \tag{5}$$

As $P_hR(u_h) = 0$, it follows that

$$\begin{aligned} J(u) - J(u_h) &= (R(u_h), z) \\ &= (R(u_h), z - z_h) \qquad \forall z_h \in \mathcal{H}^h\,; \end{aligned} \tag{6}$$

this identity is referred to as the *error representation formula.* A possible choice of z_h may be P_hz, although there are other viable alternatives. For example, if (1) represents a finite element method, then z_h may be chosen as a finite element interpolant or quasi-interpolant of z. It should be borne in mind, however, that z is the analytical solution to the dual problem, and even though $R(u_h)$ will become 'computable' after u_h has been calculated, $z - z_h$ will not be available. Hence, the right-hand side of the error representation formula (6) is *not* computable.

One approach to obtaining a computable error bound from (6) (see, e.g., [14]) proceeds by attempting to eliminate $z - z_h$. Briefly, the argument is as follows. Suppose that $\mathcal{H}_s$ is a Hilbert subspace of $\mathcal{H}$ with norm $\|\cdot\|_{\mathcal{H}_s}$; here s is a positive index related to smoothness (for example, when $\mathcal{H} = L_2(\Omega)$, $\mathcal{H}_s$ might be the Sobolev space $H^s(\Omega) \subset L_2(\Omega)$ of index $s > 0$); assume that $z \in \mathcal{H}_s$ and that we have found a constant $C_{\text{stab}} = C_{\text{stab}}(\zeta)$ such that

$$\|z\|_{\mathcal{H}_s} \le C_{\text{stab}}\,.$$

Suppose also that there exists a positive constant C_{int} and $z_h \in \mathcal{H}^h$ such that

$$\|h^{-s}(z - z_h)\|_{\mathcal{H}} \le C_{\text{int}}\|z\|_{\mathcal{H}_s}\,. \tag{7}$$

Then, on applying the Cauchy-Schwarz inequality in $\mathcal{H}$, we get

$$\begin{aligned} |J(u) - J(u_h)| &\le \|h^sR(u_h)\|_{\mathcal{H}}\,\|h^{-s}(z - z_h)\|_{\mathcal{H}} \\ &\le C_{\text{int}}\,\|h^sR(u_h)\|_{\mathcal{H}}\,\|z\|_{\mathcal{H}_s} \\ &\le C_{\text{int}}\,C_{\text{stab}}\,\|h^sR(u_h)\|_{\mathcal{H}}\,. \end{aligned}$$

The constants C_{int} and C_{stab} are referred to as the *interpolation constant* and *strong stability constant*, respectively. Typical values of C_{int} are available from approximation theory, while C_{stab} has to be determined either analytically or computationally by studying the well-posedness of the dual problem (4) in the norm $\|\cdot\|_{\mathcal{H}_s}$. Thus we have established the following *a posteriori* error bound:

$$\boxed{|J(u) - J(u_h)| \le C_{\text{int}}C_{\text{stab}}\|h^sR(u_h)\|_{\mathcal{H}}} \qquad s \ge 0\,. \tag{8}$$

An error bound of this kind which does not directly involve the dual solution z will be referred to here as a Type II *a posteriori* error bound, cf. [35].

Let us pause for a moment at this point to reflect on some of the steps in our argument and examine the validity of the hypotheses we have made. Our first assumption concerned the existence of a Hilbert subspace $\mathcal{H}_s$ contained in $\mathcal{H}$ such that the dual solution z is contained in $\mathcal{H}_s$; in particular, when $\mathcal{H} = L_2(\Omega)$, it would be convenient to choose $\mathcal{H}_s$ as a Sobolev space $H^s(\Omega)$ of positive index $s > 0$ for it is in these spaces that approximation properties of the kind (7) are readily available. We shall consider two examples.

Example A. Suppose that $\Omega = (0,1)^2$, the unit square in the $x = (x_1, x_2)$ plane, and consider the following Dirichlet boundary value problem for Poisson's equation

$$\Delta u = 1 \quad \text{in } \Omega\,, \qquad u = 0 \quad \text{on } \partial\Omega\,.$$

Here the dual problem is

$$\Delta z = \zeta \quad \text{in } \Omega\,, \qquad z = 0 \quad \text{on } \partial\Omega.$$

As $\zeta \in L_2(\Omega)$, it follows from regularity theory for elliptic partial differential equations in polygonal domains that $z \in H^s(\Omega) \cap H_0^1(\Omega)$, $1 \leq s < 3$; thus we may select $\mathcal{H}_s = H^s(\Omega) \cap H_0^1(\Omega)$, $1 \leq s < 3$.

Example B. Suppose, again, that $\Omega = (0,1)^2$, the unit square in the $x = (x_1, x_2)$ plane, and consider the following boundary value problem for a scalar first-order hyperbolic equation for a function u of the variable $(x_1, x_2) \in \Omega$:

$$u_{x_1} = 1 \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \Gamma_-\,,$$

where $\Gamma_- = \{(0, x_2) \,:\, 0 \leq x_2 \leq 1\}$. This time, the dual problem is

$$-z_{x_1} = \zeta \quad \text{in } \Omega\,, \qquad z = 0 \quad \text{on } \Gamma_+ = \{(1, x_2) \,:\, 0 \leq x_2 \leq 1\}.$$

Unlike Example A, here z need not belong to $H^s(\Omega)$ for any $s > 0$. This is easily seen by choosing ζ as a function of x_2 only; say, $\zeta(x_1, x_2) = \varphi(x_2)$ where φ belongs to $L_2(0,1)$ but does not belong to $H^s(0,1)$ for any $s > 0$ (e.g., $\varphi(x_2) = 1/(|x_2 - (1/2)|^{1/2} \log |x_2 - (1/2)|)$ belongs to $L_2(0,1)$ but not to $L_p(0,1)$ for $p > 2$; by the Sobolev embedding theorem, φ cannot belong to any $H^s(0,1)$, $s > 0$, for otherwise it would be an element of $L_p(0,1)$ for some $p > 2$, which is not the case). While this is a slightly extreme example, it nevertheless highlights the stark difference between the regularity properties of solutions to elliptic and hyperbolic boundary value problems.

We see that while in Example A our hypothesis about the existence of a Hilbert subspace $\mathcal{H}_s$, $s > 0$, of $\mathcal{H} = L_2(\Omega)$ containing the dual solution z can be satisfied, for Example B this hypothesis has failed. Although our examples are trivial, the observed problem is generic, and is due to the fact that unlike elliptic partial differential operators, first-order hyperbolic operators have no inherent smoothing properties in isotropic Sobolev spaces.

In fact, our examples allow us to gain further insight into the difficulties. Let us select a point $\hat{x}$ in $\Omega = (0,1)^2$ and suppose that the aim of the computation is now to ensure that

$$|u(\hat{x}) - u_h(\hat{x})| \leq \mathtt{TOL}\,,$$

where u is the solution to the boundary value problem from either Example A or B.

Example A′. For the elliptic boundary value problem from Example A, the appropriate dual problem associated with $J(u) = u(\hat{x})$ is

$$\Delta z = \delta(x - \hat{x}) \qquad \text{for } x \in \Omega\,, \qquad z = 0 \quad \text{on } \partial\Omega\,;$$

i.e., z is the classical Green's function. It is well known that z is a smooth function in $\Omega \setminus \{\hat{x}\}$ and it exhibits a logarithmic singularity near $\hat{x}$. A simple calculation reveals that $z \in H^s(\Omega)$ for $s \in [0, 1)$.

Example B′. For the hyperbolic boundary value problem from Example B, the appropriate dual problem associated with $J(u) = u(\hat{x})$ is now

$$-z_{x_1} = \delta(x - \hat{x}) \qquad \text{for } x \in \Omega\,, \qquad z = 0 \quad \text{on } \Gamma_+.$$

Hence,

$$z(x_1, x_2) = (1 - H(x_1 - \hat{x}_1))\,\delta(x_2 - \hat{x}_2)\,.$$

Clearly, z is not a locally integrable function but a measure supported on the line-segment $x_2 = \hat{x}_2$, $0 \leq x_1 \leq \hat{x}_1$; in particular, z does not belong to any Sobolev space on Ω of index $s \geq 0$. Incidentally, had we chosen to consider the more general scalar linear first-order hyperbolic differential equation $a_1 u_{x_1} + a_2 u_{x_2} = 1$, where $a_1 > 0$ and $a_2 \geq 0$ are smooth functions of $x = (x_1, x_2)$, a similar conclusion would have resulted except that z would have been concentrated on a piece of the characteristic curve that passes through the point $(\hat{x}_1, \hat{x}_2)$. Thus, while the Green's function for an elliptic boundary value problem on Ω has singularity at a single point within Ω and decays away from that point, for a scalar linear hyperbolic boundary value problem the Green's function (or, *Green's measure*, to be more precise) has nonlocal singularity concentrated along a characteristic and does not, in general, decay along this characteristic, cf. Fig. 1, for example, for a sketch of z.

Remark 1. Strictly speaking, Examples A′ and B′ do not fit into the general theoretical framework we had presented given that $\zeta = \delta(\cdot - \hat{x})$ is not an element of $\mathcal{H} = L_2(\Omega)$, so (4) is meaningless for this choice of ζ. However, the effects which we observed in Examples A′ and B′ can be emulated within the theory presented above by approximating, for $\varepsilon > 0$ small, $J(u) = u(\hat{x})$ by $J_\varepsilon(u) = (u, \zeta_\varepsilon(\cdot - \hat{x}))$ and $\delta(\cdot - \hat{x})$ with $\zeta_\varepsilon(\cdot - \hat{x})$, where $\zeta_\varepsilon(x) = \varepsilon^{-2}\zeta(x/\varepsilon)$ and ζ is a nonnegative function whose support is contained in the unit circle centered at the origin and such that the integral of ζ over its support is equal to 1.

If now, with the observations from Examples A′ and B′ in mind, we return to (5) we conclude, at least formally[1], that for an elliptic operator the size

[1] The argument can be made rigorous by local mollification of $\delta(\cdot - \hat{x})$ as indicated in the previous Remark.

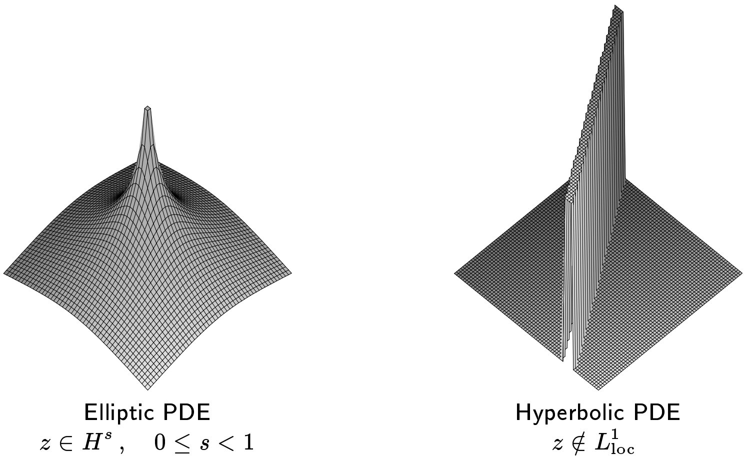


Fig. 1. Green's function for an elliptic (left) and a hyperbolic problem (right)

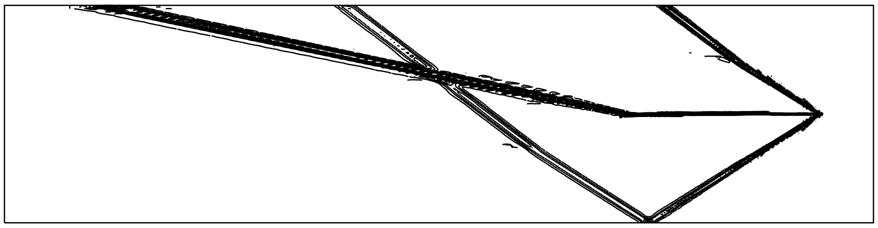


Fig. 2. Green's function for the linearized adjoint compressible Euler equations of gas dynamics

of the error $(u - u_h)$ at $\hat{x}$ will be strongly dependent on the size of $R(u_h)$ in the neighborhood of the point $\hat{x}$ and, given that z decays away from $\hat{x}$, values of $R(u_h)$ at more distant points will have less influence on $(u-u_h)(\hat{x})$. On the other hand, for a hyperbolic operator, the singularity of z is nonlocal and z does not, in general, decay along the characteristic curve on which the singularity is concentrated, so the size of $(u - u_h)$ at $\hat{x}$ can be strongly influenced by values of $R(u_h)$ at points that are quite distant from $\hat{x}$. This is further demonstrated in Fig. 2 where we show the Green's function for the linearized adjoint compressible Euler equations of gas dynamics in the case of a shock reflecting off a solid wall, cf. Kröner [39], p. 401.

In order to give an experimental illustration of the issue raised in the last paragraph, we note that, letting $e = u - u_h$ denote the *global error*, the residual and the error are related via

$$Ae = R(u_h)\,.$$

Now suppose that A is a scalar first-order linear hyperbolic operator. Following [29], we decompose the error $e_\kappa = e|_\kappa$ on an open subset κ of the computational domain Ω as

$$e_\kappa = e_\kappa^{\mathrm{trans}} + e_\kappa^{\mathrm{cell}}.$$

We shall refer to e_κ^{cell} as the *cell error* and $e_\kappa^{\mathrm{trans}}$ will be called the *transmitted error*; they are defined as the respective solutions to the following problems:

$$\begin{aligned} A e_\kappa^{\mathrm{cell}} &= R(u_h) \quad \text{on } \kappa; & e_\kappa^{\mathrm{cell}} &= 0 \quad \text{on the inflow boundary of } \kappa\,, \\ A e_\kappa^{\mathrm{trans}} &= 0 \quad \text{on } \kappa; & e_\kappa^{\mathrm{trans}} &= e_\kappa \quad \text{on the inflow boundary of } \kappa\,. \end{aligned}$$

The cell error e_κ^{cell} is the part of the error $e_\kappa = e|_\kappa$ that is governed by the residual $R(u_h)|_\kappa$ and is not influenced by pollution effects from outside κ. On the other hand, the transmitted error $e_\kappa^{\mathrm{trans}}$ is independent of the residual on κ and is driven completely by the error that has been created outside κ and transmitted into κ through the inflow boundary of κ.

Let us consider the model scalar hyperbolic problem $a_1 u_{x_1} + a_2 u_{x_2} = 0$, with $\mathbf{a} = (a_1, a_2) = (1/10 + \sin(\pi x_2), 2)$, on the unit square $\Omega = (0,1)^2$ in the (x_1, x_2)–plane, with inflow boundary condition

$$u(x_1, 0) = \begin{cases} \mathrm{e}^{-300(0.4 - x_1)^4}\,, & \text{for } 0.0 \le x_1 \le 0.4, \\ 1\,, & \text{for } 0.4 \le x_1 \le 0.6, \\ 0\,, & \text{for } 0.6 < x_1 \le 1.0, \end{cases}$$

with the compatible boundary condition $u(0, x_2) = \mathrm{e}^{-300(0.4)^4}\,(= \mathrm{e}^{-7.68})$ along $x_1 = 0$, $0 \le x_2 \le 1$. This hyperbolic problem has been solved numerically using the streamline-diffusion finite element method on a sequence of uniform meshes consisting of triangular elements κ. The streamline-diffusion parameter was selected according to the standard formula: $\delta|_\kappa = 0.5h/|\bar{\mathbf{a}}_\kappa|$, where $\bar{\mathbf{a}}_\kappa$ denotes the average of $\mathbf{a}$ on element κ and $|\bar{\mathbf{a}}_\kappa|$ is the Euclidean norm of $\bar{\mathbf{a}}_\kappa$. Since the exact solution u can be calculated using the method of characteristics, we may compute $\|e_\kappa\|_{L_2(\kappa)}$, $\|e_\kappa^{\mathrm{cell}}\|_{L_2(\kappa)}$ and $\|e_\kappa^{\mathrm{trans}}\|_{L_2(\kappa)}$ on each triangle κ in the mesh, thus obtaining three sets of data. Our aim is now to compute the coefficients of correlation[2] $\rho(e, e^{trans})$, $\rho(e, e^{cell})$, $\rho(e, R(u_h))$ between the data set $\{\|e_\kappa\|_{L_2(\kappa)}\}_\kappa$ and each of the three data sets $\{\|e_\kappa^{\mathrm{trans}}\|_{L_2(\kappa)}\}_\kappa$, $\{\|e_\kappa^{\mathrm{cell}}\|_{L_2(\kappa)}\}_\kappa$ and $\{\|R(u_h)|_\kappa\|_{L_2(\kappa)}\}_\kappa$, respectively.

[2]
$$\rho(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{cov}(X, X)}\,\sqrt{\mathrm{cov}(Y, Y)}}\,,$$

where the covariance of the data sets X and Y is defined by

$$\mathrm{cov}(X, Y) = \mathrm{E}([X - \mathrm{E}(X)]\,[Y - \mathrm{E}(Y)]);$$

here $\mathrm{E}(X)$ signifies the arithmetic mean of the data set X (i.e., the expectation of X, assuming that X is uniformly distributed).

Table 1. Correlation coefficients under mesh refinement

Mesh	Elements	$\rho(e, e^{\mathrm{trans}})$	$\rho(e, e^{\mathrm{cell}})$	$\rho(e, R(u_h))$
17×17	512	0.99	0.31	0.34
33×33	2048	1.00	0.42	0.43
65×65	8192	1.00	0.44	0.44
129×129	32768	1.00	0.45	0.45
257×257	131072	1.00	0.44	0.44

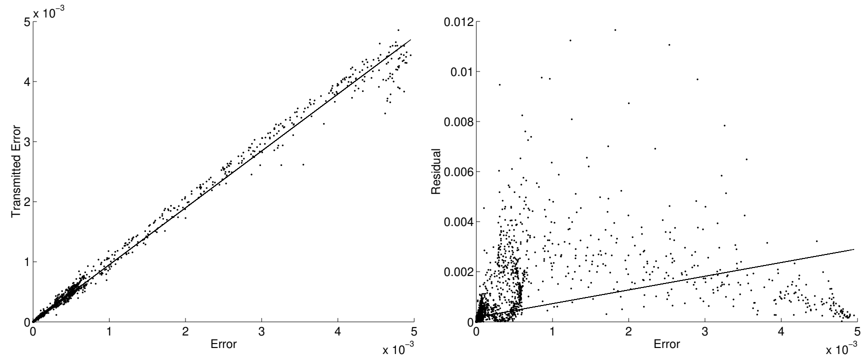


Fig. 3. Scatter plots for the data-set-pairs $(\|e_\kappa\|_{L_2(\kappa)}, \|e_\kappa^{\mathrm{trans}}\|_{L_2(\kappa)})_\kappa$ (left) and $(\|e_\kappa\|_{L_2(\kappa)}, \|R(u_h)|_\kappa\|_{L_2(\kappa)})_\kappa$ (right). Each dot in the left scatter plot corresponds to a triangle κ in the mesh. In the plot on the left, the first co-ordinate of a point is $\|e_\kappa\|_{L_2(\kappa)}$ and the second co-ordinate is $\|e_\kappa^{\mathrm{trans}}\|_{L_2(\kappa)}$; similarly, in the right-hand figure, the first co-ordinate of a point in the scatter plot is $\|e_\kappa\|_{L_2(\kappa)}$ and the second co-ordinate is $\|R(u_h)|_\kappa\|_{L_2(\kappa)}$

In Table 1 we display the correlation coefficients $\rho(e, e^{trans})$, $\rho(e, e^{cell})$ and $\rho(e, R(u_h))$ for these data sets, on a sequence of uniform structured triangular meshes: in each case the mesh is constructed from a uniform $N \times N$ mesh by connecting the bottom–left corner of each mesh square with its top–right corner. We observe that there is a very strong correlation between the error and the transmitted error, while there is only very weak correlation between the error and the cell error, as well as between the error and the residual.

This is further confirmed in the scatter diagrams presented in Fig. 3; here we plot each of $\|e_\kappa^{\mathrm{trans}}\|_{L_2(\kappa)}$ and $\|R(u_h)|_\kappa\|_{L_2(\kappa)}$ against the error $\|e_\kappa\|_{L_2(\kappa)}$ as κ ranges through the triangulation of the computational domain, together with the linear regression line, on the third mesh with 8192 elements. In particular, we see from the right-hand figure that there is a cluster of points lying on the horizontal axis; i.e., as the error increases, the residual remains small in magnitude. Clearly, this has serious implications in the context of mesh design; an adaptive algorithm based on refining elements κ where the *local* residual is large may fail to refine all the elements where $\|e_\kappa\|_{L_2(\kappa)}$ is large, hence leading to poor resolution of localized structures in the solution.

One can draw a number conclusions from this discussion. First, the dual solution measures the sensitivity of the error in the quantity of interest to local perturbations in the residual. Second, the smoothness and decay properties of the dual solution in a hyperbolic problem may be completely different than in an elliptic problem; indeed, the lack of decay of the dual solution to a hyperbolic problem induces nonlocal dependence between the error and the residual; hence the local size of the residual may give little or no information about the local size of the error. From the point of view of sharpness of the *a posteriori* error bound, it therefore appears beneficial to retain the dual solution in the bound as a local weight function. In Sections 5 and 6, we shall revisit the error representation formula (6) that gave rise to the Type II *a posteriori* error bound (8), and attempt to derive a new bound which still includes z (or, rather, a numerical approximation to z); such bounds will be referred to henceforth as Type I *a posteriori* error bounds, cf. [35]. Before doing so, however, we shall make a brief diversion in order to introduce some basic results from function space theory and formulate the mathematical models of fluid dynamics whose discretizations we shall consider.

3 Elements of function space theory

The aim of this section is to provide a brief overview of some elementary results from the theory of function spaces and to introduce the notation which will be used throughout. For proofs and further technical details on classical function spaces the reader is referred to the monograph of Adams [1], for example.

Spaces of continuous functions. Let $\mathbb{N}$ denote the set of all nonnegative integers. An n–tuple $\alpha = (\alpha_1, \ldots, \alpha_n)$ in $\mathbb{N}^n$ will be referred to as a *multi-index*; the nonnegative integer $|\alpha| = |\alpha_1| + \ldots + |\alpha_n|$ is called the length of the multi-index α. We define $\partial^\alpha = \partial_1^{\alpha_1} \ldots \partial_n^{\alpha_n}$, where $\partial_j = \partial/\partial x_j$ for $j = 1, \ldots, n$.

Suppose that Ω is an open set in $\mathbb{R}^n$. For $k \in \mathbb{N}$, we denote by $C^k(\Omega)$ the set of all continuous real-valued functions u defined on Ω such that $\partial^\alpha u$ is continuous on Ω for every multi-index α, $|\alpha| \leq k$. When $k = 0$, we shall write $C(\Omega)$ in place of $C^0(\Omega)$. For $k = \infty$, $C^\infty(\Omega)$ will denote the intersection $\bigcap_{k\geq 0} C^k(\Omega)$.

We shall also require spaces of functions defined over the closure $\bar{\Omega}$ of an open set $\Omega \subset \mathbb{R}^n$. For $k \in \mathbb{N}$, $C^k(\bar{\Omega})$ will signify the set of all $u \in C^k(\Omega)$ such that $\partial^\alpha u$ can be continuously extended from Ω onto $\bar{\Omega}$ for every multi-index α, $|\alpha| \leq k$. Further, we define $C^\infty(\bar{\Omega})$ as the intersection $\bigcap_{k\geq 0} C^k(\bar{\Omega})$. The notation $C^0(\bar{\Omega})$ is abbreviated to $C(\bar{\Omega})$.

Assuming that Ω is a *bounded* open set in $\mathbb{R}^n$ and $k \in \mathbb{N}$, the linear space $C^k(\bar{\Omega})$ is a Banach space equipped with the norm

$$\|u\|_{C^k(\bar{\Omega})} = \max_{|\alpha| \le k} \sup_{x \in \Omega} |\partial^\alpha u(x)| \,.$$

For $k \in \mathbb{N}$ we denote by $C^{k,1}(\bar{\Omega})$ the set of all $u \in C^k(\bar{\Omega})$ such that the quantity

$$|u|_{C^{k,1}(\bar{\Omega})} = \max_{|\alpha| = k} \sup_{x \neq y,\ x,y \in \Omega} \frac{|\partial^\alpha u(x) - \partial^\alpha u(y)|}{|x-y|}$$

is finite. $C^{k,1}(\bar{\Omega})$ is a Banach space with the norm

$$\|u\|_{C^{k,1}(\bar{\Omega})} = \|u\|_{C^k(\bar{\Omega})} + |u|_{C^{k,1}(\bar{\Omega})} \,.$$

Clearly, $C^{k+1}(\bar{\Omega}) \subset C^{k,1}(\bar{\Omega})$. When u belongs to $C^{0,1}(\bar{\Omega})$, it is said to be *Lipschitz continuous* on $\bar{\Omega}$.

The *support*, $\operatorname{supp} u$, of a continuous function u defined on an open set Ω is the closure in Ω of the set $\{x \in \Omega \ : \ u(x) \neq 0\}$; in other words, $\operatorname{supp} u$ is the smallest closed subset of Ω such that $u = 0$ on $\Omega \setminus \operatorname{supp} u$. For $k = 0, 1, \ldots, \infty$, $C_0^k(\Omega)$ denotes the set of all $u \in C^k(\Omega)$ whose support is a bounded (and, by definition, closed) subset of Ω.

Spaces of integrable functions. For $p \ge 1$ and an open set $\Omega \subset \mathbb{R}^n$, $L_p(\Omega)$ will denote the set of all real-valued Lebesgue measurable functions u defined on Ω such that $|u|^p$ is integrable on Ω with respect to the Lebesgue measure $\mathrm{d}x = \mathrm{d}x_1 \ldots \mathrm{d}x_n$; it is implicitly assumed that any two functions which are equal almost everywhere (i.e., equal, except maybe on a set of zero Lebesgue measure) are identified. $L_p(\Omega)$ is a Banach space equipped with the norm

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

When $p = 2$, $L_2(\Omega)$ is a Hilbert space with the inner product

$$(u, v) = \int_\Omega u(x)\, v(x) \ \mathrm{d}x \,.$$

$L_\infty(\Omega)$ denotes the set of all real-valued Lebesgue measurable functions u defined on Ω such that $|u|$ has finite essential supremum; the essential supremum of $|u|$ is defined as the infimum of the set of all positive real numbers M such that $|u| \le M$ almost everywhere on Ω. Again, any two functions that are equal almost everywhere on Ω are identified. $L_\infty(\Omega)$ is a Banach space with norm

$$\|u\|_{L_\infty(\Omega)} = \text{ess.sup}_{x \in \Omega} |u(x)| \,.$$

Hölder's Inequality. Let $u \in L_p(\Omega)$ and $v \in L_q(\Omega)$, where $1/p + 1/q = 1$, $1 \le p, q \le \infty$. Then $uv \in L_1(\Omega)$ and

$$\left| \int_\Omega u(x)\, v(x) \ \mathrm{d}x \right| \le \|u\|_{L_p(\Omega)} \|v\|_{L_q(\Omega)} \,.$$

In the special case when $p = q = 2$, this inequality is referred to as the *Cauchy-Schwarz Inequality.*

Sobolev spaces. Given that Ω is an open set in $\mathbb{R}^n$, k a non-negative integer and $1 \le p \le \infty$, we define the *Sobolev space*

$$W_p^k(\Omega) = \{u \in L_p(\Omega) \,:\, \partial^\alpha u \in L_p(\Omega)\,, \quad |\alpha| \le k\}\,,$$

and equip it with the *Sobolev norm* defined by

$$\|u\|_{W_p^k(\Omega)} = \begin{cases} \left(\sum_{|\alpha|\le k} \|\partial^\alpha u\|^p_{L_p(\Omega)}\right)^{1/p}, & \text{if } 1 \le p < \infty\,, \\ \|u\|_{W^k_\infty(\Omega)} = \max_{|\alpha|\le k} \|\partial^\alpha u\|_{L_\infty(\Omega)}\,, & \text{if } p = \infty\,. \end{cases}$$

The associated *Sobolev seminorm* is defined by

$$|u|_{W_p^k(\Omega)} = \begin{cases} \left(\sum_{|\alpha|= k} \|\partial^\alpha u\|^p_{L_p(\Omega)}\right)^{1/p}, & \text{if } 1 \le p < \infty\,, \\ \max_{|\alpha|= k} \|\partial^\alpha u\|_{L_\infty(\Omega)}\,, & \text{if } p = \infty\,. \end{cases}$$

In these definitions the derivatives are to be understood in the sense of distributions. The Sobolev space $W_p^k(\Omega)$ is a Banach space with the norm $\|\cdot\|_{W_p^k(\Omega)}$, $1 \le p \le \infty$, $k \ge 0$. Specifically, for $p = 2$, the normed linear space $W_2^k(\Omega)$ is a Hilbert space with the inner product

$$(u,v)_{W_2^k(\Omega)} = \sum_{|\alpha|\le k} (\partial^\alpha u, \partial^\alpha v)\,,$$

where $(\cdot,\cdot)$ denotes the inner product in $L_2(\Omega)$.

Finer smoothness properties of integrable functions can be detected by considering fractional-order Sobolev spaces. Given that s is a positive real number, $s \notin \mathbb{N}$, let us write $s = m + \sigma$, where $0 < \sigma < 1$ and $m = [s]$ is the integer part of s. The fractional-order Sobolev space $W_p^s(\Omega)$, $1 \le p < \infty$, is the set of all $u \in W_p^m(\Omega)$ such that

$$|u|_{W_p^s(\Omega)} = \left\{ \sum_{|\alpha|=m} \int_\Omega \int_\Omega \frac{|D^\alpha u(x) - D^\alpha u(y)|^p}{|x-y|^{n+\sigma p}} \,\mathrm{d}x\,\mathrm{d}y \right\}^{1/p} < \infty\,,$$

with the usual modification when $p = \infty$. The fractional-order Sobolev norm of index s is defined by

$$\|u\|_{W_p^s(\Omega)} = \begin{cases} \left\{\|u\|^p_{W_p^m(\Omega)} + |u|^p_{W_p^s(\Omega)}\right\}^{1/p}, & \text{if } 1 \le p < \infty\,, \\ \|u\|_{W^m_\infty(\Omega)} + |u|_{W^s_\infty(\Omega)}\,, & \text{if } p = \infty\,. \end{cases}$$

The fractional-order Sobolev space $W_p^s(\Omega)$ is a Banach space with this norm.

The function spaces defined above will be helpful for characterizing the smoothness properties of a solution to a partial differential equation. In physical applications, however, a PDE is rarely considered in isolation, but is coupled with side–conditions which typically take the form of boundary conditions. The PDE in tandem with boundary conditions is referred to as a *boundary value problem*. In order to guarantee that the boundary condition is mathematically meaningful, it is necessary to ensure that the boundary is sufficiently smooth (in the sense of the following definition, for example).

Definition 1. Suppose that Ω is an open set in $\mathbb{R}^n$. The boundary $\partial\Omega$ of Ω is said to be Lipschitz continuous if, for every $x \in \partial\Omega$, there is an open set $\mathcal{O} \subset \mathbb{R}^n$ with $x \in \mathcal{O}$ and a local orthogonal coordinate system with coordinate $\zeta = (\zeta_1, \ldots, \zeta_n) \equiv (\zeta', \zeta_n)$ and $a \in \mathbb{R}^n$, such that

$$\mathcal{O} = \{\zeta \,:\, -a_j < \zeta_j < a_j, \quad 1 \le j \le n\},$$

and there is a Lipschitz continuous function φ defined on

$$\mathcal{O}' = \{\zeta' \in \mathbb{R}^{n-1} \,:\, -a_j < \zeta_j < a_j, \quad 1 \le j \le n-1\},$$

with

$$|\varphi(\zeta')| \le a_n/2, \qquad \zeta' \in \mathcal{O}',$$

$$\Omega \cap \mathcal{O} = \{\zeta : \zeta_n < \varphi(\zeta'),\ \ \zeta' \in \mathcal{O}'\} \text{ and } \partial\Omega \cap \mathcal{O} = \{\zeta : \zeta_n = \varphi(\zeta'),\ \ \zeta' \in \mathcal{O}'\}.$$

A bounded open set with a Lipschitz continuous boundary is called a *Lipschitz domain.* [Analogously, if φ above belongs to $C^\infty(\mathcal{O}')$ for each $\mathcal{O}'$ then we say that Ω has a C^∞ boundary. A bounded open set with C^∞ boundary is called a C^∞ *domain.*]

An important property of a Lipschitz domain Ω is that the unit outward normal vector to $\partial\Omega$ is defined almost everywhere with respect to the $(n-1)$–dimensional measure on $\partial\Omega$; thus, for example, one may meaningfully talk about the normal flux of a physical quantity across $\partial\Omega$. A simple example of a Lipschitz domain is a bounded open polyhedron in $\mathbb{R}^n$, $n \ge 2$.

Having characterized the smoothness of the boundary, next we need to make sure that the boundary data are sufficiently smooth so as to be meaningful. In order to do so, we give a brief description of Sobolev spaces on the boundary $\partial\Omega$ of a Lipschitz domain Ω. Let us first recall from Definition 1 that for every x on $\partial\Omega$ there exists a Lipschitz continuous function $\varphi : \mathcal{O}' \subset \mathbb{R}^{n-1} \to \mathbb{R}$ such that, using the notation from Definition 1,

$$\partial\Omega \cap \mathcal{O} = \{\zeta = (\zeta', \varphi(\zeta')) \,:\, \zeta' \in \mathcal{O}'\},$$

so that, locally, $\partial\Omega$ is an $(n-1)$–dimensional hypersurface in $\mathbb{R}^n$. We define the mapping ϕ by

$$\phi(\zeta') = (\zeta', \varphi(\zeta')).$$

Then ϕ^{-1} exists and it is Lipschitz continuous on $\phi(\mathcal{O}')$, which leads us to the following definition.

Definition 2. Let Ω be a Lipschitz domain in $\mathbb{R}^n$. For $0 \le s \le 1$ and $1 \le p < \infty$ we denote by $W_p^s(\partial\Omega)$ the set of all $u \in L_p(\partial\Omega)$ such that the composition $u \circ \phi$ belongs to $W_p^s(\mathcal{O}' \cap \phi^{-1}(\partial\Omega \cap \mathcal{O}))$ for all possible $\mathcal{O}'$ and φ satisfying the conditions of Definition 1, where $\phi(\zeta') = (\zeta', \varphi(\zeta'))$ for $\zeta' \in \mathcal{O}'$.

Next, we shall equip $W_p^s(\partial\Omega)$ with a norm; for this purpose we consider any *atlas* $(\mathcal{O}_j, \varphi_j)_{j=1}^J$ for $\partial\Omega$ such that $\mathcal{O}_j$ and φ_j, $j = 1, \dots, J$, satisfy the conditions of Definition 1. We define $\|\cdot\|_{W_p^s(\partial\Omega)}$ by

$$\|u\|_{W_p^s(\partial\Omega)} = \left(\sum_{j=1}^J \|u \circ \phi_j\|^p_{W_p^s(\mathcal{O}_j' \cap \phi_j^{-1}(\partial\Omega \cap \mathcal{O}_j))} \right)^{1/p},$$

where $\phi_j(\zeta') = (\zeta', \varphi_j(\zeta'))$ for $\zeta' \in \mathcal{O}_j'$, $j = 1, \dots, J$. The defining expression for $\|\cdot\|_{W_p^s(\partial\Omega)}$ is not particularly convenient from the practical point of view as it explicitly involves the atlas $(\mathcal{O}_j, \varphi_j)_{j=1}^J$. Fortunately, for $0 < s < 1$ the norm $\|\cdot\|_{W_p^s(\partial\Omega)}$ defined above is equivalent to the following norm

$$\|u\|_{W_p^s(\partial\Omega)} = \left(\int_{\partial\Omega} |u|^p \, \mathrm{d}\sigma + \int_{\partial\Omega} \int_{\partial\Omega} \frac{|u(x) - u(y)|^p}{|x - y|^{n-1+sp}} \, \mathrm{d}\sigma(x) \, \mathrm{d}\sigma(y) \right)^{1/p},$$

where $\mathrm{d}\sigma$ denotes the $(n-1)$–dimensional surface measure on $\partial\Omega$. If Ω has a C^∞ boundary then the restriction $s \le 1$ is redundant; in fact, if $\Gamma_0 \subset \partial\Omega$ and Γ_0 is part of an $(n-1)$–dimensional C^∞ hypersurface in $\mathbb{R}^n$, then $\|u\|_{W_p^s(\Gamma_0)}$ can be defined, analogously as above, for all $s \ge 0$ and $1 \le p < \infty$.

Proposition 1. *Suppose that Ω is a Lipschitz domain contained in $\mathbb{R}^n$ and let $1 \le p < \infty$. Then $C^\infty(\bar{\Omega})$ is dense in $W_p^s(\Omega)$ for $s \ge 0$; in other words, the closure of $C^\infty(\bar{\Omega})$ in the norm of $W_p^s(\Omega)$ is equal to $W_p^s(\Omega)$.*

We note that while $C^\infty(\bar{\Omega})$ is dense in $W_p^s(\Omega)$ for $s \ge 0$ and $1 \le p < \infty$, $C_0^\infty(\Omega)$ is not dense in $W_p^s(\Omega)$ for $s > 1/p$; i.e., for $s > 1/p$, the closure of $C_0^\infty(\Omega)$ in the norm of $W_p^s(\Omega)$ is a proper closed linear subspace of $W_p^s(\Omega)$.

Now we are ready to discuss the notion of trace of a function on the boundary $\partial\Omega$ of a Lipschitz domain $\Omega \subset \mathbb{R}^n$. If ψ belongs to $C^\infty(\bar{\Omega})$ then we put

$$\gamma_0(\psi) = \psi|_{\partial\Omega} \,. \tag{9}$$

The *trace* of a function u in $W_p^s(\Omega)$ with $s > 1/p$, $1 < p < \infty$, is then defined by extending the operator γ_0 from the dense subspace $C^\infty(\bar{\Omega})$ to the whole of $W_p^s(\Omega)$.

Proposition 2. *Suppose that Ω is a Lipschitz domain in $\mathbb{R}^n$, and let $1 < p < \infty$. Assuming that $1/p < s \le 1$, the mapping γ_0 defined on $C^\infty(\bar{\Omega})$ by (9) has a unique continuous extension to a linear operator, called the trace* *operator, still denoted* γ_0*, from* $W_p^s(\Omega)$ *onto* $W_p^{s-(1/p)}(\partial\Omega)$*. If* Γ_0 *is an open subset of* $\partial\Omega$ *and* Γ_0 *is part of an* $(n-1)$*–dimensional* C^∞ *hypersurface in* $\mathbb{R}^n$*, then* $\gamma_0 : W_p^s(\Omega) \to W_p^{s-(1/p)}(\Gamma_0)$ *for* $1 < p < \infty$ *and all* $s > 1/p$ *(the restriction* $s \leq 1$ *being redundant).*

When $p = 2$ we shall write H^s in place of W_2^s to signify the fact that we are dealing with a Hilbert space. We denote by $H_0^s(\Omega)$ the closure of $C_0^\infty(\Omega)$ in the norm of $H^s(\Omega)$; when Ω is a Lipschitz domain and $1/2 < s < 3/2$, this space coincides with the set of all those functions in $H^s(\Omega)$ whose trace on $\partial\Omega$ is equal to zero.

4 Mathematical models of compressible fluid dynamics

The purpose of this section is to give a brief overview of mathematical models of compressible fluid dynamics and to list some simple model hyperbolic problems derived from these equations whose numerical approximations will be considered here. The computational tools developed in this work, however, are by no means limited to problems in compressible fluid dynamics, and the partial differential equations listed in this section should only be thought of as typical continuum models to which these computational tools apply.

4.1 Compressible Navier–Stokes equations

The motion of a compressible fluid occupying an open set $\Omega \subset \mathbb{R}^n$, $1 \leq n \leq 3$, is governed by the *compressible Navier–Stokes equations.* The equations express key physical properties, such as the conservation of mass, momentum and energy in the flow, and involve the density $\rho : \Omega \to \mathbb{R}$, the velocity field $\mathbf{u} : \Omega \to \mathbb{R}^n$, and the total energy per unit mass $e : \Omega \to \mathbb{R}$.

In Eulerian co-ordinates, the compressible Navier–Stokes equations have the following form:

Conservation of mass

$$\frac{\partial \rho}{\partial t} + \sum_{j=1}^{n} \frac{\partial}{\partial x_j}(\rho u_j) = 0\,;$$

Conservation of momentum

$$\frac{\partial}{\partial t}(\rho u_i) + \sum_{j=1}^{n} \frac{\partial}{\partial x_j}(\rho u_i u_j + p\delta_{ij}) = \sum_{j=1}^{n} \frac{\partial}{\partial x_j}\tau_{ij} + \hat{S}_i\,, \qquad \text{for } i = 1, \dots, n\,;$$

Conservation of energy

$$\frac{\partial}{\partial t}(\rho e) + \sum_{j=1}^{n} \frac{\partial}{\partial x_j}\left((\rho e + p)u_j\right) = \sum_{j=1}^{n} \frac{\partial}{\partial x_j}\left(k\frac{\partial T}{\partial x_j} + \sum_{\ell=1}^{n} \tau_{jl} u_l\right).$$

In these equations, $k > 0$ denotes the thermal diffusivity, τ is the stress tensor, p is the pressure, $\hat{\mathbf{S}} \in [L_2(\Omega)]^n$ is a vector of given sources, δ_{ij} is the Kronecker delta, and $T > 0$ is the temperature.

4.2 Compressible Euler equations

If viscous effects are negligible, the compressible Navier–Stokes equations can be simplified. For example, in the case of an inviscid compressible ideal gas, they reduce to:

$$\frac{\partial \mathbf{U}}{\partial t} + \sum_{j=1}^{n} \frac{\partial}{\partial x_j} \mathbf{F}_j(\mathbf{U}) = \mathbf{S}\,, \tag{10}$$

where

$$\mathbf{U} = [\rho, \rho u_1, \dots, \rho u_n, \rho e]^{\mathrm{T}}$$

is the vector of *conserved variables*,

$$\mathbf{F}_j = [\rho u_j, \rho u_1 u_j + \delta_{1j} p, \dots, \rho u_n u_j + \delta_{nj} p, (\rho e + p) u_j]^{\mathrm{T}}\,, \qquad j = 1, \dots, n\,,$$

are the fluxes, and

$$\mathbf{S} = \left[0, \hat{S}_1, \dots, \hat{S}_n, 0\right]^{\mathrm{T}} \in [L_2(\Omega)]^{n+2}$$

is a source vector. The range of the conserved variables is a subset $\Sigma \subset \mathbb{R}^{n+2}$ called the *state space*. For an ideal gas, density and pressure are related through the *equation of state*

$$p = (\kappa - 1)\rho \left(e - \frac{1}{2}|\mathbf{u}|^2\right)\,,$$

where $\mathbf{u} = (u_1, \dots, u_n)^{\mathrm{T}}$ and κ is the ratio of specific heats; for dry air, $\kappa = 1.405$. The nonlinear partial differential equations (10) are called the *compressible Euler equations.*

If $\mathbf{U}$ is sufficiently smooth, say $\mathbf{U} \in [C^1(\Omega)]^{n+2}$, then, by applying the chain rule, (10) can be restated in the equivalent quasilinear (nonconservative) form:

$$\frac{\partial \mathbf{U}}{\partial t} + \sum_{j=1}^{n} A_j(\mathbf{U}) \frac{\partial \mathbf{U}}{\partial x_j} = \mathbf{S}\,,$$

where $A_j(\mathbf{U}) = \nabla_{\mathbf{U}} \mathbf{F}_j(\mathbf{U}) \in \mathbb{R}^{(n+2)\times(n+2)}$ denotes the Jacobi matrix associated with the flux vector $\mathbf{F}_j$, $j = 1, \dots, n$. The system of compressible Euler equations is *hyperbolic*, i.e., the matrix

$$\sum_{j=1}^{n} \mu_j A_j(\mathbf{U}) \in \mathbb{R}^{(n+2)\times(n+2)}$$

possesses $n+2$ real eigenvalues λ_i, $i = 1, \dots, n+2$, and a complete set of linearly independent eigenvectors for all values of $\mathbf{U} \in \Sigma$ and for all $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^{\mathrm{T}} \in \mathbb{R}^n$.

4.3 Symmetrization of the compressible Euler equations

Symmetric hyperbolic systems frequently arise in mathematical models of continuum mechanics; later on in these notes we shall consider their finite element approximation in more detail. Here, as a motivating example, we show how the compressible Euler equations can be restated as a symmetric system of partial differential equations using entropy variables. Our exposition of the symmetrization process follows [52] and is based on the work of Hughes, Franca and Mallet [36] and Mock [43]; see also the paper of Barth [6] where symmetric forms of conservation laws have been extensively used in numerical computations. For the sake of simplicity and brevity, we shall confine our presentation to the case of two space dimensions ($n = 2$).

The starting point of the symmetrization process is to hypothesize the existence of a scalar *entropy function*

$$\eta \,:\, \mathbf{U} \in \Sigma \mapsto \eta(\mathbf{U}) \in \mathbb{R}$$

and the existence of the *entropy fluxes*

$$q_i \,:\, \mathbf{U} \in \Sigma \to q_i(\mathbf{U}) \in \mathbb{R}\,, \qquad i = 1, 2\,,$$

which are compatible with η in the sense that

$$\nabla_{\mathbf{U}}\eta(\mathbf{U}) \cdot A_i(\mathbf{U}) = \nabla_U q_i(\mathbf{U})\,, \qquad i = 1, 2\,.$$

The entropy η and the entropy fluxes q_i, $i = 1, 2$, are assumed to satisfy the entropy inequality

$$\frac{\partial}{\partial t}\eta(\mathbf{U}) + \sum_{j=1}^{2} \frac{\partial}{\partial x_j} q_j(\mathbf{U}) \leq 0$$

in the sense of distributions. For the Euler equations, the thermodynamical entropy density

$$\eta(\mathbf{U}) = -\rho s$$

with entropy

$$s = \log \frac{p}{\rho^\kappa}$$

is a candidate for an entropy function; $\mathbf{U} \mapsto \eta(\mathbf{U})$ is then a convex function of

$$\mathbf{U} = [\rho, \rho u_1, \rho u_2, \rho e]^{\mathrm{T}}\,.$$

The compressible Euler equations (10) can be transformed into a symmetric system of nonlinear partial differential equations using the invertible change of variables

$$\mathbf{U} \mapsto \mathbf{V}(\mathbf{U}) = \nabla_{\mathbf{U}}\eta(\mathbf{U})\,.$$

Indeed, on writing

$$\frac{\partial}{\partial t}\mathbf{U}(\mathbf{V}) + \sum_{j=1}^{2} \frac{\partial}{\partial x_j}\mathbf{F}_j(\mathbf{U}(\mathbf{V})) = \mathbf{S}$$

and applying the chain rule, we deduce that

$$(\nabla_{\mathbf{V}}\mathbf{U})\frac{\partial \mathbf{V}}{\partial t} + \sum_{j=1}^{2}(\nabla_U \mathbf{F}_j(\mathbf{U}(\mathbf{V})))\,(\nabla_{\mathbf{V}}\mathbf{U})\frac{\partial \mathbf{V}}{\partial x_j} = \mathbf{S}\,.$$

On defining $A^0(\mathbf{V}) = \nabla_{\mathbf{V}}\mathbf{U}$, the last equation can be rewritten in the following, more compact, form:

$$A^0(\mathbf{V})\frac{\partial \mathbf{V}}{\partial t} + \sum_{j=1}^{2} A_j(\mathbf{U}(\mathbf{V}))A^0(\mathbf{V})\frac{\partial \mathbf{V}}{\partial x_j} = \mathbf{S}\,.$$

After some calculations we find that the mapping $\mathbf{U} \mapsto \mathbf{V}(\mathbf{U}) = \nabla_{\mathbf{U}}\eta(\mathbf{U})$ is given by

$$\mathbf{V}(\mathbf{U}) = \frac{\kappa - 1}{p}\begin{bmatrix} \frac{p}{\kappa-1}(\kappa + 1 - s) - \rho e \\ \rho u_1 \\ \rho u_2 \\ -\rho \end{bmatrix} \equiv \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix},$$

while the inverse $\mathbf{V} \mapsto \mathbf{U}(\mathbf{V})$ of this mapping can be expressed as

$$\mathbf{U}(\mathbf{V}) = \frac{p}{\kappa - 1}\begin{bmatrix} -V_4 \\ V_2 \\ V_3 \\ 1 - \frac{1}{2}\frac{V_2^2 + V_3^2}{V_4} \end{bmatrix}.$$

Consequently,

$$A^0 = \frac{1}{\kappa - 1}\begin{bmatrix} \rho & \rho u_1 & \rho u_2 & \frac{p}{\kappa-1} + \frac{1}{2}\rho|\mathbf{u}|^2 \\ & p + \rho u_1^2 & \rho u_1 u_2 & \frac{1}{2}\rho u_1|\mathbf{u}|^2 + \frac{\kappa u_1 p}{\kappa-1} \\ & & p + \rho u_2^2 & \frac{1}{2}\rho u_2|\mathbf{u}|^2 + \frac{\kappa u_2 p}{\kappa-1} \\ \text{symm} & & & \frac{1}{4}\rho|\mathbf{u}|^4 - \frac{\kappa|\mathbf{u}|^2 p}{\kappa-1} + \frac{\kappa p^2}{\rho(\kappa-1)^2} \end{bmatrix}.$$

On the other hand,

$$A_1(\mathbf{U}) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{\kappa-3}{2}u_1^2 + \frac{\kappa-1}{2}u_2^2 & (3-\kappa)u_1 & (1-\kappa)u_2 & \kappa - 1 \\ -u_1 u_2 & u_2 & u_1 & 0 \\ (\kappa-1)u_1|\mathbf{u}|^2 - \kappa u_1 e & \kappa e - \frac{\kappa-1}{2}(u_2^2 + 3u_1^2) & (1-\kappa)u_1 u_2 & \kappa u_1 \end{bmatrix},$$

$$A_2(\mathbf{U}) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ -u_1u_2 & u_2 & u_1 & 0 \\ \frac{\kappa-3}{2}u_2^2 + \frac{\kappa-1}{2}u_1^2 & (1-\kappa)u_1 & (3-\kappa)u_2 & \kappa-1 \\ (\kappa-1)u_2|\mathbf{u}|^2 - \kappa u_2 e & (1-\kappa)u_1u_2 & \kappa e - \frac{\kappa-1}{2}(u_1^2+3u_2^2) & \kappa u_2 \end{bmatrix}.$$

Thus, we have transformed the compressible Euler equations (10) into the form

$$\sum_{j=0}^{2} B_j(\mathbf{V})\frac{\partial \mathbf{V}}{\partial x_j} = \mathbf{S}\,, \tag{11}$$

with $x_0 = t$, where $B_0(\mathbf{V}) = A^0(\mathbf{V})$ and $B_j(\mathbf{V}) = A_j(\mathbf{U}(\mathbf{V}))A^0(\mathbf{V})$, $j = 1, 2$, are symmetric 4×4 matrices[3].

4.4 Symmetric linear hyperbolic systems

Suppose that the solution to the nonlinear system (11) can be decomposed as

$$\mathbf{V} = \hat{\mathbf{V}} + \mathbf{W}\,,$$

where $\hat{\mathbf{V}}$ is a (4–component vector) mean state and $\mathbf{W}$ is a small nonconstant perturbation to $\hat{\mathbf{V}}$. In many physical situations the mean state is known and one is only interested in modeling the evolution of the small perturbations $\mathbf{W}$. After some simple algebraic manipulations (see [52], p. 627), we deduce that

$$\sum_{j=0}^{2} B_j(\hat{\mathbf{V}})\frac{\partial \mathbf{W}}{\partial x_j} = \mathbf{S}\,,$$

with $x_0 = t$, which is a symmetric linear hyperbolic system; equivalently, the last set of equations can be rewritten in the following *conservative* (or *conservation*) *form*:

$$\sum_{j=0}^{2} \frac{\partial}{\partial x_j}(B_j(\hat{\mathbf{V}})\mathbf{W}) + C\mathbf{W} = \mathbf{S}\,, \tag{12}$$

where

$$C = -\sum_{j=0}^{2} \frac{\partial}{\partial x_j} B_j(\hat{\mathbf{V}})\,.$$

In particular, when the mean state $\hat{\mathbf{V}}$ is a constant vector then C is the 4×4 zero-matrix.

We explained how the fundamental equations of compressible fluid dynamics lead to symmetric (linear or nonlinear) hyperbolic systems. However, symmetric linear hyperbolic systems also arise in many other physical

[3] The symmetry of the matrices B_1 and B_2 may be verified by evaluating the matrix products $A_j(\mathbf{U}(\mathbf{V}))A^0(\mathbf{V})$, $j = 1, 2$.

applications such as acoustics, electromagnetism and linear elasticity. This motivates us to adopt, as linear model problem, the following symmetric hyperbolic system where the principal part of the operator is in conservation form:

$$\sum_{j=1}^{n} \frac{\partial}{\partial x_j}(\mathbf{B}_j \mathbf{u}) + \mathbf{C}\mathbf{u} = \mathbf{f} \qquad \text{on } \Omega\,;$$

here Ω is a bounded open polyhedral domain in $\mathbb{R}^n$. This system of partial differential equation will be considered in tandem with appropriate boundary conditions which will be stated below after we have made appropriate assumptions on the smoothness of the data.

Let Γ denote the union of open faces of the polyhedral domain Ω and suppose that $\mathbf{B} = (\mathbf{B}_1, \ldots, \mathbf{B}_n)$ is an n–component matrix function defined on $\bar{\Omega}$ with $\mathbf{B}_j \in [W^1_\infty(\Omega)]^{m\times m}_{\mathrm{symm}}$, $j = 1, \ldots, n$. Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ denote the unit outward normal vector to Γ, and consider the symmetric matrix

$$\mathbf{B}(\boldsymbol{\mu}) \equiv \mathbf{B}\cdot\boldsymbol{\mu} = \mu_1\mathbf{B}_1 + \ldots + \mu_n\mathbf{B}_n\,.$$

Suppose that $\mathbf{B}$ is diagonalized *via* $\mathbf{B}(\boldsymbol{\mu}) = X(\boldsymbol{\mu})^{-1}\Lambda(\boldsymbol{\mu})X(\boldsymbol{\mu})$, where $\Lambda(\boldsymbol{\mu})$ is a diagonal matrix, with the (real) eigenvalues of $\mathbf{B}(\boldsymbol{\mu})$ appearing along its diagonal. We shall suppose that Γ is nowhere characteristic in the sense that none of the diagonal entries of $\Lambda(\boldsymbol{\mu})$ is zero for any choice of the unit outward normal vector $\boldsymbol{\mu}$ on Γ. We decompose $\Lambda(\boldsymbol{\mu})$ as

$$\Lambda(\boldsymbol{\mu}) = \Lambda_-(\boldsymbol{\mu}) + \Lambda_+(\boldsymbol{\mu})\,,$$

where $\Lambda_-(\boldsymbol{\mu})$ is diagonal and negative semidefinite, and $\Lambda_+(\boldsymbol{\mu})$ is diagonal and positive semidefinite. Thus we define

$$\mathbf{B}_-(\boldsymbol{\mu}) = X(\boldsymbol{\mu})^{-1}\Lambda_-(\boldsymbol{\mu})X(\boldsymbol{\mu}) \quad \text{and} \quad \mathbf{B}_+(\boldsymbol{\mu}) = X(\boldsymbol{\mu})^{-1}\Lambda_+(\boldsymbol{\mu})X(\boldsymbol{\mu})\,.$$

Clearly, for each unit outward normal vector $\boldsymbol{\mu}$ on Γ,

$$\mathbf{B}(\boldsymbol{\mu}) = \mathbf{B}_-(\boldsymbol{\mu}) + \mathbf{B}_+(\boldsymbol{\mu})\,.$$

Given $\mathbf{C} \in [L_\infty(\Omega)]^{m\times m}$, $\mathbf{f} \in [L_2(\Omega)]^m$ and $\mathbf{g} \in [L_2(\Gamma)]^m$, we consider the following hyperbolic boundary value problem: find $\mathbf{u} \in H(\mathcal{L}, \Omega)$ such that

$$\mathcal{L}\mathbf{u} \equiv \sum_{j=1}^{n} \frac{\partial}{\partial x_j}(\mathbf{B}_j\mathbf{u}) + \mathbf{C}\mathbf{u} = \mathbf{f} \quad \text{in } \Omega\,, \quad \mathbf{B}_-(\boldsymbol{\mu})\mathbf{u} = \mathbf{B}_-(\boldsymbol{\mu})\mathbf{g} \quad \text{on } \Gamma\,, \tag{13}$$

where $H(\mathcal{L}, \Omega) = \{\mathbf{v} \in [L_2(\Omega)]^m : \mathcal{L}\mathbf{v} \in [L_2(\Omega)]^m\}$ denotes the *graph space* of the partial differential operator $\mathcal{L}$ in $[L_2(\Omega)]^m$. The analysis of the well-posedness of this linear system is technical; for details, the interested reader is referred to the papers [16,41,50]. When $m = 1$, the multidimensional symmetric linear hyperbolic system (13) collapses to a scalar linear first-order hyperbolic PDE, an equation that is of great physical importance in its own right.

4.5 Scalar linear hyperbolic equations

As a scalar hyperbolic model problem, we consider the hyperbolic boundary-value problem

$$\mathcal{L}u \equiv \operatorname{div}(\mathbf{b}u) + c\,u = f \quad \text{in } \Omega\,, \qquad u = g \quad \text{on } \Gamma_-\,, \tag{14}$$

where Ω is a bounded open polyhedron in $\mathbb{R}^n$, with *inflow boundary*

$$\Gamma_- = \{x \in \Gamma \,:\, \mathbf{b}(x) \cdot \boldsymbol{\mu}(x) < 0\}\,;$$

here Γ denotes the union of open faces of Ω, $\boldsymbol{\mu}(x)$ denotes the unit outward normal vector at $x \in \Gamma$. The complement of Γ_- with respect to Γ will be denoted Γ_+ and will be referred to as the *outflow boundary.* We shall suppose that $\mathbf{b} = (b_1, \ldots, b_n) \in [W^1_\infty(\Omega)]^n$, $c \in L_\infty(\Omega)$, $f \in L_2(\Omega)$, and $g \in L_2(\Gamma_-)$.

For the sake of simplicity we adopt the following additional hypothesis.

Hypothesis 1 *Suppose that $\Omega = (0,1)^n$ and the components $b_1, \ldots, b_n$ of the vector field* $\mathbf{b}$ *are strictly positive functions on $\bar{\Omega}$.*

This assumption ensures that Γ is non-characteristic for the differential operator $v \mapsto \mathcal{L}v$, and leads to the following well-posedness result whose proof, in the case of $g = 0$, may be found in [53]; in the case of $g \in L_2(\Gamma_-)$ nonzero, it is a special case of a general result of Rauch [50] (see also, [5] and [13]).

Proposition 3. *Under Hypothesis 1 and given that $f \in L_2(\Omega)$, $g \in L_2(\Gamma_-)$, and $c \in L_\infty(\Omega)$, problem (14) has a unique solution u in the graph-space $H(\mathcal{L}, \Omega) = \{v \in L_2(\Omega) \,:\, \mathcal{L}v \in L_2(\Omega)\}$.*

We note that an analogous result holds for the scalar multi-dimensional linear hyperbolic equation written in *nonconservative form*:

$$\mathbf{b} \cdot \nabla u + c\,u = f \quad \text{in } \Omega\,, \qquad u = g \quad \text{on } \Gamma_-\,,$$

where $\mathbf{b}$, c, f, g and Ω are as in Proposition 3. Scalar hyperbolic PDEs in nonconservative form arise in the mathematical modeling of transport phenomena.

5 Stabilized FE methods for transport problems

In this section we return to the difficulties raised at the beginning of the paper concerning the *a posteriori* error analysis of hyperbolic problems. As we have seen there, the dual solution carries important information about the sensitivity of the error to local perturbations of the residual. It is therefore essential that the local variation of the dual solution is reflected by the error bound. Guided by this observation, here we shall develop Type I *a posteriori* error bounds for a class of stabilized finite element approximations of first-order scalar linear hyperbolic PDEs in nonconservative form. The discussion in this section is a brief survey of our paper with Rolf Rannacher [30].

Let us consider the transport problem:

$$\mathcal{L}u \equiv \mathbf{b} \cdot \nabla u + cu = f\,, \quad x \in \Omega\,, \qquad u = g\,, \quad x \in \Gamma_-\,, \tag{15}$$

where $\Omega = (0,1)^n$. With a slight abuse of notation, here we have used the symbol $\mathcal{L}$ to denote the scalar hyperbolic differential operator $\mathbf{b} \cdot \nabla + c$ in nonconservative form, even though in the previous section $\mathcal{L}$ signified the related partial differential operator in conservation form. Since $\mathbf{b} \in [W^1_\infty(\Omega)]^n$, we can safely move back and forth between the two forms, only having to alter the coefficient c if necessary; therefore, we have chosen not to distinguish between the two forms in terms of $\mathcal{L}$. Of course, if $\nabla \cdot \mathbf{b} = 0$ then the conservative- and nonconservative-form differential operators are identical.

As in the previous section, we assume that the entries $b_1, \dots, b_n$ of the n–component vector function $\mathbf{b} \in [W^1_\infty(\Omega)]^n$ are positive on $\bar{\Omega} = [0,1]^n$. This hypothesis ensures that Γ, the union of open faces of Ω, is noncharacteristic for the operator $\mathcal{L}$ at each point x of Γ. Also, we shall suppose that $c \in L_\infty(\Omega)$, $f \in L_2(\Omega)$, and $g \in L_2(\Gamma_-)$. In addition, it will be assumed that there exists $\gamma > 0$ such that

$$c_0^2(x) \equiv c(x) - \frac{1}{2}\nabla \cdot \mathbf{b}(x) \geq \gamma^2 \quad \text{for all } x \in \bar{\Omega}\,. \tag{16}$$

In order to set up the variational formulation of the boundary value problem (15), we associate with $\mathcal{L}$ the graph space

$$H(\mathcal{L},\Omega) = \{v \in L_2(\Omega) \,:\, \mathcal{L}v \in L_2(\Omega)\}\,.$$

We consider the bilinear form $B(\cdot,\cdot) \,:\, H(\mathcal{L},\Omega) \times H(\mathcal{L},\Omega) \to \mathbb{R}$ defined by

$$B(w,v) = (\mathcal{L}w, v) - ((\mathbf{b} \cdot \boldsymbol{\mu})g, v)_{\Gamma_-}$$

and introduce the linear functional $\ell \,:\, H(\mathcal{L},\Omega) \to \mathbb{R}$ by

$$\ell(v) = (f, v) - ((\mathbf{b} \cdot \boldsymbol{\mu})g, v)_{\Gamma_-}\,.$$

In these definitions, $(\cdot,\cdot)$ denotes the $L_2(\Omega)$ inner product and $(\cdot,\cdot)_{\Gamma_-}$ is the $L_2(\Gamma_-)$ inner product with respect to the surface measure $\mathrm{d}\sigma$ (with analogous definition of $(\cdot,\cdot)_{\Gamma_+}$). In terms of this notation, the boundary value problem (15) can be restated in the following variational form: find $u \in H(\mathcal{L},\Omega)$ such that

$$B(u,v) = \ell(v) \qquad \forall v \in H(\mathcal{L},\Omega)\,. \tag{17}$$

Suppose that $\mathcal{T}_h$ is a finite element partition of the computational domain Ω into open element domains κ. It will be assumed that the family $\mathcal{T}_h$ is shape regular (see, [10], for example); namely, there exists a positive constant α

such that $h_\kappa/\rho_\kappa \le \alpha$ for all $\kappa \in \bigcup_h \mathcal{T}_h$, where $h_\kappa = \mathrm{diam}(\bar{\kappa})$ and ρ_κ is the diameter of the largest ball contained in $\bar{\kappa}$. Let us denote by $h(x)$ the mesh function defined on the union of the elements κ in $\mathcal{T}_h$ whose value on element κ is identically equal to h_κ. We then consider on $\mathcal{T}_h$ the finite element space $S_h \subset H^1(\Omega) \subset H(\mathcal{L},\Omega)$ consisting of continuous piecewise polynomial functions of maximum degree p, $p \ge 1$. It will be assumed that S_h possesses the following approximation property:

(H) Given that $v \in H^{s+1}(\Omega)$ and $v|_{\Gamma_-} \in H^{s+1}(\Gamma_-)$ for some s, $0 \le s \le p$, there exists $P_h v$ in S_h and a positive constant C_{int}, independent of v and the mesh function h, such that

$$\|v - P_h v\|_{L_2(\kappa)} + h_\kappa |v - P_h v|_{H^1(\kappa)} \le C_{\mathrm{int}} h_\kappa^{s+1} |v|_{H^{s+1}(\hat{\kappa})} \quad \forall \kappa \in \mathcal{T}_h\,,$$
$$\|v - P_h v\|_{L_2(\partial\kappa \cap \Gamma_-)} \le C_{\mathrm{int}} h_\kappa^{s+1} |v|_{H^{s+1}(\partial\hat{\kappa} \cap \Gamma_-)} \;\; \forall \kappa \in \mathcal{T}_h \text{ s.t. } \partial\kappa \cap \Gamma_- \neq \emptyset\,.$$

In this hypothesis $\hat{\kappa}$ denotes the union of all such elements (including κ itself) whose closure has nonempty intersection with the closure of κ. Hypothesis (H) may be satisfied by choosing $P_h v$ as a finite element quasi-interpolation operator based on local averaging that involves the neighbors of κ. A further possibility is to define $P_h v \in S_h$ at the degrees of freedom interior to $\Omega \cup \Gamma_+$ as indicated in the previous sentence, while on Γ_- one can define $P_h v|_{\Gamma_-}$ as the orthogonal projection of $v|_{\Gamma_-}$ with respect to the inner product

$$\langle w, v \rangle_- = ((\mathbf{b}\cdot\boldsymbol{\mu})w, v)_{\Gamma_-}\,;$$

the inner product $\langle w, v\rangle_+$ is defined analogously.

We note that, according to Proposition 2, the regularity hypothesis $v \in H^{s+1}(\Omega)$ implies $v \in H^{s+1/2}(\Gamma_-)$ only; therefore the regularity of v on Γ_- assumed in (H) marginally exceeds the regularity on Γ_- that is implied by Proposition 2.

We are now ready to consider stabilized finite element approximations of our model problem. Let δ be a positive function contained in $L_\infty(\Omega)$; δ will be referred to as the *stabilization parameter*. The stabilized finite element approximation of (15) is then defined as follows: find $u_h \in S_h$ such that

$$B_\delta(u_h, v_h) = l_\delta(v_h) \quad \forall v_h \in S_h\,, \tag{18}$$

where the bilinear functional $B_\delta : H(\mathcal{L},\Omega) \times H(\mathcal{L},\Omega) \to \mathbb{R}$ and the linear functional $\ell_\delta : H(\mathcal{L},\Omega) \to \mathbb{R}$ are defined by

$$B_\delta(w,v) = (\mathcal{L}w, v + \delta\hat{\mathcal{L}}v) - \langle w, v\rangle_-\,, \qquad l_\delta(v) = (f, v + \delta\hat{\mathcal{L}}v) - \langle g, v\rangle_-\,,$$

respectively, with $\mathcal{L}w = \mathbf{b}\cdot\nabla w + cw$ and $\hat{\mathcal{L}}w = \mathbf{b}\cdot\nabla w + \hat{c}w$. Depending on the choice of $\hat{c}$ we obtain different stabilization techniques; two typical choices are listed below:

$$\hat{c} = \begin{cases} 0 & \text{Streamline-Diffusion FEM (SDFEM)}\,, \\ c & \text{Least-Squares FEM (LSFEM)}\,. \end{cases}$$

The standard choice of the stabilization parameter, based on *a priori* error analysis, is $\delta = C_\delta h$, where C_δ is a positive constant which should be selected by the user. Condition (16) implies that $B_\delta(v,v) > 0$ for all $v \in S_h \setminus \{0\}$. As (18) is a linear problem over a finite–dimensional space S_h, the existence of a unique solution u_h to (18) follows from the positivity of $B_\delta(v,v)$, $v \in S_h \setminus \{0\}$.

5.1 *A posteriori* error estimation for linear functionals

In many problems of physical interest the quantity of interest is an output or target functional of the solution to the partial differential equation rather than the solution itself. Relevant physical examples have already been discussed in the Introduction. For simplicity, here we restrict ourselves to linear target functionals; the extension to general nonlinear output functionals will be discussed in Section 6.4.

Let us suppose, for example, that we wish to control the discretization error in some linear functional $J(\cdot)$ defined on $H(\mathcal{L},\Omega) + S_h$. A typical situation might be that a certain tolerance $\mathtt{TOL} > 0$ is given and the aim of the computation is to find a subdivision $\mathcal{T}_h$ of the computational domain Ω and u_h in the finite element space S_h associated with $\mathcal{T}_h$ such that

$$|J(u) - J(u_h)| < \mathtt{TOL}\,.$$

Problems of this kind are referred to as *measurement problems*. In order to solve a measurement problem, we resort to *a posteriori* error analysis to derive a 'computable' bound on $|J(u) - J(u_h)|$ and then perform adaptive mesh refinement until the *a posteriori* error bound drops below the specified tolerance.

Here we shall discuss the derivation of an *a posteriori* error bound on $|J(u) - J(u_h)|$ in the case when a stabilized finite element method is used for the computation of u_h. The derivation of the *a posteriori* error bound will be based on a duality argument. Two approaches will be considered: the first approach involves the formal adjoint of the partial differential operator, while the second approach exploits the adjoint of the stabilized primal operator. Thus the two alternative dual problems are as follows:

$$\text{Find } z \in H(\mathcal{L},\Omega) \text{ such that} \qquad B(w,z) = J(w) \quad \forall w \in H(\mathcal{L},\Omega)\,; \quad (19)$$

$$\text{Find } z_\delta \in H(\mathcal{L},\Omega) \text{ such that} \quad B_\delta(w,z_\delta) = J(w) \quad \forall w \in H(\mathcal{L},\Omega)\,. \quad (20)$$

These two dual problems will lead to two different *a posteriori* error bounds, and one of our aims will be to assess which of the two provides a sharper estimate of the error. The starting point for the error analysis is the following result which states the error representation formulas that arise from the two dual problems.

Theorem 1. *The dual problems (19), (20) give rise to the following error representation formulas, respectively:*

$$J(u) - J(u_h) = -\langle r_h^-, z - z_h\rangle_- + (r_h, z - z_h) - (r_h, \delta\hat{\mathcal{L}} z_h)\,, \tag{21}$$

$$J(u) - J(u_h) = -\langle r_h^-, z_\delta - z_{\delta,h}\rangle_- + (r_h, z_\delta - z_{\delta,h}) + (r_h, \delta\hat{\mathcal{L}}(z_\delta - z_{\delta,h}))\,, \tag{22}$$

where $r_h = f - \mathcal{L}u_h$ *is the* internal residual, $r_h^- = g - u_h$ *is the boundary residual,* $z_h = P_h z$ *and* $z_{\delta,h} = P_h z_\delta$.

Proof. In order to prove (22), we begin by noting the following *Galerkin orthogonality property*:

$$B_\delta(u - u_h, v_h) = (r_h, v_h + \delta\hat{\mathcal{L}} v_h) - \langle r_h^-, v_h\rangle_- = 0 \quad \forall v_h \in S_h\,.$$

Then, by virtue of the linearity of $J(\cdot)$ and the definition of the dual problem (20), we have that

$$J(u) - J(u_h) = J(u - u_h) = B_\delta(u - u_h, z_\delta) = B_\delta(u - u_h, z_\delta - z_{\delta,h})\,,$$

where $z_{\delta,h} = P_h z_\delta$. Hence,

$$J(u) - J(u_h) = \ell_\delta(z_\delta - z_{\delta,h}) - B_\delta(u_h, z_\delta - z_{\delta,h})\,.$$

We then deduce (22) by recalling the definitions of the linear functional $\ell_\delta(\cdot)$ and the bilinear form $B_\delta(\cdot,\cdot)$.

The proof of (21), using the standard dual problem (19), is completely analogous, except that Galerkin orthogonality does not enter into the argument. □

It follows from Theorem 1 that

$$J(u) - J(u_h) = \sum_{\kappa\in\mathcal{T}_h} \eta_\kappa\,,$$

where, letting $z_h = P_h z$ and $z_{h,\delta} = P_h z_\delta$, we have the following definitions of the *local refinement indicator* η_κ corresponding to the dual problems (19) and (20), respectively:

$$\eta_\kappa = \begin{cases} -((\mathbf{b}\cdot\boldsymbol{\mu})\, r_h^-, z - z_h)_{\partial\kappa\cap\Gamma_-} + (r_h, z - z_h)_\kappa - (r_h, \delta\hat{\mathcal{L}} z_h)_\kappa, \\ -((\mathbf{b}\cdot\boldsymbol{\mu})\, r_h^-, z_\delta - z_{\delta,h})_{\partial\kappa\cap\Gamma_-} + (r_h, z_\delta - z_{\delta,h})_\kappa + (r_h, \delta\hat{\mathcal{L}}(z_\delta - z_{\delta,h}))_\kappa. \end{cases}$$

We adopt the following notational convention: $\eta_\kappa^{(1)}$ will denote the first line in the definition of η_κ corresponding to the standard dual problem, while $\eta_\kappa^{(2)}$ will signify the second line in the definition of η_κ associated with the stabilization-dependent dual problem.

We see from this discussion that a necessary and sufficient condition for ensuring that

$$|J(u) - J(u_h)| < \mathtt{TOL} \tag{23}$$

is that

$$|\mathcal{E}_\Omega(u_h)| < \mathtt{TOL}\,, \qquad \text{where } \mathcal{E}_\Omega(u_h) \equiv \textstyle\sum_{\kappa\in\mathcal{T}_h} \eta_\kappa\,. \tag{24}$$

If (24) holds then $J(u_h)$ is accepted as an accurate representation of $J(u)$; otherwise, u_h is discarded and a new (hopefully more accurate) approximation is computed on a new partition which is obtained from $\mathcal{T}_h$ by refining certain elements $\kappa \in \mathcal{T}_h$. In order to ensure that refinement is performed only where it is necessary, a local decision has to be made on each element κ in $\mathcal{T}_h$ as to whether h_κ is acceptable in relation to TOL. A convenient approach to obtaining a refinement criterion that relates the local element size h_κ to the global tolerance TOL is to *localize* $\mathcal{E}_\Omega(u_h)$ by noting that the following is a sufficient condition for (24) to hold (and thereby also for (23)):

$$\mathcal{E}_{|\Omega|}(u_h) < \mathtt{TOL}\,, \qquad \text{where } \mathcal{E}_{|\Omega|}(u_h) \equiv \textstyle\sum_{\kappa\in\mathcal{T}_h} |\eta_\kappa|\,. \tag{25}$$

In the following section we discuss the implementation of this bound into an adaptive mesh refinement algorithm. However, before proceeding, a word of caution is in order: the seemingly harmless transition from (24) to (25) can be detrimental in the sense that the asymptotic rate of convergence of the resulting *a posteriori* error bound (25) may be suboptimal, leading to uneconomical mesh design. Indeed, the presence of such adverse effects in the *a posteriori* error bound is closely related to *global* superconvergence phenomena in the three terms constituting (21) and (22), respectively, which may be destroyed in the course of localization. We end this section by showing that the persistence of superconvergence under localization in the three terms of the error representation formula is closely related to the choice of the dual problem. To this end, let us first introduce some notation. We write I_1, II_1 and III_1 to denote the three terms (in modulus) on the right-hand side of (21), respectively; i.e.

$$\mathrm{I}_1 = |\langle r_h^-, z - z_h\rangle_-|\,, \quad \mathrm{II}_1 = |(r_h, z - z_h)|\,, \quad \mathrm{III}_1 = |(r_h, \delta\hat{\mathcal{L}} z_h)|\,. \tag{26}$$

Similarly, we write

$$\begin{aligned} \mathrm{I}_{1,\delta} = |\langle r_h^-, z_\delta - z_{\delta,h}\rangle_-|\,, \qquad & \mathrm{II}_{1,\delta} = |(r_h, z_\delta - z_{\delta,h})|\,, \\ & \mathrm{III}_{1,\delta} = |(r_h, \delta\hat{\mathcal{L}}(z_\delta - z_{\delta,h}))| \end{aligned} \tag{27}$$

to denote the moduli of the three terms on the right-hand side of (22), respectively.

Next we recall the following result which is a minor variation on the *a priori* error bound derived in [38].

Lemma 1. *Assume that (H) holds and let u and u_h denote the solutions of (15) and (18), respectively. Suppose that $u \in H^{s+1}(\Omega)$ for some s, $0 \le s \le p$, and there exist positive constants c_δ and C_δ such that $c_\delta h \le \delta \le C_\delta h$ on Ω; then*

$$|||u - u_h||| \le C h^{s+1/2} |u|_{H^{s+1}(\Omega)}\,,$$

where

$$|||v|||^2 = \|\sqrt{\delta}\,\mathcal{L}v\|^2_{L_2(\Omega)} + \|v\|^2_{L_2(\Omega)} + \|v\|^2_{L_2(\Gamma)}$$

and C is a constant independent of the mesh function h.

Thereby, from Hypothesis (H) and Lemma 1, we deduce the following *a priori* bound for the error in the computed linear functional $J(\cdot)$.

Theorem 2. *Assume that (H) holds and let u and u_h denote the solutions of (15) and (18), respectively. Given that $u \in H^{s+1}(\Omega)$ for some s, $0 \leq s \leq p$, the dual solution z satisfying (19) is sufficiently regular, and there exist positive constants c_δ and C_δ such that $c_\delta h \leq \delta \leq C_\delta h$ on Ω, we have that*

$$\begin{aligned}
\mathrm{I}_1 &\leq Ch^{2s+3/2}\,|u|_{H^{s+1}(\Omega)}\,|z|_{H^{s+1}(\Gamma_-)}\,,\\
\mathrm{II}_1 &\leq Ch^{2s+1}\,|u|_{H^{s+1}(\Omega)}\,|z|_{H^{s+1}(\Omega)}\,,\\
\mathrm{III}_1 &\leq Ch^{2s+1}\,|u|_{H^{s+1}(\Omega)}\,|z|_{H^{s+1}(\Omega)} + Ch^2\,|u|_{H^{s+1}(\Omega)}\,|z|_{H^1(\Omega)}\,,
\end{aligned}$$

where C is a constant independent of the mesh function h. Analogously, assuming that the solution z_δ of the stabilization–dependent dual problem (20) is sufficiently regular, we deduce that

$$\begin{aligned}
\mathrm{I}_{1,\delta} &\leq Ch^{2s+3/2}\,|u|_{H^{s+1}(\Omega)}\,|z_\delta|_{H^{s+1}(\Gamma_-)}\,,\\
\mathrm{II}_{1,\delta} &\leq Ch^{2s+1}\,|u|_{H^{s+1}(\Omega)}\,|z_\delta|_{H^{s+1}(\Omega)}\,,\\
\mathrm{III}_{1,\delta} &\leq Ch^{2s+1}\,|u|_{H^{s+1}(\Omega)}\,|z_\delta|_{H^{s+1}(\Omega)}\,.
\end{aligned}$$

Thereby, Theorem 2 indicates that under suitable smoothness for both primal and dual solutions u and z, respectively, the error in the computed target functional $J(\cdot)$ is of size $\mathcal{O}(h^2)$, when the standard dual problem is employed, *irrespective* of the degree p of the approximating polynomial used in the finite element space S_h. In contrast, employing the error representation formula based on the stabilization–dependent dual problem, cf. (22), we find that

$$|J(u) - J(u_h)| \leq Ch^{2s+1}\,,$$

for $0 \leq s \leq p$, provided that $z_\delta \in H^{s+1}(\Omega)$ with $\|z_\delta\|_{H^{s+1}(\Omega)} \leq$ *Const.*, independent of h.

Finally, we note that if we assume that $u \in H^{s+1}(\Gamma_-)$ for some s, $0 \leq s \leq p$, then the *a priori* error bound for the boundary terms I_1 and $I_{1,\delta}$ stated in Theorem 2 may be improved provided that the boundary residual is redefined as $r_h^- = g - P_h g$. In this case, we deduce that

$$I_1 \leq Ch^{2s+2}\,|u|_{H^{s+1}(\Gamma_-)}\,|z|_{H^{s+1}(\Gamma_-)}\,,$$

provided that $z \in H^{s+1}(\Gamma_-)$, and

$$I_{1,\delta} \leq Ch^{2s+2}\,|u|_{H^{s+1}(\Gamma_-)}\,|z_\delta|_{H^{s+1}(\Gamma_-)}\,,$$

provided that $z_\delta \in H^{s+1}(\Gamma_-)$.

5.2 Mesh refinement criteria and mesh modification strategies

Given a user defined tolerance TOL, in this section we provide an outline of mesh refinement criteria and mesh modification strategies that may be employed to ensure the reliable and efficient control of the error in the computed functional $J(\cdot)$.

Local tolerance criterion. A possible mesh refinement criterion might consist of checking whether on each element κ in the partition $\mathcal{T}_h$ the following inequality holds:

$$|\eta_\kappa| \le \frac{\mathtt{TOL}}{N}\,, \tag{28}$$

where N is the number of elements in $\mathcal{T}_h$. If the inequality (28) is violated on an element $\kappa \in \mathcal{T}_h$, then κ is refined; otherwise κ is accepted as being of adequate size. It is also possible to incorporate derefinement into the algorithm by selecting θ, $0 < \theta \ll 1$, and marking elements κ with

$$|\eta_\kappa| \le \theta \frac{\mathtt{TOL}}{N}$$

for derefinement.

Fixed fraction criterion. This involves choosing two numbers φ_{ref} and φ_{deref} in the interval $(0, 100)$ with $\varphi_{\mathrm{deref}} + \varphi_{\mathrm{ref}} < 100$, ordering the *local refinement indicators* $|\eta_\kappa|$, $\kappa \in \mathcal{T}_h$, according to their size, and then refining those elements κ which correspond to φ_{ref}% of the largest entries in the ordered sequence (the top 20%, say), and derefining those elements κ which correspond to the φ_{deref}% of the smallest entries in this ordered sequence (the bottom 10%, say); further variations on this strategy, with dynamically varying φ_{ref} and φ_{deref}, are also possible.

Optimized mesh criterion. Yet a further technique, called the *optimized mesh strategy* [49] (see, also, [19] and [22]) aims to design a subdivision $\mathcal{T}_h$ for the primal problem so that

$$\mathcal{E}_{|\Omega|}(u_h) \le \mathtt{TOL}$$

subject to the constraint that the number N of elements in the subdivision $\mathcal{T}_h$ is minimized. Suppose, for example, that we have chosen to use the stabilization-parameter-dependent dual problem (20); then,

$$\eta_\kappa = -((\mathbf{b}\cdot\boldsymbol{\mu})\, r_h^-, z_\delta - z_{\delta,h})_{\partial\kappa\cap\Gamma_-} + (r_h, z_\delta - z_{\delta,h})_\kappa + (r_h, \delta\hat{\mathcal{L}}(z_\delta - z_{\delta,h}))_\kappa\,.$$

Let us suppose that the finite element space S_h consists of continuous piecewise polynomials of degree $p = 1$. We then define

$$\begin{aligned} A(x) &= |r_h(z_\delta - z_{\delta,h}) + \delta r_h \hat{\mathcal{L}}(z_\delta - z_{\delta,h})|/h_\Omega^3(x)\,,\\ B(x) &= |(\mathbf{b}\cdot\boldsymbol{\mu})(g - u_h)(z_\delta - z_{\delta,h})|/h_{\Gamma_-}^4(x)\,, \end{aligned}$$

where h_Ω and h_{Γ_-} are the mesh functions on $\Omega \cup \Gamma_+$ and Γ_-, respectively. After an elementary calculation based on the use of Lagrange multipliers, we arrive at the following optimal mesh functions $h_\Omega^{\mathrm{opt}}(x)$ and $h_{\Gamma_-}^{\mathrm{opt}}(x)$:

$$h_\Omega^{\mathrm{opt}} = \left(\frac{2}{3\lambda A}\right)^{1/5}, \qquad h_{\Gamma_-}^{\mathrm{opt}} = \left(\frac{1}{4\lambda B}\right)^{1/5},$$

where λ is the positive root of

$$\left(\frac{2}{3\lambda}\right)^{3/5} \int_\Omega A^{2/5}\,\mathrm{d}x + \left(\frac{1}{4\lambda}\right)^{4/5} \int_{\Gamma_-} B^{1/5}\,\mathrm{d}\sigma = \mathtt{TOL}\,.$$

For $\mathtt{TOL} \ll 1$ we expect $\lambda \gg 1$, so that $(1/\lambda)^{4/5} \ll (1/\lambda)^{3/5}$. Thus, for simplicity, we may neglect the boundary integral term in the last equality. Thereby, we can explicitly determine λ in terms of TOL and the integral of $A^{2/5}$; substituting the resulting expression into the formulae for h_Ω^{opt}, we obtain

$$h_\Omega^{\mathrm{opt}}(x) \approx \left(\frac{\mathtt{TOL}}{W}\right)^{1/3} \frac{1}{A^{1/5}(x)}, \qquad \text{where} \qquad W = \int_\Omega A^{2/5}(x)\,\mathrm{d}x\,,$$

with a similar expression for $h_{\Gamma_-}^{\mathrm{opt}}$.

Any of these criteria can be coupled with a suitable mesh modification algorithm. For example, in two space dimensions a popular mesh modification strategy is a red-green refinement. Here, the user must first specify a coarse *background mesh* upon which any future refinement will be based. Red refinement corresponds to dividing a certain triangle into four similar triangles by connecting the midpoints of the three sides. Since red refinement is performed only locally (rather than in each element in the triangulation) hanging nodes are created in the mesh; green refinement is then used to remove any hanging nodes in the mesh created in the course of red refinement by connecting a hanging node on an edge to the opposite vertex of the triangle. Green refinement is temporary and is only applied to elements which contain one hanging node; on elements with two or more hanging nodes red refinement is performed. Within this mesh modification algorithm elements may also be removed from the mesh through derefinement provided they do not lie in the original background mesh. It is perhaps worth noting here that the removal of hanging nodes through green refinement is necessary only if S_h is contained in $C(\bar{\Omega})$, as is the case for SDFEM considered here. In the discontinuous Galerkin finite element method to be discussed in the next section it is not assumed that S_h is contained in $C(\bar{\Omega})$, so the existence of hanging nodes in the mesh is perfectly acceptable there.

An important issue concerning practical implementation of the *stopping criterion* (25) and the *refinement criterion*, cf. (28) for example, into the adaptive algorithm is that the calculation of η_κ presupposes knowledge of

z or z_δ, depending on which error representation formula is selected. These functions, however, are the analytical solutions to the dual problems (19) and (20), respectively, and therefore they are not available explicitly but have to be computed numerically. The numerical approximation of the dual solution requires special care and will be discussed in more detail in the next section. We shall comment on the numerical approximation of the solution z_δ of (20) only, as problem (19) is a special case of (20) upon formally setting $\delta = 0$ in (20), so identical comments apply to z.

5.3 Numerical approximation of z_δ

Let us recall that the function z_δ is defined as the solution to the following dual problem: find $z_\delta \in H(\mathcal{L}, \Omega)$ such that

$$B_\delta(w, z_\delta) = J(w) \quad \forall w \in H(\mathcal{L}, \Omega) . \tag{29}$$

In order to be more specific, let us consider two relevant examples.

- In the case when the measurement problem concerns a local weighted mean value of the analytical solution u, we have
$$J(u) \equiv M_\zeta(u) = \int_\Omega u(x)\zeta(x) \, \mathrm{d}x ,$$
where $\zeta \in L_2(\Omega)$ is a given nonnegative weight function defined on Ω.
- In the case when the measurement problem concerns the normal advective flux of u across the outflow boundary Γ_+, we have
$$J(u) \equiv N_\psi(u) = \int_{\Gamma_+} (\mathbf{b} \cdot \boldsymbol{\mu}) \, u\psi \, \mathrm{d}\sigma ,$$
where $\psi \in L_2(\Gamma_+)$ is a nonnegative weight function defined on Γ_+.

In the case of the first of these examples, (29) can be rewritten in the following strong form:

$$\begin{aligned} \mathcal{L}^*(z_\delta + \delta\hat{\mathcal{L}} z_\delta) &= \zeta , & x &\in \Omega , \\ \delta\hat{\mathcal{L}} z_\delta &= 0 , & x &\in \Gamma_- , \\ z_\delta + \delta\hat{\mathcal{L}} z_\delta &= 0 , & x &\in \Gamma_+ , \end{aligned} \tag{30}$$

whereas in the case of the second example the strong form of (29) is

$$\begin{aligned} \mathcal{L}^*(z_\delta + \delta\hat{\mathcal{L}} z_\delta) &= 0 , & x &\in \Omega , \\ \delta\hat{\mathcal{L}} z_\delta &= 0 , & x &\in \Gamma_- , \\ z_\delta + \delta\hat{\mathcal{L}} z_\delta &= \psi , & x &\in \Gamma_+ . \end{aligned} \tag{31}$$

The differential operator $\mathcal{L}_\delta$ defined by

$$\mathcal{L}_\delta v \equiv \mathcal{L}^*(v + \delta\hat{\mathcal{L}}v) = -\nabla\cdot(a_\delta\nabla v) + \mathbf{b}_\delta\cdot\nabla v + c_\delta v\,,$$

where

$$a_\delta = \delta\mathbf{b}\mathbf{b}^{\mathrm{T}}, \qquad \mathbf{b}_\delta = (\delta(c-\hat{c})-1)\mathbf{b}\,, \qquad c_\delta = \mathcal{L}^*(1+\delta\hat{c})\,,$$

is a second-order partial differential operator with nonnegative characteristic form, given that the matrix a_δ is only negative semidefinite but not positive definite.

Problem (30) or (31), as the case may be, can be solved numerically using the streamline–diffusion finite element method, say, by considering a shape regular subdivision $\mathcal{T}_{h_d}$ of Ω with mesh function h_δ, where, for $\kappa\in\mathcal{T}_{h_d}$, $h_\delta|_\kappa = \mathrm{diam}(\kappa)$ (see, [34]). We note that, typically, $\mathcal{T}_h$ and $\mathcal{T}_{h_d}$ will differ; after all, the primal and dual problems are of completely different nature so there is no reason (other than that of minimizing the overall cost of the computation) to use the same sequence of computational meshes for the primal and dual problems. Further, we define S_{h_d} to be a finite element space consisting of piecewise polynomials of fixed degree p_d, $p_d\geq 1$ on $\mathcal{T}_{h_d}$. Again, in general, $p_d\neq p$; more importantly, however, if for any reason $\mathcal{T}_{h_d}$ has been chosen to be the same as $\mathcal{T}_h$, then it is necessary to select $p_d > p \geq 1$, for otherwise our numerical approximation to η_κ will be equal to zero and the computed error bound will then become meaningless.

A further question is: *What constitutes an appropriate computational mesh $\mathcal{T}_{h_d}$ for the dual problem?* A suitable mesh for the numerical solution of the dual problem may be designed adaptively through the use of an auxiliary *a posteriori* error bound to ensure that the additional error incurred through the replacement of z (respectively, z_δ) by its finite element approximation $\hat{z}_{h_d}$ (respectively, $\hat{z}_{\delta,h_d}$) does not exceed a small fraction $\varepsilon\mathtt{TOL}$ (with $0<\varepsilon\ll 1$) of the global tolerance $\mathtt{TOL}$. We shall say more about this approach in the next section when we discuss the discontinuous Galerkin finite element method. For the sake of simplicity of exposition, here we shall adopt a pragmatic and completely empirical approach: the sequence of meshes $\{\mathcal{T}_{h_d}\}$ for the finite element approximation of the dual problem will be generated from a triangular background mesh of the computational domain Ω (which is geometrically different from the background mesh that gives rise to the family $\{\mathcal{T}_h\}$); the background mesh for the dual problem will contain about 50% more elements than the background mesh for the primal problem so as to ensure that the error committed in the numerical solution of the dual problem has little or no effect on the local refinement indicators η_κ, $\kappa\in\mathcal{T}_h$. Furthermore, for simplicity, we set $p_d = p = 1$.

5.4 Numerical experiments

In this section we present a series of numerical experiments to both compare the asymptotic rate of convergence of the *a posteriori* error bound (25) us-

Table 2. Example 1: Convergence of $\|u - u_h\|_{L_2(\Omega)}$ and $|N_\psi(u) - N_\psi(u_h)|$ with $\delta = h/4$, and the rate of convergence k

Mesh	$\|u - u_h\|_{L_2(\Omega)}$	k	$\|N_\psi(u) - N_\psi(u_h)\|$	k
17×17	1.723×10^{-3}	-	2.631×10^{-4}	-
33×33	3.541×10^{-4}	2.28	3.581×10^{-5}	2.88
65×65	8.101×10^{-5}	2.13	4.672×10^{-6}	2.94
129×129	1.967×10^{-5}	2.04	5.965×10^{-7}	2.97
257×257	4.871×10^{-6}	2.01	7.534×10^{-8}	2.99

ing both the standard and stabilization–dependent dual problems (19) and (20), respectively, as well as to demonstrate the performance of an adaptive algorithm employing the local refinement indicator $|\eta_\kappa^{(2)}|$.

Example 1. The purpose of this example is to highlight the difference between using the standard dual problem (19) and the stabilization–dependent dual problem (20). Let us take $\Omega = (0,1)^2$, $\mathbf{b} = (1+x, 1+y)$, $c = 0$ and $f = 0$ with boundary condition

$$u(x,y) = \begin{cases} 1 - y^6\,, & \text{for } x = 0,\ 0 \le y \le 1\,, \\ \mathrm{e}^{-10x^4}\,, & \text{for } 0 \le x \le 1,\ y = 0\,. \end{cases}$$

We set $C_\delta = 1/4$ in the definition of the stabilization parameter δ and let

$$\psi = \begin{cases} 1 - \sin^2(\pi(1-y)/2)\cos(\pi y/2)\,, & \text{for } x = 1,\ 0 \le y \le 1\,, \\ 1 - (1-x)^3 - (1-x)^4/2\,, & \text{for } 0 \le x \le 1,\ y = 1\,. \end{cases}$$

The goal of the exercise is to compute the weighted outflow normal flux, $J(u) = N_\psi(u) = ((\mathbf{b}\cdot\boldsymbol{\mu})u, \psi)_{\Gamma_+}$, of the analytical solution u. For purposes of comparison, the analytical solution u and the dual solutions z and z_δ have been computed to high accuracy using the method of characteristics; in particular, the exact value of the weighted outward normal flux was found to be $N_\psi(u) = 2.641445145716141$.

In Table 2 we have displayed the orders of convergence of the error in the $L_2(\Omega)$ norm as well as in the functional $N_\psi(\cdot)$ as h tends to zero on a sequence of uniform triangular meshes obtained from uniform square meshes by cutting each mesh square into two triangles, and S_h consisting of continuous piecewise polynomials of degree 1 ($p = 1$). We observe that $|N_\psi(u) - N_\psi(u_h)|$ converges like $\mathcal{O}(h^3)$ with $\mathcal{O}(h)$ stabilization, while the $L_2(\Omega)$ norm of $u - u_h$ is $\mathcal{O}(h^2)$. In Tables 3 and 4 we show the convergence of each of the terms in the error representation formulas (21) and (22), respectively. Here, we recall the notation

$$\mathrm{I}_1 = |\langle r_h^-, z - z_h\rangle_-|\,, \quad \mathrm{II}_1 = |(r_h, z - z_h)|\,, \quad \mathrm{III}_1 = |(r_h, \delta\hat{\mathcal{L}} z_h)|\,,$$

Table 3. Example 1: Convergence of the terms in the error representation formula (21) with $\delta = h/4$, and the rate of convergence k

Mesh	I_1	k	II_1	k	III_1	k
17×17	4.396×10^{-6}	-	5.667×10^{-6}	-	2.644×10^{-4}	-
33×33	2.980×10^{-7}	3.88	3.569×10^{-7}	3.99	3.587×10^{-5}	2.88
65×65	1.951×10^{-8}	3.93	2.238×10^{-8}	4.00	4.675×10^{-6}	2.94
129×129	1.249×10^{-9}	3.97	1.401×10^{-9}	4.00	5.967×10^{-7}	2.97
257×257	7.903×10^{-11}	3.98	8.764×10^{-11}	4.00	7.535×10^{-8}	2.99

Table 4. Example 1: Convergence of the terms in the error representation formula (22) with $\delta = h/4$, and the rate of convergence k

Mesh	$\mathrm{I}_{1,\delta}$	k	$\mathrm{II}_{1,\delta}$	k	$\mathrm{III}_{1,\delta}$	k
17×17	4.396×10^{-6}	-	1.400×10^{-5}	-	2.447×10^{-4}	-
33×33	2.980×10^{-7}	3.88	2.130×10^{-6}	2.72	3.339×10^{-5}	2.87
65×65	1.951×10^{-8}	3.93	2.748×10^{-7}	2.95	4.378×10^{-6}	2.93
129×129	1.249×10^{-9}	3.97	3.414×10^{-8}	3.01	5.611×10^{-7}	2.96
257×257	7.903×10^{-11}	3.98	4.225×10^{-9}	3.01	7.104×10^{-8}	2.98

and

$$\mathrm{I}_{1,\delta} = |\langle r_h^-, z_\delta - z_{\delta,h}\rangle_-| , \qquad \mathrm{II}_{1,\delta} = |(r_h, z_\delta - z_{\delta,h})| ,$$
$$\mathrm{III}_{1,\delta} = |(r_h, \delta\hat{\mathcal{L}}(z_\delta - z_{\delta,h}))|$$

introduced in Section 5.1, cf. (26) and (27), respectively.

Table 3 shows that the second term in the error representation formula (21), i.e. term II_1, is superconvergent; here $\mathrm{II}_1 = \mathcal{O}(h^4)$ as h tends to zero. Term III_1 which arises as the result of the stabilization employed exhibits (superconvergent) $\mathcal{O}(h^3)$ convergence and entirely dominates the error in the weighted outward normal flux. On the other hand, Table 4 indicates that while the first term, term $\mathrm{I}_{1,\delta}$, in the error representation formula (22) associated with the stabilization-dependent dual problem (20) is fourth-order convergent as h tends to zero, the second term, term $\mathrm{II}_{1,\delta}$, is now only $\mathcal{O}(h^3)$ as the mesh is refined, as is term $\mathrm{III}_{1,\delta}$. Therefore, by using the stabilization-dependent dual problem (20) the terms involving integration over the entire computational domain (namely, terms $\mathrm{II}_{1,\delta}$ and $\mathrm{III}_{1,\delta}$) have now become 'balanced'. This balancing of the terms in the error representation formula (22) will be essential for ensuring that any estimate obtained by further bounding $\mathrm{I}_{1,\delta}$, $\mathrm{II}_{1,\delta}$ and $\mathrm{III}_{1,\delta}$ from above upon localization exhibits the same asymptotic convergence rate as $|N_\psi(u) - N_\psi(u_h)|$.

Table 5. Example 1: Comparison of the *a posteriori* error bounds $\mathcal{E}_{|\Omega|}(u_h)$ and $\mathcal{E}_{|\Omega|,\delta}(u_h)$ arising from the dual problems (19) and (20), respectively, and the corresponding convergence rates k and effectivity indices θ and θ_δ

Mesh	$\mathcal{E}_{\vert\Omega\vert}(u_h)$	k	θ	$\mathcal{E}_{\vert\Omega\vert,\delta}(u_h)$	k	θ_δ
17×17	1.921×10^{-3}	-	7.30	2.642×10^{-4}	-	1.00
33×33	4.936×10^{-4}	1.96	13.8	3.639×10^{-5}	2.86	1.02
65×65	1.246×10^{-4}	1.99	26.7	4.912×10^{-6}	2.89	1.05
129×129	3.128×10^{-5}	1.99	52.4	6.420×10^{-7}	2.94	1.08
257×257	7.833×10^{-6}	2.00	104.0	8.624×10^{-8}	2.90	1.14

To demonstrate that this is so when the stabilization-dependent dual problem (20) is used, but not so when the standard dual problem (19) is employed, we show in Table 5 the effectivity indices:

$$\theta = \mathcal{E}_{|\Omega|}(u_h)/|N_\psi(u) - N_\psi(u_h)| \quad \text{and} \quad \theta_\delta = \mathcal{E}_{|\Omega|,\delta}(u_h)/|N_\psi(u) - N_\psi(u_h)|\,,$$

where $\mathcal{E}_{|\Omega|}(u_h)$ (respectively, $\mathcal{E}_{|\Omega|,\delta}(u_h)$) is defined through the use $\eta_\kappa = \eta_\kappa^{(1)}$ (respectively, $\eta_\kappa = \eta_\kappa^{(2)}$). Here, very high accuracy approximations to the dual solutions z and z_δ, computed by means of the method of characteristics, have been used in the calculation of $\eta_\kappa^{(1)}$ and $\eta_\kappa^{(2)}$, respectively, to ensure that the effects observed are not due to additional errors which arise from the numerical solution of the dual problem. We observe that $\mathcal{E}_{|\Omega|}(u_h)$ converges at the suboptimal rate of $\mathcal{O}(h^2)$ as h tends to zero while $\mathcal{E}_{|\Omega|,\delta}(u_h)$ exhibits the same $\mathcal{O}(h^3)$ rate as the true error $|N_\psi(u) - N_\psi(u_h)|$. The deterioration of the asymptotic convergence rate of $\mathcal{E}_{|\Omega|}(u_h)$ results in large effectivity indices which grow under mesh refinement, while θ_δ remains in the range $1 - 1.2$.

The final two tables concerning this example demonstrate that localization adversely affects the *a posteriori* error bound which stems from using the standard dual problem, while the bound that arises from the error representation formula based on the stabilization-dependent dual problem is insensitive to localization. We define

$$\mathrm{I}_2 = \sum_{\kappa \in \mathcal{T}_h} |\langle r_h^-, z - z_h \rangle_{\partial\kappa \cap \Gamma_-}|\,,$$
$$\mathrm{II}_2 = \sum_{\kappa \in \mathcal{T}_h} |(r_h, z - z_h)_\kappa|\,, \qquad \mathrm{III}_2 = \sum_{\kappa \in \mathcal{T}_h} |(r_h, \delta \hat{\mathcal{L}} z_h)_\kappa|\,,$$

and, analogously, we introduce

$$\mathrm{I}_{2,\delta} = \sum_{\kappa \in \mathcal{T}_h} |\langle r_h^-, z_\delta - z_{\delta,h} \rangle_{\partial\kappa \cap \Gamma_-}|\,,$$
$$\mathrm{II}_{2,\delta} = \sum_{\kappa \in \mathcal{T}_h} |(r_h, z_\delta - z_{\delta,h})_\kappa|\,, \qquad \mathrm{III}_{2,\delta} = \sum_{\kappa \in \mathcal{T}_h} |(r_h, \delta \hat{\mathcal{L}}(z_\delta - z_{\delta,h}))_\kappa|\,.$$

Table 6. Example 1: Convergence of the terms I_2, II_2 and III_2 in $\mathcal{E}_{|\Omega|}(u_h)$

Mesh	I_2	k	II_2	k	III_2	k
17×17	7.088×10^{-6}	-	5.985×10^{-5}	-	1.954×10^{-3}	-
33×33	4.513×10^{-7}	3.97	7.596×10^{-6}	2.98	4.986×10^{-4}	1.97
65×65	2.923×10^{-8}	3.95	9.502×10^{-7}	3.00	1.253×10^{-4}	1.99
129×129	1.872×10^{-9}	3.96	1.188×10^{-7}	3.00	3.137×10^{-5}	2.00
257×257	1.186×10^{-10}	3.98	1.484×10^{-8}	3.00	7.844×10^{-6}	2.00

Table 7. Example 1: Convergence of the terms $\mathrm{I}_{2,\delta}$, $\mathrm{II}_{2,\delta}$, $\mathrm{III}_{2,\delta}$ in $\mathcal{E}_{|\Omega|,\delta}(u_h)$

Mesh	$\mathrm{I}_{2,\delta}$	k	$\mathrm{II}_{2,\delta}$	k	$\mathrm{III}_{2,\delta}$	k
17×17	7.088×10^{-6}	-	1.497×10^{-4}	-	2.447×10^{-4}	-
33×33	4.513×10^{-7}	3.97	2.297×10^{-5}	2.70	3.339×10^{-5}	2.87
65×65	2.923×10^{-8}	3.95	3.250×10^{-6}	2.82	4.398×10^{-6}	2.92
129×129	1.872×10^{-9}	3.96	4.357×10^{-7}	2.90	5.680×10^{-7}	2.95
257×257	1.186×10^{-10}	3.98	5.657×10^{-8}	2.95	7.609×10^{-8}	2.90

Table 6 shows each of the terms I_2, II_2 and III_2 in the error bound $\mathcal{E}_{|\Omega|}(u_h)$ together with their respective rates of convergence. We see that term I_2 arising from the localization of term I_1 still converges like $\mathcal{O}(h^4)$ as h tends to zero, and term II_2 which stems from the localization of term II_1 is now third-order convergent; however, unlike III_1, term III_2 is now only second-order convergent. Table 7 displays the analogous terms $\mathrm{I}_{2,\delta}$, $\mathrm{II}_{2,\delta}$ and $\mathrm{III}_{2,\delta}$; clearly, these exhibit exactly the same convergence rates as $\mathrm{I}_{1,\delta}$, $\mathrm{II}_{1,\delta}$ and $\mathrm{III}_{1,\delta}$, respectively. This insensitivity of the convergence rate of the *a posteriori* error bound based on the dual problem (29) to localization is of great importance from the point of view of adaptive algorithms, given that any standard mesh refinement criterion will require a local quantity, such as $|\eta_\kappa|$, to decide which elements need to be refined/derefined.

Example 2. Let us again consider the transport equation

$$\mathbf{b} \cdot \nabla u + cu = f$$

in $\Omega = (0,1)^2$, but now with $\mathbf{b} = (10y^2 - 12x + 1, 1 + y)$, $c = 0$ and $f = 0$. Here the characteristics enter Ω through the bottom of the square and through the two vertical sides, and exit through the top edge. We impose the following

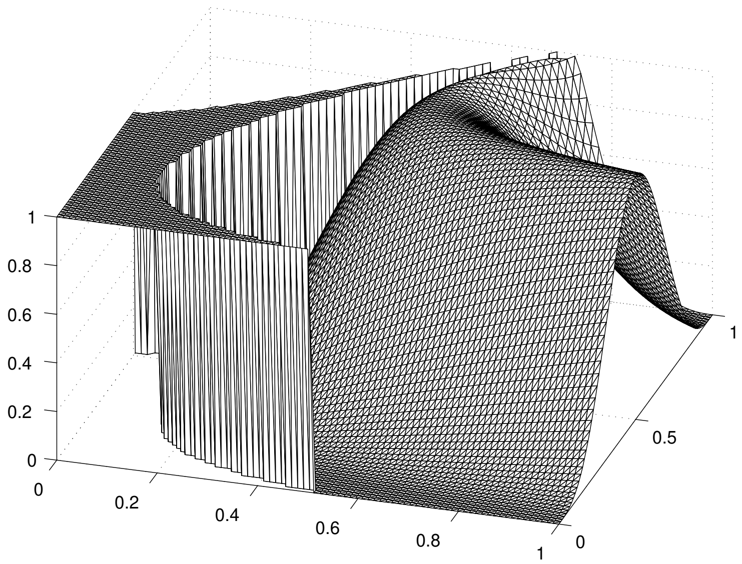


Fig. 4. Example 2: The analytical solution to the primal problem

inflow boundary conditions:

$$u(x,y) = \begin{cases} 0\,, & \text{for } x=0,\ 0.5<y\le 1\,, \\ 1\,, & \text{for } x=0,\ 0<y\le 0.5\,, \\ 1\,, & \text{for } 0\le x\le 0.5,\ y=0\,, \\ 0\,, & \text{for } 0.5<x\le 1,\ y=0\,, \\ \sin^3(\pi y)\,, & \text{for } x=1,\ 0\le y\le 1\,. \end{cases}$$

The analytical solution is depicted in Fig. 4. Let us suppose that the aim of the computation is to calculate the weighted normal flux $N_\psi(u)$ of the analytical solution u through the outflow edge of the square, where the weight-function ψ is defined by

$$\psi(x) = \sin(\pi x/2)\,, \qquad \text{for } 0\le x\le 1,\, y=1\,.$$

For purposes of comparison, using the method of characteristics one may compute a highly accurate approximation to u and thereby deduce that $N_\psi(u) = 0.198826929742166869$. In this example we shall only use the *a posteriori* error bound arising from the stabilization-dependent dual problem (20).

Numerical experiments analogous to those in Example 1 indicate that on both uniform triangular and unstructured triangular meshes the convergence rate of the error measured in the $L_2(\Omega)$ norm is $k\approx 0.4$ due to the presence of the two discontinuities in the solution u. In contrast, the error in the weighted outward normal flux $N_\psi(\cdot)$ is $\mathcal{O}(h^3)$ as h tends to zero, cf. Fig. 5. The high

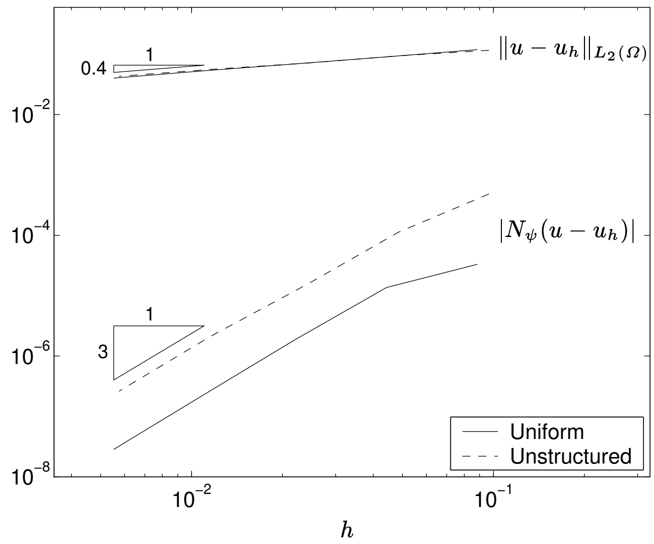
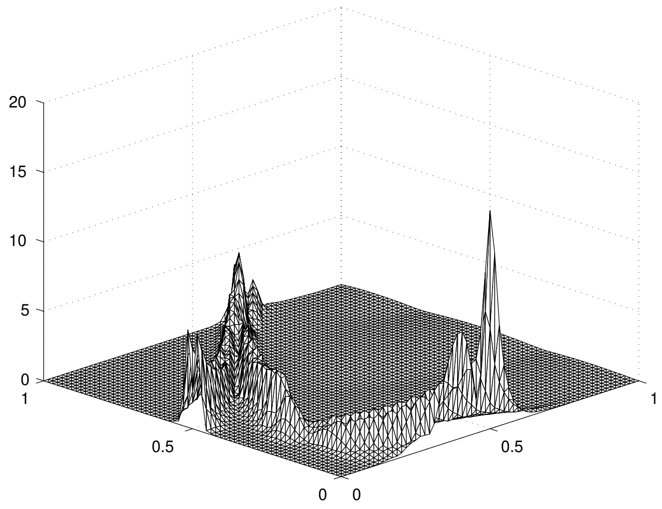


Fig. 5. Example 2: Convergence of $\|u - u_h\|_{L_2(\Omega)}$ and $|N_\psi(u - u_h)|$ with uniform mesh refinement

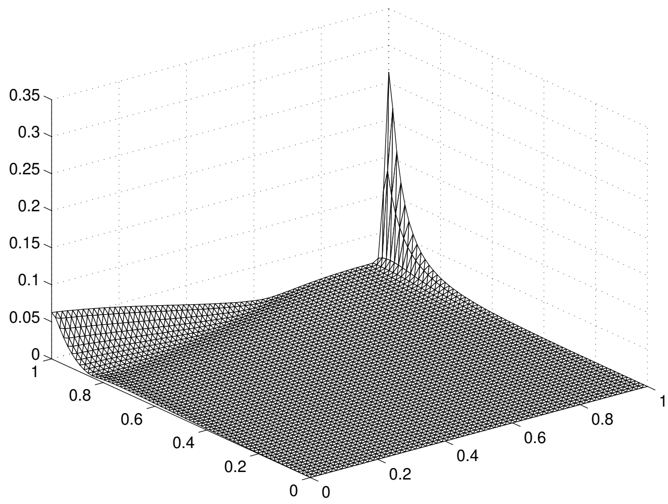
rate of convergence of $N_\psi(u_h)$ to $N_\psi(u)$ in the presence of a discontinuous primal solution u is attributed to the fact that while the residual terms r_h and r_h^- are large in the neighborhood of the discontinuities in u (cf. Fig. 6(a)), the local weighting terms in $\eta_\kappa^{(2)}$ involving $z_\delta - z_{\delta,h}$ are inactive in these regions of the computational domain, since the dual solution is smooth with 'small' support concentrated near the boundary $y = 1$, cf. Fig. 6(b).

The mesh design in our adaptive algorithm is based on the *optimal mesh criterion* with $\texttt{TOL} = 5.0 \times 10^{-5}$, and $\delta = h/4$. The background meshes for the primal and dual problems as well as the adaptively refined meshes resulting from these are depicted in Fig. 7. Here, Fig. 7(c) shows that most of the nodes in the adaptively refined mesh for the primal problem are concentrated near the outflow boundary where the quantity of interest, $N_\psi(u)$, is defined. It is interesting to note the lack of mesh refinement in the discontinuities as they enter the computational domain from $y = 0$ and $x = 0$. Any refinement indicator based on the residual *alone* will induce heavy (and clearly unnecessary) refinement along the discontinuities in the primal solution, cf. Fig. 8, for example. The adaptively refined mesh for the stabilization–dependent dual problem is, again, finer in the upper half of the computational domain, with most of the elements concentrated near the corners $(x, y) = (0, 1)$ and $(x, y) = (1, 1)$ due to the presence of boundary layers in the dual solution z_δ along Γ_- of thickness $|\mathbf{b} \cdot \boldsymbol{\mu}|\delta/|\mathbf{b}|$.

Finally, in Fig. 9 we show the performance of our adaptive algorithm with $\texttt{TOL} = 5.0 \times 10^{-6}$. The initial meshes are as shown in Fig. 7. It is important to note that even though the stabilization–dependent dual problem has been



(a)



(b)

Fig. 6. Example 2: (a) Scaled residual term $\|r_h\|_{L_2(\kappa)}/h_\kappa$; (b) Scaled weighting term $\|z_\delta - z_{\delta,h}\|_{L_2(\kappa)}/h_\kappa^2$; from [30]

solved numerically, our *a posteriori* error bound $\hat{\mathcal{E}}_{|\Omega|,\delta}$, based on the use of the numerically computed dual solution $\hat{z}_{\delta,h_d}$ and $P_h(\hat{z}_{\delta,h_d})$ in place of z_δ and $z_{\delta,h} = P_h z_\delta$, respectively, remains an upper bound on the true error in our approximation of $N_\psi(u)$. In Fig. 9, $N_\psi(e(\hat{\mathcal{E}}_{|\Omega|,\delta}))$ denotes the true error in the outward normal flux on the sequence of adaptively refined meshes which have been generated using the optimized mesh criterion. Here, Fig. 9 also shows

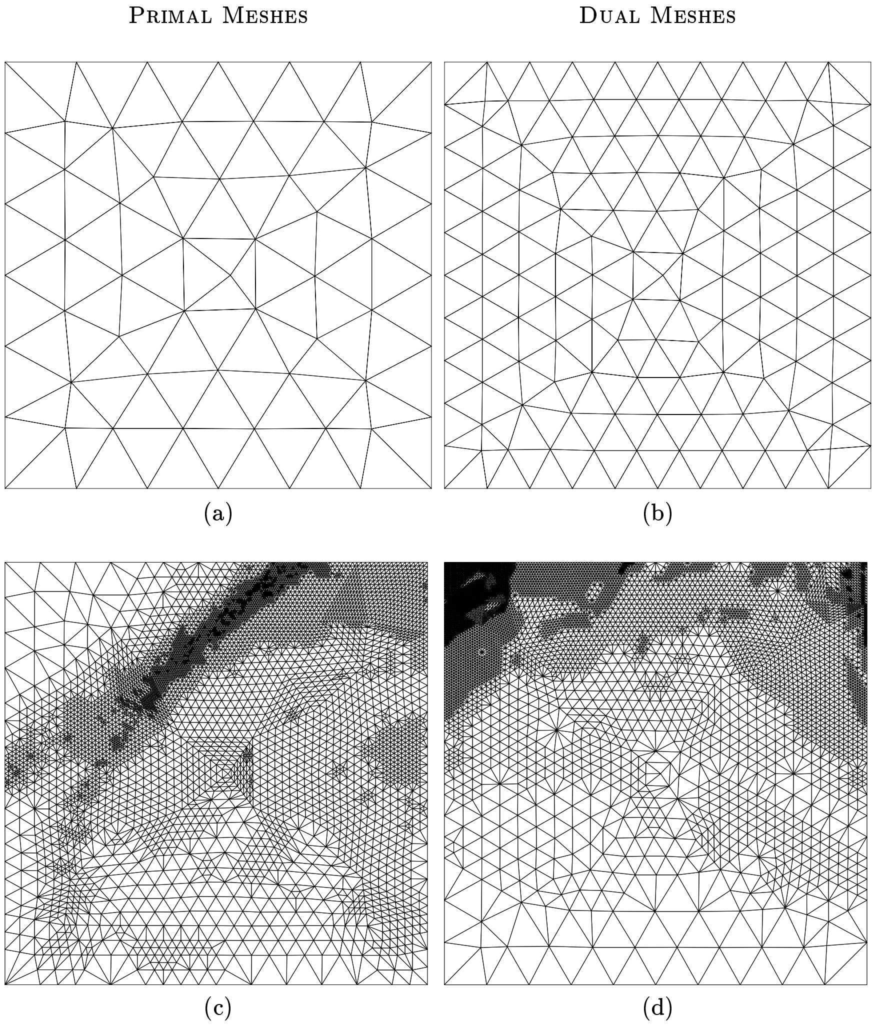


Fig. 7. Example 2: (a) Background mesh for the primal problem with 61 nodes and 96 elements; (b) Background mesh for stabilization–dependent dual problem with 137 nodes and 232 elements; (c) Primal mesh, resulting from the background mesh in (a) using the optimal mesh criterion for $\mathtt{TOL} = 5.0 \times 10^{-5}$, with 5996 nodes and 11827 elements ($|N_\psi(u) - N_\psi(u_h)| = 3.299 \times 10^{-6}$); (d) Dual mesh, resulting from the background mesh in (b) using the fixed fraction criterion, with 7631 nodes and 14341 elements

the true error in the outward normal flux on a sequence of adaptively refined meshes which have been constructed on a sequence of meshes using the energy refinement indicator $\|r_h\|_{L_2(\kappa)}$ on each element κ in $\mathcal{T}_h$ in conjunction with a fixed fraction strategy, cf. Fig. 8. Clearly, the latter approach is inferior:

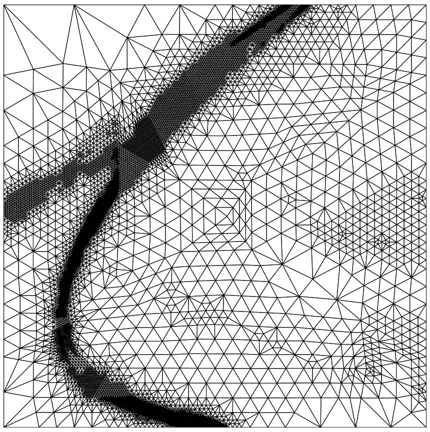


Fig. 8. Example 2: Mesh constructed using energy refinement indicator with 9147 nodes and 18113 elements ($|N_\psi(u-u_h)| = 4.778 \times 10^{-5}$)

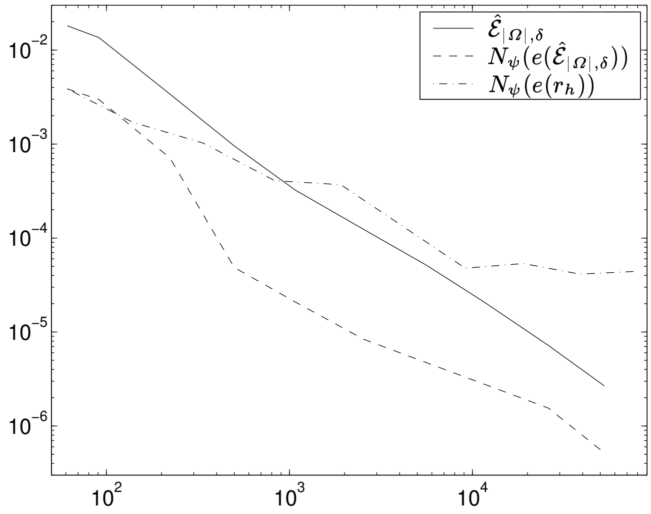


Fig. 9. Example 2: Performance of the adaptive algorithm for $\mathtt{TOL} = 5.0 \times 10^{-6}$ and $\delta = h/4$. $N_\psi(e(\hat{\mathcal{E}}_{|\Omega|,\delta}))$ denotes the error in the functional on a sequence of adaptively refined meshes using the stabilization–dependent dual problem; $\hat{\mathcal{E}}_{|\Omega|,\delta}$ is the corresponding Type I *a posteriori* error bound; $N_\psi(e(r_h))$ error denotes the error in the functional when the adaptive meshes are generated by using $\|r_h\|_{L_2(\kappa)}$ only as refinement indicator

stagnation of the error in the functional is observed in the course of the adaptive mesh refinement. The inferiority of the energy refinement indicator

$\|r_h\|_{L_2(\kappa)}$, $\kappa \in \mathcal{T}_h$, is due to the fact that it does not take into account the behavior of the solution to the dual problem.

In conclusion then, the use of a stabilization–dependent dual problem in the *a posteriori* error analysis of a stabilized Galerkin finite element method, such as the streamline–diffusion method, leads to sharper *a posteriori* error bounds than the standard dual problem. This is achieved through 'balancing' the magnitudes of the various terms in the error representation formula, which then leads to economical meshes when implemented into an adaptive algorithm. We have also seen that a Type I *a posteriori* error bound leads to more economical computational meshes than empirical residual-based refinement indicators (such as the energy refinement indicator $\|r_h\|_{L_2(\kappa)}$ that arises from a Type II error bound).

6 *hp*–DGFEM for systems of hyperbolic PDEs

Adaptive finite element methods that exploit both local polynomial degree variation (p-refinement) and local mesh subdivision (h-refinement) offer much greater flexibility and improved efficiency than mesh refinement algorithms which only incorporate h–refinement or p–refinement in isolation.

Since the early analytical paper of Gui and Babuška [23], the benefits of *hp*–version finite element methods have been clearly established for elliptic boundary value problems (see, for example, the monograph of Schwab [51]), particularly in the field of linear elasticity. The application of *hp*-version finite element methods to hyperbolic problems is less standard, although their potential in compressible gas dynamics has been demonstrated recently by Flaherty and collaborators; see [2], for example. The argument in favor of using an *hp*–version finite element method for the numerical solution of a hyperbolic equation rests on the observation that while solutions to these equations may exhibit local singularities and discontinuities, in large parts of the computational domain the solution is typically a real analytic function. Such large variations in the smoothness of the solution can be captured in a particularly simple and flexible manner by using a finite element method based on discontinuous piecewise polynomials, such as the discontinuous Galerkin finite element method (DGFEM). The method admits local variation of the degree of the approximating polynomial and meshes with an arbitrary number of interelement hanging nodes. In its lowest order form, with piecewise polynomials of degree 0, the discontinuous Galerkin finite element method becomes a finite volume method; therefore the *hp*–version of DGFEM can be thought of as a high-order extension of classical finite volume schemes. For a comprehensive review of discontinuous Galerkin methods, including a detailed historical survey, we refer to [12] and the references cited therein.

The aim of this section is to discuss the *a posteriori* error analysis of the *hp*–version of the discontinuous Galerkin finite element method; we refer to [9] for earlier work in this area. As in the previous section, we shall be concerned with the derivation of computable error bounds for target functionals of the solution; in Sections 6.1–6.3 we consider first–order symmetric linear hyperbolic systems; Section 6.4 is devoted to the extension of these ideas to nonlinear systems of hyperbolic conservation laws. The work presented here is based on the series of papers [33,35,54,55].

6.1 The model problem and its *hp*–DGFEM approximation

Let Ω be a bounded open polyhedral domain in $\mathbb{R}^n$, $n \geq 2$, and let Γ denote the union of open faces of Ω. We consider the boundary value problem (13); using the notation introduced in Section 4.4, we recall the following symmetric hyperbolic system: find $\mathbf{u} \in H(\mathcal{L}, \Omega)$ such that

$$\mathcal{L}\mathbf{u} \equiv \sum_{j=1}^{n} \frac{\partial}{\partial x_j}(\mathbf{B}_j \mathbf{u}) + \mathbf{C}\mathbf{u} = \mathbf{f} \quad \text{in } \Omega\,, \quad \mathbf{B}_-(\boldsymbol{\mu})\mathbf{u} = \mathbf{B}_-(\boldsymbol{\mu})\mathbf{g} \quad \text{on } \Gamma\,, \tag{32}$$

where $H(\mathcal{L}, \Omega) = \{\mathbf{v} \in [L_2(\Omega)]^m : \mathcal{L}\mathbf{v} \in [L_2(\Omega)]^m\}$ denotes the graph space of the partial differential operator $\mathcal{L}$ in $[L_2(\Omega)]^m$. Here, $\mathbf{B}_-(\boldsymbol{\mu})$ is the negative part of the matrix

$$\mathbf{B}(\boldsymbol{\mu}) \equiv \mathbf{B} \cdot \boldsymbol{\mu} = \mu_1 \mathbf{B}_1 + \ldots + \mu_n \mathbf{B}_n\,,$$

where $\boldsymbol{\mu}$ denotes the unit outward normal vector to the boundary Γ.

We begin by formulating the *hp*–version of the discontinuous Galerkin finite element method (*hp*–DGFEM, for short) for the numerical solution of (32). Let us suppose that $\mathcal{T}_h$ is a regular or 1-irregular subdivision of the computational domain Ω into disjoint open element domains κ such that $\bar{\Omega} = \cup_{\kappa \in \mathcal{T}_h} \bar{\kappa}$. By *1-irregular* we mean that an $(n-1)$–dimensional face of each element κ in $\mathcal{T}_h$ is allowed to contain at most one hanging (irregular) node — typically the barycenter of the face. Further, we shall suppose that the family of subdivisions $\mathcal{T}_h$ is shape–regular and that each $\kappa \in \mathcal{T}_h$ is a smooth bijective image of a fixed master element $\hat{\kappa}$, that is, $\kappa = F_\kappa(\hat{\kappa})$ for all $\kappa \in \mathcal{T}_h$, where $\hat{\kappa}$ is either the open unit simplex or the open unit hypercube in $\mathbb{R}^n$.

On the reference element $\hat{\kappa}$, with $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_n) \in \hat{\kappa}$ and a multi-index $\alpha \in \mathbb{N}^n$, we define spaces of polynomials of degree $p \geq 1$ as follows:

$$\mathcal{Q}_p = \operatorname{span}\{\hat{x}^\alpha : 0 \leq \alpha_i \leq p,\ \ 1 \leq i \leq n\}\,, \quad \mathcal{P}_p = \operatorname{span}\{\hat{x}^\alpha : 0 \leq |\alpha| \leq p\}\,.$$

With each element κ in $\mathcal{T}_h$ we associate an integer $p_\kappa \geq 1$; collecting the p_κ and F_κ in the vectors $\mathbf{p} = \{p_\kappa : \kappa \in \mathcal{T}_h\}$ and $\mathbf{F} = \{F_\kappa : \kappa \in \mathcal{T}_h\}$, respectively, we introduce the finite element space

$$\begin{aligned} S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F}) = \{\mathbf{v} \in [L_2(\Omega)]^m \ :\ & \mathbf{v}|_\kappa \circ F_\kappa \in [\mathcal{Q}_{p_\kappa}]^m \text{ if } F_\kappa^{-1}(\kappa) \text{ is} \\ & \text{the open unit hypercube and} \quad \mathbf{v}|_\kappa \circ F_\kappa \in [\mathcal{P}_{p_\kappa}]^m \\ & \text{if } F_\kappa^{-1}(\kappa) \text{ is the open unit simplex; } \kappa \in \mathcal{T}_h\}\,. \end{aligned} \tag{33}$$

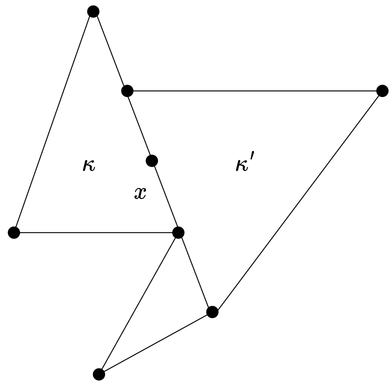


Fig. 10. A point x such that $x \in \partial\kappa$ and $x \in \partial\kappa'$

Assuming that $\mathcal{T}_h$ is a subdivision of Ω, we consider the broken Sobolev space $H^{\mathbf{s}}(\Omega, \mathcal{T}_h)$ of composite index $\mathbf{s}$ with nonnegative components s_κ, $\kappa \in \mathcal{T}_h$, defined by

$$H^{\mathbf{s}}(\Omega, \mathcal{T}_h) = \{\mathbf{v} \in [L_2(\Omega)]^m : \mathbf{v}|_\kappa \in [H^{s_\kappa}(\kappa)]^m \quad \forall \kappa \in \mathcal{T}_h\}\,.$$

If $s_\kappa = s \geq 0$ for all $\kappa \in \mathcal{T}_h$, we shall simply write $H^s(\Omega, \mathcal{T}_h)$.

For each element κ in $\mathcal{T}_h$, we denote by $\partial\kappa$ the union of $(n-1)$–dimensional open faces of κ. Given that $x \in \partial\kappa$, we label by $\boldsymbol{\mu}_\kappa(x)$ the unit outward normal vector to $\partial\kappa$ at x. With these conventions, we define $\mathbf{B}(\boldsymbol{\mu}_\kappa)$, $\mathbf{B}_-(\boldsymbol{\mu}_\kappa)$ and $\mathbf{B}_+(\boldsymbol{\mu}_\kappa)$ analogously to $\mathbf{B}(\boldsymbol{\mu})$, $\mathbf{B}_-(\boldsymbol{\mu})$ and $\mathbf{B}_+(\boldsymbol{\mu})$ in Section 4.4, respectively.

For each $\kappa \in \mathcal{T}_h$ and any $\mathbf{v} \in [H^1(\kappa)]^m$ we denote by $\mathbf{v}_\kappa^+$ the interior trace of $\mathbf{v}$ on $\partial\kappa$ (i.e., the trace of $\mathbf{v}$ on $\partial\kappa$ taken from within κ). Let us consider an element κ such that the set $\partial\kappa\backslash\Gamma$ is nonempty; then for each $x \in \partial\kappa\backslash\Gamma$ (with the exception of a set of $(n-1)$–dimensional measure zero) there exists a unique element κ', depending on the choice of x, such that $x \in \partial\kappa'$, cf. Fig. 10. Suppose that $\mathbf{v} \in H^1(\Omega, \mathcal{T}_h)$. If $\partial\kappa\backslash\Gamma$ is nonempty for some $\kappa \in \mathcal{T}_h$, then we define the outer trace $\mathbf{v}_\kappa^-$ of $\mathbf{v}$ on $\partial\kappa\backslash\Gamma$ relative to κ as the inner trace $\mathbf{v}_{\kappa'}^+$ relative to those elements κ' for which $\partial\kappa'$ has intersection with $\partial\kappa\backslash\Gamma$ of positive $(n-1)$–dimensional measure. Below, it will always be clear from the context which element κ in the subdivision $\mathcal{T}_h$ the quantities $\boldsymbol{\mu}_\kappa$, $\mathbf{v}_\kappa^+$ and $\mathbf{v}_\kappa^-$ correspond to; therefore, for the sake of notational simplicity, we shall suppress the letter κ in the subscript and write, respectively, $\boldsymbol{\mu}$, $\mathbf{v}^+$, and $\mathbf{v}^-$ instead.

For $\mathbf{v}, \mathbf{w} \in H^1(\Omega, \mathcal{T}_h)$, we define the bilinear form

$$\begin{aligned} B_{\mathrm{DG}}(\mathbf{w}, \mathbf{v}) = \sum_{\kappa \in \mathcal{T}_h} \int_\kappa \mathbf{w} \cdot \mathcal{L}^* \mathbf{v} \,\mathrm{d}x + \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa\backslash\Gamma} \mathcal{H}(\mathbf{w}^+, \mathbf{w}^-, \boldsymbol{\mu}) \cdot \mathbf{v}^+ \,\mathrm{d}\sigma \\ + \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa\cap\Gamma} \mathbf{B}_+(\boldsymbol{\mu}) \mathbf{w}^+ \cdot \mathbf{v}^+ \,\mathrm{d}\sigma\,, \end{aligned}$$

where $\mathcal{L}^*$ is the formal adjoint of $\mathcal{L}$ defined by

$$\mathcal{L}^*\mathbf{v} \equiv -(\mathbf{B}\cdot\nabla)\mathbf{v} + \mathbf{C}^T\mathbf{v}\,,$$

and $\mathcal{H}(\cdot,\cdot,\cdot)$ is a numerical flux function, assumed to be Lipschitz continuous, and such that:

(i) $\mathcal{H}$ is consistent; i.e., $\mathcal{H}(\mathbf{u},\mathbf{u},\boldsymbol{\mu})|_{\partial\kappa\setminus\Gamma} = \mathbf{B}(\boldsymbol{\mu})\mathbf{u}|_{\partial\kappa\setminus\Gamma}$ for all κ in $\mathcal{T}_h$;
(ii) $\mathcal{H}(\cdot,\cdot,\cdot)$ is conservative; i.e., $\mathcal{H}(\mathbf{u}^+,\mathbf{u}^-,\boldsymbol{\mu})|_{\partial\kappa\setminus\Gamma} = -\mathcal{H}(\mathbf{u}^-,\mathbf{u}^+,-\boldsymbol{\mu})|_{\partial\kappa\setminus\Gamma}$.

The choice of the numerical flux function is completely independent of the finite element space employed. For example, one may choose

$$\begin{aligned}\mathcal{H}(\mathbf{u}^+,\mathbf{u}^-,\boldsymbol{\mu}) &= \mathbf{B}_+(\boldsymbol{\mu})\mathbf{u}^+ + \mathbf{B}_-(\boldsymbol{\mu})\mathbf{u}^- \\ &= \frac{1}{2}\left(\mathbf{B}(\boldsymbol{\mu})\mathbf{u}^+ + \mathbf{B}(\boldsymbol{\mu})\mathbf{u}^-\right) - \frac{1}{2}|\mathbf{B}(\boldsymbol{\mu})|(\mathbf{u}^- - \mathbf{u}^+)\,,\end{aligned}$$

where $|\mathbf{B}(\boldsymbol{\mu})| = \mathbf{B}_+(\boldsymbol{\mu}) - \mathbf{B}_-(\boldsymbol{\mu})$. For $\mathbf{v}\in H^1(\Omega,\mathcal{T}_h)$, we consider the linear functional

$$\ell_{\mathrm{DG}}(\mathbf{v}) = \sum_{\kappa\in\mathcal{T}_h}\int_\kappa \mathbf{f}\cdot\mathbf{v}\;\mathrm{d}x - \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\cap\Gamma}\mathbf{B}_-(\boldsymbol{\mu})\,\mathbf{g}\cdot\mathbf{v}^+\;\mathrm{d}\sigma\,.$$

The *hp*–DGFEM for the numerical solution of the symmetric hyperbolic system (32) is defined as follows: find $\mathbf{u}_{\mathrm{DG}}\in S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$ such that

$$B_{\mathrm{DG}}(\mathbf{u}_{\mathrm{DG}},\mathbf{v}) = \ell_{\mathrm{DG}}(\mathbf{v}) \quad \forall\mathbf{v}\in S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})\,. \tag{34}$$

For the *a priori* error analysis of *hp*–DGFEM, including *hp*–optimal error bounds, we refer to [31,32].

6.2 *A posteriori* error analysis

Suppose that $\mathtt{TOL} > 0$ is a prescribed tolerance and $J(\cdot)$ is a linear functional defined on a linear space which contains $H(\mathcal{L},\Omega) + S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$. We shall consider the measurement problem concerned with computing a numerical approximation $\mathbf{u}_{\mathrm{DG}}$ from $S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$ such that

$$|J(\mathbf{u}) - J(\mathbf{u}_{\mathrm{DG}})| \le \mathtt{TOL}\,. \tag{35}$$

As in the previous section where we discussed stabilized finite element methods (see, also [30]), we proceed by deriving an *a posteriori* bound on the error between $J(\mathbf{u})$ and $J(\mathbf{u}_{\mathrm{DG}})$ and demanding that this bound does not exceed the prescribed tolerance.

Let us first define the following *dual* problem: find $\mathbf{z}$ in $H(\mathcal{L}^*,\Omega)$ such that

$$B_{\mathrm{DG}}(\mathbf{w},\mathbf{z}) = J(\mathbf{w}) \quad \forall\mathbf{w}\in H(\mathcal{L},\Omega)\,, \tag{36}$$

where $H(\mathcal{L}^*, \Omega)$ denotes the graph space of the adjoint operator $\mathcal{L}^*$ in $L_2(\Omega)$. It will be assumed throughout that (36) possesses a unique solution.

Our *a posteriori* bound on the error between $J(\mathbf{u})$ and $J(\mathbf{u}_{\mathrm{DG}})$ will be expressed in terms of the finite element residual $\mathbf{r}_{h,p}$ defined on $\kappa \in \mathcal{T}_h$ by

$$\mathbf{r}_{h,p}|_\kappa = (\mathbf{f} - \mathcal{L}\mathbf{u}_{\mathrm{DG}})|_\kappa \,,$$

which measures the extent to which $\mathbf{u}_{\mathrm{DG}}$ fails to satisfy the differential equation on the union of the elements κ in the mesh $\mathcal{T}_h$; thus we refer to $\mathbf{r}_{h,p}$ as the *internal residual.* Also, since $\mathbf{u}_{\mathrm{DG}}$ satisfies the boundary conditions only approximately, the vector $\mathbf{B}_-(\boldsymbol{\mu})(\mathbf{g} - \mathbf{u}_{\mathrm{DG}})$ need not be zero on Γ; thus for each element κ with $\partial\kappa \cap \Gamma$ nonempty, we define the *boundary residual* $\boldsymbol{\rho}_{h,p}$ by

$$\boldsymbol{\rho}_{h,p}|_{\partial\kappa\cap\Gamma} = \mathbf{B}_-(\boldsymbol{\mu})(\mathbf{u}^+_{\mathrm{DG}} - \mathbf{g})|_{\partial\kappa\cap\Gamma} \,.$$

If our finite element space consisted of continuous piecewise polynomial functions, as in the streamline–diffusion finite element method with weakly imposed boundary condition, these two residuals would be the only ones we'd encounter. Here, however, due to the fact that a discontinuous piecewise polynomial approximation is used, we need to introduce a third residual which measures the extent to which the numerical solution fails to satisfy continuity of the normal flux across element interfaces. Thus, on $\partial\kappa \setminus \Gamma$, we define the *flux residual*

$$\boldsymbol{\sigma}_{h,p}|_{\partial\kappa\setminus\Gamma} = \left(\mathbf{B}(\boldsymbol{\mu})\mathbf{u}^+_{\mathrm{DG}} - \mathcal{H}(\mathbf{u}^+_{\mathrm{DG}}, \mathbf{u}^-_{\mathrm{DG}}, \boldsymbol{\mu})\right)|_{\partial\kappa\setminus\Gamma} \,.$$

With these definitions we deduce from (34) the following *Galerkin orthogonality property* of the *hp*–DGFEM:

$$\begin{aligned} B_{\mathrm{DG}}(\mathbf{u} - \mathbf{u}_{\mathrm{DG}}, \mathbf{v}) = &\sum_{\kappa\in\mathcal{T}_h} (\mathbf{r}_{h,p}, \mathbf{v})_\kappa + \sum_{\kappa\in\mathcal{T}_h} (\boldsymbol{\sigma}_{h,p}, \mathbf{v}^+)_{\partial\kappa\setminus\Gamma} \\ &+ \sum_{\kappa\in\mathcal{T}_h} (\boldsymbol{\rho}_{h,p}, \mathbf{v}^+)_{\partial\kappa\cap\Gamma} = 0 \end{aligned} \tag{37}$$

for all $\mathbf{v}$ in $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$. Upon selecting $\mathbf{w} = \mathbf{u} - \mathbf{u}_{\mathrm{DG}}$ in (36), the linearity of $J(\cdot)$ and (37) give rise to the following (primal) error representation formula:

$$\begin{aligned} J(\mathbf{u}) - J(\mathbf{u}_{\mathrm{DG}}) &= J(\mathbf{u} - \mathbf{u}_{\mathrm{DG}}) \\ &= B_{\mathrm{DG}}(\mathbf{u} - \mathbf{u}_{\mathrm{DG}}, \mathbf{z}) = B_{\mathrm{DG}}(\mathbf{u} - \mathbf{u}_{\mathrm{DG}}, \mathbf{z} - \mathbf{z}_{h,p}) \\ &\equiv \mathcal{E}_\Omega(\mathbf{u}_{\mathrm{DG}}, h, \mathbf{p}, \mathbf{z} - \mathbf{z}_{hp}) = \sum_{\kappa\in\mathcal{T}_h} \eta_\kappa \,, \end{aligned} \tag{38}$$

where

$$\eta_\kappa = (\mathbf{r}_{h,p}, \mathbf{z} - \mathbf{z}_{h,p})_\kappa + (\boldsymbol{\sigma}_{h,p}, (\mathbf{z} - \mathbf{z}_{h,p})^+)_{\partial\kappa\setminus\Gamma} + (\boldsymbol{\rho}_{h,p}, (\mathbf{z} - \mathbf{z}_{h,p})^+)_{\partial\kappa\cap\Gamma} \tag{39}$$

for all $\mathbf{z}_{h,p}$ in $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$.

The measurement problem (35) is equivalent to demanding that the *stopping criterion*

$$|\mathcal{E}_{\Omega}(\mathbf{u}_{\mathrm{DG}}, h, \mathbf{p}, \mathbf{z} - \mathbf{z}_{h,p})| \leq \mathtt{TOL} \tag{40}$$

is satisfied, cf. equation (38).

As it stands, the left-hand side of (40) is not 'computable' since it involves the unknown analytical solution $\mathbf{z}$ to the dual problem (36). Let us therefore suppose that problem (36) has been solved numerically, using an hp–adaptive DGFEM for example, on a sequence of suitable 'dual' finite element spaces $S^{\tilde{\mathbf{p}}}(\Omega, \tilde{\mathcal{T}}_h, \tilde{\mathbf{F}})$, based on a 'dual' partition $\tilde{\mathcal{T}}_h$ and 'dual' polynomial vector $\tilde{\mathbf{p}}$, which may differ from the 'primal' partition $\mathcal{T}_h$ and 'primal' polynomial vector $\mathbf{p}$, respectively. We shall denote by $\tilde{\mathbf{z}}_{\mathrm{DG}}$ the corresponding approximation to the analytical dual solution $\mathbf{z}$ from the finite element space $S^{\tilde{\mathbf{p}}}(\Omega, \tilde{\mathcal{T}}_h, \tilde{\mathbf{F}})$. Further, $\mathbf{z}_{h,p}$ will denote the $L_2(\Omega)$ orthogonal projection of the numerically computed dual solution $\tilde{\mathbf{z}}_{\mathrm{DG}}$ onto the primal finite element space $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$ defined over the primal subdivision $\mathcal{T}_h$.

Let us now replace $\mathbf{z}$ by $\tilde{\mathbf{z}}_{\mathrm{DG}}$ in η_κ and decompose the error representation formula (38) into terms which are computable, namely those involving the numerical approximation $\tilde{\mathbf{z}}_{\mathrm{DG}}$ to the dual solution, and those that require the analytical dual solution $\mathbf{z}$: on noting that

$$\mathcal{E}_{\Omega}(\mathbf{u}_{\mathrm{DG}}, h, \mathbf{p}, \mathbf{z} - \mathbf{z}_{h,p}) = \mathcal{E}_{\Omega}(\mathbf{u}_{\mathrm{DG}}, h, \mathbf{p}, \tilde{\mathbf{z}}_{\mathrm{DG}} - \mathbf{z}_{h,p}) + \mathcal{E}_{\Omega}(\mathbf{u}_{\mathrm{DG}}, h, \mathbf{p}, \mathbf{z} - \tilde{\mathbf{z}}_{\mathrm{DG}}),$$

we deduce that

$$\begin{aligned} |J(\mathbf{u}) - J(\mathbf{u}_{\mathrm{DG}})| &\leq \mathcal{E}_{\mathrm{P}}^{\mathrm{loc}} + \mathcal{E}_{\mathrm{D}} \\ &\equiv \sum_{\kappa \in \mathcal{T}_h} |\tilde{\eta}_\kappa| + |\mathcal{E}_{\Omega}(\mathbf{u}_{\mathrm{DG}}, h, \mathbf{p}, \mathbf{z} - \tilde{\mathbf{z}}_{\mathrm{DG}})| \,, \end{aligned} \tag{41}$$

where $\tilde{\eta}_\kappa$ is defined analogously to η_κ, cf. (39), with $\mathbf{z}$ replaced by $\tilde{\mathbf{z}}_{\mathrm{DG}}$. Hence a sufficient condition for the validity of our original (noncomputable) stopping criterion (40) is that

$$\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}} + \mathcal{E}_{\mathrm{D}} \leq \mathtt{TOL} \,. \tag{42}$$

We emphasize here that the fundamental difference between the terms $\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}}$ and $\mathcal{E}_{\mathrm{D}}$ is that in the former the absolute value signs appear under the summation over the elements $\kappa \in \mathcal{T}_h$, while in the latter the absolute value sign is outside the sum. It has been shown through numerical experiments in [35] (cf. also [25–27]) that, with a sensible choice of the dual finite element space $S^{\tilde{\mathbf{p}}}(\Omega, \tilde{\mathcal{T}}_h, \tilde{\mathbf{F}})$, $\mathcal{E}_{\mathrm{D}}$ is typically an order of magnitude smaller than $\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}}$. Therefore, $\mathcal{E}_{\mathrm{D}}$ can be safely absorbed into $\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}}$ without compromising the reliability of the adaptive algorithm when the stopping criterion (40) is replaced by

$$\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}} \leq \mathtt{TOL} \,. \tag{43}$$

A bound of the form

$$|J(\mathbf{u}) - J(\mathbf{u}_{\mathrm{DG}})| \lesssim \mathcal{E}_{\mathrm{P}}^{\mathrm{loc}} \equiv \sum_{\kappa \in \mathcal{T}_h} |\tilde{\eta}_\kappa| \tag{44}$$

which explicitly involves the numerical solution $\tilde{\mathbf{z}}_{\mathrm{DG}}$ to the dual problem (through $\tilde{\eta}_\kappa$ defined above) is referred to as a *Type I* (cf. [35]) *a posteriori* error bound.

A possible alternative to using a Type I *a posteriori* error bound would be to completely eliminate the dual solution from the *a posteriori* error bound, through the use of the Cauchy–Schwarz inequality on each of the terms appearing in η_κ, choosing $\mathbf{z}_{h,p}$ as the finite element interpolant (or quasi-interpolant) of $\mathbf{z}$ from $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$, applying interpolation error estimates to bound $\|\mathbf{z} - \mathbf{z}_{h,p}\|_{L_2(\kappa)}$ and $\|\mathbf{z} - \mathbf{z}_{h,p}\|_{L_2(\partial\kappa)}$ in terms of powers of h and p and Sobolev seminorms of $\mathbf{z}$, and finally employing strong stability estimates of Sobolev seminorms of the dual solution in terms of Sobolev norms of the data for the dual problem. Hence, the resulting so-called *Type II a posteriori* error bound will not involve the dual solution. However, such an approach is not without practical drawbacks: quite apart from the fact that the derivation of a Type II *a posteriori* error bound requires a considerable amount of analytical work in order to establish strong stability results for the dual problem in various Sobolev norms and to determine the associated strong stability constants, typically, Type II bounds give a pessimistic over-estimate of the error. Consequently, when employed in an adaptive algorithm to approximate $J(\mathbf{u})$ by $J(\mathbf{u}_{\mathrm{DG}})$ to within a prescribed tolerance, a Type II *a posteriori* error bound may lead to uneconomical meshes; see, [30]. Furthermore, as indicated in Section 2, for hyperbolic problems at least, the presence of the dual solution (or its numerical approximation) as local weight function in the *a posteriori* error bound is highly desirable.

On the other hand, for the Type I error bound (44) to be an accurate approximation of (41) it is essential that $\mathcal{E}_{\mathrm{D}} \ll \mathcal{E}_{\mathrm{P}}^{\mathrm{loc}}$, and for this to be true it is necessary to ensure that the dual finite element space $S^{\tilde{\mathbf{p}}}(\Omega, \tilde{\mathcal{T}}_h, \tilde{\mathbf{F}})$ is sufficiently different from the primal finite element space $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$; for example, if $S^{\tilde{\mathbf{p}}}(\Omega, \tilde{\mathcal{T}}_h, \tilde{\mathbf{F}})$ is chosen to coincide with $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$ then $\mathbf{z}_{h,p} = \tilde{\mathbf{z}}_{\mathrm{DG}}$ and thereby $\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}} = 0$, so $0 < \mathcal{E}_{\mathrm{D}} \ll \mathcal{E}_{\mathrm{P}}^{\mathrm{loc}}$ cannot hold. How can then one select an appropriate dual finite element space $S^{\tilde{\mathbf{p}}}(\Omega, \tilde{\mathcal{T}}_h, \tilde{\mathbf{F}})$?

To answer this question, we begin by noting that

$$N(\cdot) = \mathcal{E}_\Omega(\mathbf{u}_{\mathrm{DG}}, h, \mathbf{p}, \cdot)$$

is a linear functional. Thereby,

$$\mathcal{E}_\Omega(\mathbf{u}_{\mathrm{DG}}, h, \mathbf{p}, \mathbf{z} - \tilde{\mathbf{z}}_{\mathrm{DG}}) = N(\mathbf{z}) - N(\tilde{\mathbf{z}}_{\mathrm{DG}})\,.$$

Now, given the tolerance TOL we select θ, $0 < \theta \ll 1$, split the tolerance

$$\mathtt{TOL} = (1-\theta)\,\mathtt{TOL} + \theta\,\mathtt{TOL}\,,$$

demand that

$$\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}} \leq (1-\theta)\,\mathtt{TOL} \qquad \text{and} \qquad \mathcal{E}_{\mathrm{D}} \leq \theta\,\mathtt{TOL}\,,$$

and aim to satisfy the first of these inequalities with near-equality; for then, both $0 < \mathcal{E}_{\mathrm{D}} \ll \mathcal{E}_{\mathrm{P}}^{\mathrm{loc}}$ and (42) will follow. Thus, it now remains to ensure that $\mathcal{E}_{\mathrm{D}} \leq \theta\,\mathtt{TOL}$; this in turn is equivalent to requiring that

$$|N(\mathbf{z}) - N(\tilde{\mathbf{z}}_{\mathrm{DG}})| \leq \theta\,\mathtt{TOL}\,, \tag{45}$$

which is a measurement problem, analogous to (35), — the problem that we had set out to solve, — except that (45) concerns the dual solution $\mathbf{z}$ while (35) involves the primal solution $\mathbf{u}$. To ensure the validity of (45) we may proceed by deriving an *a posteriori* error bound on $|N(\mathbf{z}) - N(\tilde{\mathbf{z}}_{\mathrm{DG}})|$; the corresponding error representation formula will involve the residuals associated with the numerical solution of the dual problem (36) and the analytical solution of the dual to the dual problem. As a Type I *a posteriori* error bound on $\mathcal{E}_{\mathrm{D}}$ based on such an error representation formula would contain the solution of the dual to the dual problem which would then in turn have to be approximated numerically, in [35] a cruder Type II *a posteriori* error bound on $\mathcal{E}_{\mathrm{D}}$ was employed so as to terminate the potentially infinite succession of mutually dual problems which would otherwise arise. The crudeness of the Type II bound on $\mathcal{E}_{\mathrm{D}}$ is of no concern: from the practical point of view there appears to be little advantage in performing reliable error control for $\mathcal{E}_{\mathrm{D}}$; the aim, when using the Type II bound on $\mathcal{E}_{\mathrm{D}}$, is merely to generate an adequate sequence of finite element approximations $\tilde{\mathbf{z}}_{\mathrm{DG}}$ to the dual solution $\mathbf{z}$ which can then be used to compute $\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}}$.

Thus, given a user–defined tolerance $\mathtt{TOL}$, our approach to the problem of designing an *hp*–finite element space $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$ such that

$$|J(\mathbf{u}) - J(\mathbf{u}_{\mathrm{DG}})| \leq \mathtt{TOL}\,, \tag{46}$$

is as follows. We use the Type I *a posteriori* error bound (42) to construct $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$ such that

$$\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}} \leq \mathtt{TOL} \tag{47}$$

and simply absorb $\mathcal{E}_{\mathrm{D}}$ into $\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}}$. If the stopping criterion (47) is not satisfied, then the elements are marked for refinement/derefinement; to this end, we employ the fixed fraction mesh refinement algorithm (cf. Section 5.2), with refinement and derefinement fractions set to 20% and 10%, respectively.

In the numerical experiments that follow, for the sake of computational simplicity, the dual finite element space $\tilde{S}^{\tilde{\mathbf{p}}}(\Omega, \tilde{\mathcal{T}}_h, \tilde{\mathbf{F}})$ that is used to compute the discontinuous Galerkin approximation $\tilde{\mathbf{z}}_{\mathrm{DG}}$ to $\mathbf{z}$ will be constructed using the same mesh as the one employed for $\mathbf{u}_{\mathrm{DG}}$, i.e., $\tilde{\mathcal{T}}_h \equiv \mathcal{T}_h$, with $\tilde{\mathbf{p}} = \mathbf{p} + \mathbf{1}$ (see [35], for a more general algorithm with $\tilde{\mathcal{T}}_h \neq \mathcal{T}_h$ and $\tilde{\mathbf{p}}$ unrelated to $\mathbf{p}$).

Once an element $\kappa \in \mathcal{T}_h$ has been flagged for refinement or derefinement, a decision must be made whether the local mesh size h_κ or the local degree p_κ of the approximating polynomial should be adjusted accordingly. The choice to perform either h–refinement/derefinement or p–refinement/derefinement is dependent on the local smoothness of the primal and dual solutions $\mathbf{u}$ and $\mathbf{z}$, respectively; cf. [2] and [35].

Let us first consider the case when an element has been flagged for refinement. Clearly, if $\mathbf{u}$ or $\mathbf{z}$ are locally smooth, then p–refinement will be more effective than h–refinement, since the error will be expected to decay quickly within the current element κ as p_κ is increased. On the other hand, if $\mathbf{u}$ and $\mathbf{z}$ have low regularity within the element κ, then h–refinement will be performed. Thus, regions in the computational domain where the primal or dual solution are locally non-smooth are isolated from smooth regions, thereby reducing the influence of singularities/discontinuities as well as making p–refinement more effective. To ensure that the desired level of accuracy is achieved efficiently, in [35] an automatic procedure was developed for deciding when to h– or p–refine, based on the smoothness–estimation strategy proposed by Ainsworth & Senior [3] in the context of norm control for second–order elliptic problems. For a review of various hp-adaptive strategies as well as a description of a new algorithm based on Sobolev index estimation *via* local Legendre expansions, we refer to [33].

If an element has been flagged for derefinement, then the strategy implemented here is to coarsen the mesh in low–error–regions where either the primal or dual solutions $\mathbf{u}$ and $\mathbf{z}$, respectively, are smooth and decrease the degree of the approximating polynomial in low–error–regions when both $\mathbf{u}$ and $\mathbf{z}$ are insufficiently regular, cf. [2] and [35].

6.3 Numerical experiments

The purpose of this section is to assess the performance of our hp–adaptive algorithm for the discontinuous Galerkin finite element method on a simple model problem.

Wave equation. Let us consider the one–dimensional wave equation

$$\frac{\partial^2 \phi}{\partial t^2} - c^2 \frac{\partial^2 \phi}{\partial x^2} = 0\,. \tag{48}$$

We note that by defining $x_1 \equiv x$ and $x_2 \equiv t$, the second–order partial differential equation (48) can be rewritten as the symmetric first–order system

$$\begin{bmatrix} 0 & -c \\ -c & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}_{x_1} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}_{x_2} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{49}$$

Table 8. Adaptive algorithm for the wave equation using h–refinement

Elements	DOF	$J(\mathbf{u}-\mathbf{u}_{\mathrm{DG}})$	$\sum_\kappa \tilde{\eta}_\kappa$	θ_1	$\sum_\kappa \lvert\tilde{\eta}_\kappa\rvert$	θ_2
32	256	0.1917e-02	0.2462e-02	1.28	0.3139e-02	1.64
47	376	0.1191e-02	0.1392e-02	1.17	0.2108e-02	1.77
74	592	-0.1384e-04	-0.1734e-04	1.25	0.4570e-03	33.01
116	928	-0.6897e-05	-0.6779e-05	0.98	0.2610e-03	37.84
200	1600	-0.3506e-05	-0.3407e-05	0.97	0.6041e-04	17.23
371	2968	-0.8052e-06	-0.9045e-06	1.12	0.2481e-04	30.81
611	4888	-0.7429e-07	-0.7133e-07	0.96	0.6982e-05	93.98
926	7408	-0.1920e-06	-0.1913e-06	1.00	0.4202e-05	21.89
1607	12856	0.3443e-07	0.3382e-07	0.98	0.1218e-05	35.39
2624	20992	-0.1646e-07	-0.1649e-07	1.00	0.6956e-06	42.25
4382	35056	0.4606e-08	0.4562e-08	0.99	0.2853e-06	61.95
7247	57976	0.4882e-09	0.4800e-09	0.98	0.1173e-06	240.29

where $u_1 = \partial\phi/\partial x_2$ and $u_2 = c\partial\phi/\partial x_1$. Here we let $\Omega = (0,1)\times(0,1/2)$ and $c = 1/2$, with

$$u_1(x_1,0) = \begin{cases} 1\,, & \text{for } 0.45 < x_1 \le 0.55\,, \\ 0\,, & \text{otherwise}\ , \end{cases}$$

and

$$u_2(x_1,0) = c\,e^{-100(x_1-1/2)^2}\ ,$$

for $0 \le x_1 \le 1$; this is a variant of the test problem considered in [29].

Our aim is to compute the weighted line integral of the first component of $\mathbf{u}$ along the upper side of the boundary Γ corresponding to $0 \le x_1 \le 1$, $x_2 = 1/2$, defined as

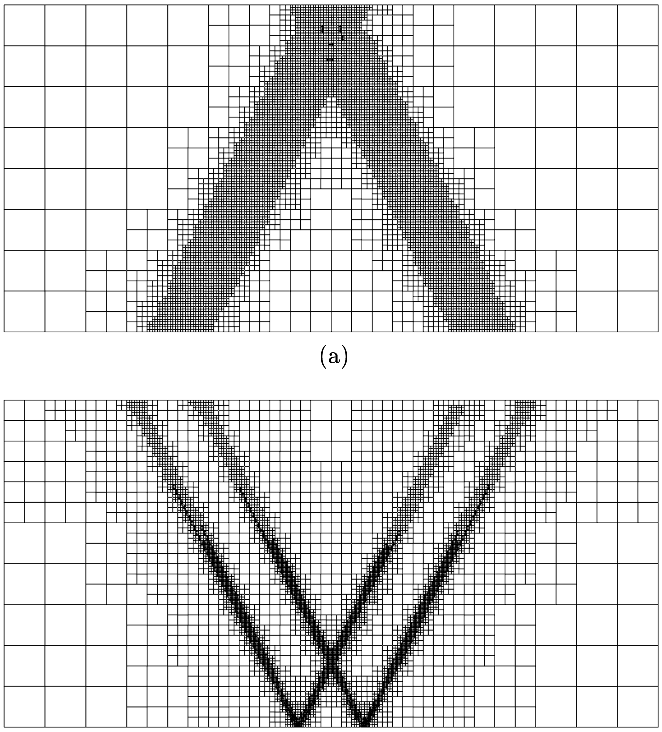
$$J(\mathbf{u}) = \int_0^1 \psi(x_1)\, u_1(x_1,1/2)\,\mathrm{d}x_1\,,$$

where the weight function ψ is defined by the Gaussian

$$\psi(x_1) = \mathrm{e}^{-\sigma^2(x_1-1/2)^2}\,, \qquad \text{for } 0 \le x_1 \le 1\,,$$

with $\sigma = 40$. The true value of the functional is $J(\mathbf{u}) \approx 0$.

In order to further highlight the advantages of designing an adaptive finite element algorithm based on the weighted refinement indicators $|\tilde{\eta}_\kappa|$ over traditional refinement strategies which do not require the solution of the dual problem (36), cf. Example 2 in Section 5.4, we first study the performance



(a)

(b)

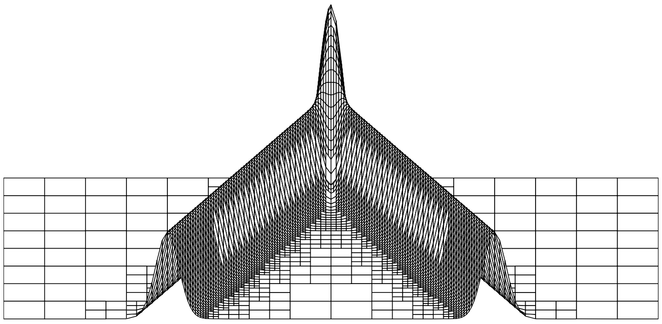
Fig. 11. Wave equation. (a) Mesh constructed using dual refinement indicator with 7247 elements and 57976 degrees of freedom ($|J(\mathbf{u}) - J(\mathbf{u}_{\mathrm{DG}})| = 4.882 \times 10^{-10}$); (b) Mesh constructed using the empirical refinement indicator with 9293 elements and 74344 degrees of freedom ($|J(\mathbf{u}) - J(\mathbf{u}_{\mathrm{DG}})| = 2.050 \times 10^{-8}$)

of our adaptive strategy with h–refinement. Here, we consider the empirical refinement indicator

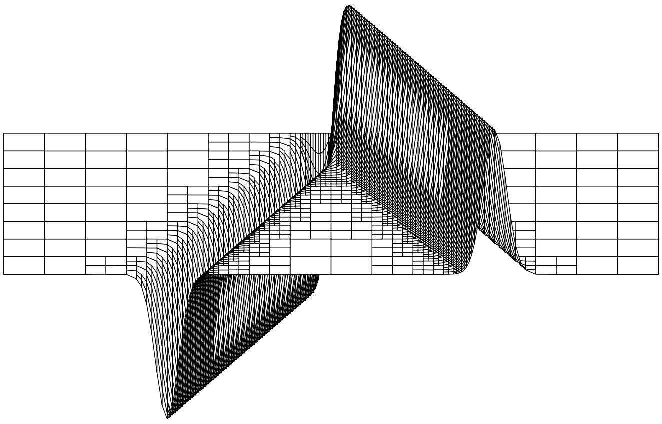
$$\eta_\kappa^{\text{empirical}} = \|h\mathbf{r}_{h,p}\|_{L_2(\kappa)} + \|h^{1/2}\boldsymbol{\sigma}_{h,p}\|_{L_2(\partial\kappa\setminus\Gamma)} + \|h^{1/2}\boldsymbol{\rho}_{h,p}\|_{L_2(\partial\kappa\cap\Gamma)} ,$$

which stems from a Type II *a posteriori* error analysis, cf. [35].

In Table 8 we illustrate the performance of the adaptive algorithm with $\mathbf{p} = \mathbf{1}$; here, we show the number of elements and degrees of freedom (DOF) in $S^{\mathbf{1}}(\Omega, \mathcal{T}_h, \mathbf{F})$, the true error in the functional $J(\mathbf{u} - \mathbf{u}_{\mathrm{DG}})$, the computed error representation formula $\tilde{\mathcal{E}}_\Omega \equiv \sum_{\kappa\in\mathcal{T}_h} \tilde{\eta}_\kappa$, the *a posteriori* error bound (44) and their respective effectivity indices θ_1 and θ_2. We see that the quality of the computed error representation formula $\tilde{\mathcal{E}}_\Omega$ is extremely good, in the sense that $\theta_1 = \tilde{\mathcal{E}}_\Omega / J(u - u_h)$ is close to one, even on relatively coarse meshes. Furthermore, we observe that the Type I *a posteriori* error bound is reliable, in the sense that the second effectivity index θ_2 overestimates the true error



(a)



(b)

Fig. 12. Dual solution for the wave equation: (a) $\mathbf{z}_1$; (b) $\mathbf{z}_2$

in the computed functional by a consistent factor as the finite element space $S^{\mathbf{1}}(\Omega, \mathcal{T}_h, \mathbf{F})$ is enriched.

In Fig. 11 we show the meshes generated using both the weighted refinement indicator $|\tilde{\eta}_\kappa|$ and the empirical refinement indicator $\eta_\kappa^{\mathrm{empirical}}$. From Fig. 11(a), we see that there is virtually no refinement in the regions of the computational domain where the discontinuities in $\mathbf{u}$ are located when $|\tilde{\eta}_\kappa|$ is employed. Indeed, most of the mesh refinement is concentrated in two regions upstream of the Gaussian weight function ψ. This is due to the presence of the weighting terms in $|\tilde{\eta}_\kappa|$, involving the difference between the approximate dual solution $\tilde{\mathbf{z}}_{\mathrm{DG}}$ and its projection $\mathbf{z}_{h,p}$ onto the primal finite element space, which multiply the computable residual terms involving the numerical solution $\mathbf{u}_{\mathrm{DG}}$, cf. (39) with $\mathbf{z}$ replaced by $\tilde{\mathbf{z}}_{\mathrm{DG}}$. We see from Fig. 12 that each component of the dual solution consists of two 'hills' transported backwards along the characteristics $x_1 \pm cx_2$. In contrast with this, when the weight-

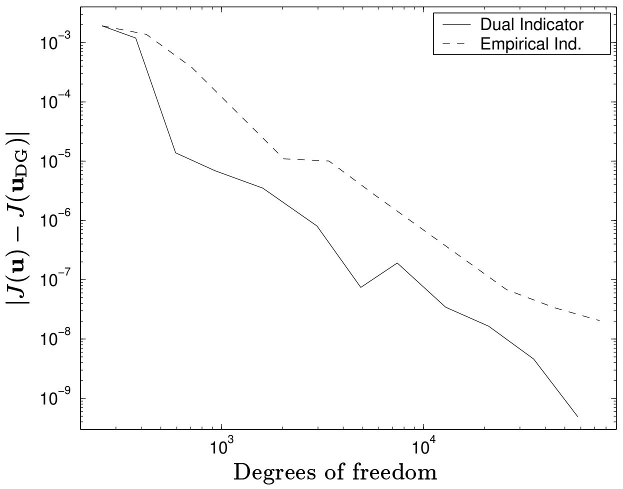


Fig. 13. Convergence of $|J(\mathbf{u}) - J(\mathbf{u}_{\mathrm{DG}})|$ using the dual and empirical refinement indicators for the wave equation

ing terms involving the dual solution are neglected, as in $\eta_\kappa^{\mathrm{empirical}}$, then the resulting mesh is concentrated in the vicinity of the discontinuities in the primal solution $\mathbf{u}$. In Fig. 13, we compare the true error in the computed functional $J(\cdot)$ using the two mesh refinement strategies. Here, we clearly observe the superiority of the Type I *a posteriori* error bound based on the weighted refinement indicators $|\tilde{\eta}_\kappa|$; on the final mesh the true error in the linear functional is almost two orders of magnitude smaller than $|J(\mathbf{u})-J(\mathbf{u}_{\mathrm{DG}})|$ computed on the sequence of meshes produced using $\eta_\kappa^{\mathrm{empirical}}$. This clearly indicates that a good numerical resolution of the discontinuities in this example is irrelevant for the accurate approximation of the functional of interest.

Let us now turn our attention to hp–adaptivity; in Table 9 we show the performance of the proposed adaptive finite element algorithm employing hp–refinement. Here, we again see that the quality of the computed error representation formula is extremely good, with $\theta_1 \approx 1$ even on very coarse meshes. Furthermore, the Type I *a posteriori* error bound (44) overestimates the true error in the computed functional by about an order of magnitude.

In Fig. 14 we plot $|J(\mathbf{u}) - J(\mathbf{u}_{\mathrm{DG}})|$, using both h– and hp–refinement against the square–root of the number of degrees of freedom on a linear–log scale. We see that after the initial transient, the error in the computed functional using hp–refinement becomes (on average) a straight line, thereby indicating exponential convergence of $J(\mathbf{u}_{\mathrm{DG}})$ to $J(\mathbf{u})$, despite the fact that $\mathbf{u}$ is only piecewise continuous; this occurs since $\mathbf{z}$ is a real analytic function on $\bar{\Omega}$. Fig. 14 also demonstrates the superiority of the adaptive hp–refinement strategy over the standard adaptive h–refinement algorithm. On the final mesh the true error between $J(\mathbf{u})$ and $J(\mathbf{u}_{\mathrm{DG}})$ using hp–refinement is al-

Table 9. Adaptive algorithm for the wave equation using hp–refinement

Elements	DOF	$J(\mathbf{u} - \mathbf{u}_{\mathrm{DG}})$	$\sum_\kappa \tilde{\eta}_\kappa$	θ_1	$\sum_\kappa \lvert\tilde{\eta}_\kappa\rvert$	θ_2
32	576	-0.3768e-03	-0.3727e-03	0.99	0.6615e-03	1.76
32	646	-0.1917e-03	-0.2558e-03	1.33	0.3495e-03	1.82
32	688	-0.3753e-04	-0.8638e-04	2.30	0.1490e-03	3.97
53	1248	0.1563e-04	0.1869e-04	1.20	0.2657e-04	1.70
86	2024	0.3341e-06	0.4492e-06	1.34	0.4501e-05	13.47
107	2868	0.1061e-06	0.9900e-07	0.93	0.7144e-06	6.73
119	3918	0.6372e-08	0.1900e-08	0.30	0.1401e-06	21.99
140	5486	0.3203e-08	0.3051e-08	0.95	0.2866e-07	8.95
203	9152	0.4293e-08	0.3655e-08	0.85	0.1476e-07	3.44
272	12648	-0.3332e-09	-0.3691e-09	1.11	0.1627e-08	4.88
344	17512	-0.3862e-11	-0.3862e-11	1.00	0.6570e-10	17.01
401	21616	0.6639e-12	0.6508e-12	0.98	0.9235e-11	13.91
464	27202	0.3406e-13	0.3477e-13	1.02	0.1799e-11	52.81

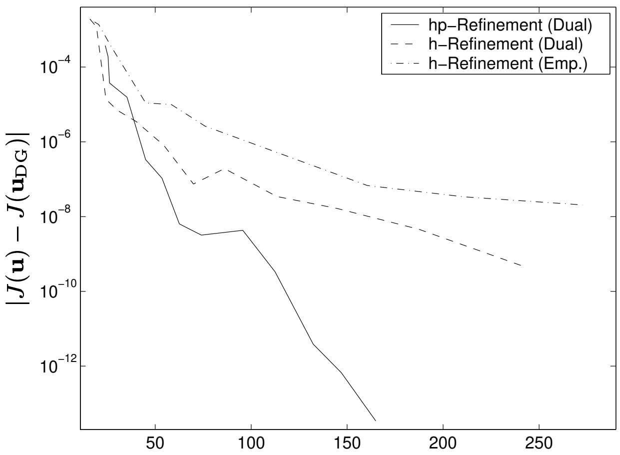


Fig. 14. Comparison between h– and hp–adaptive mesh refinement for the wave equation

most 6 orders of magnitude smaller than the corresponding quantity when h–refinement is employed alone.

Figs. 15 and 16 depict the primal mesh after 7 and 11 adaptive mesh refinement steps, respectively. For clarity, we show the h–mesh alone, as well

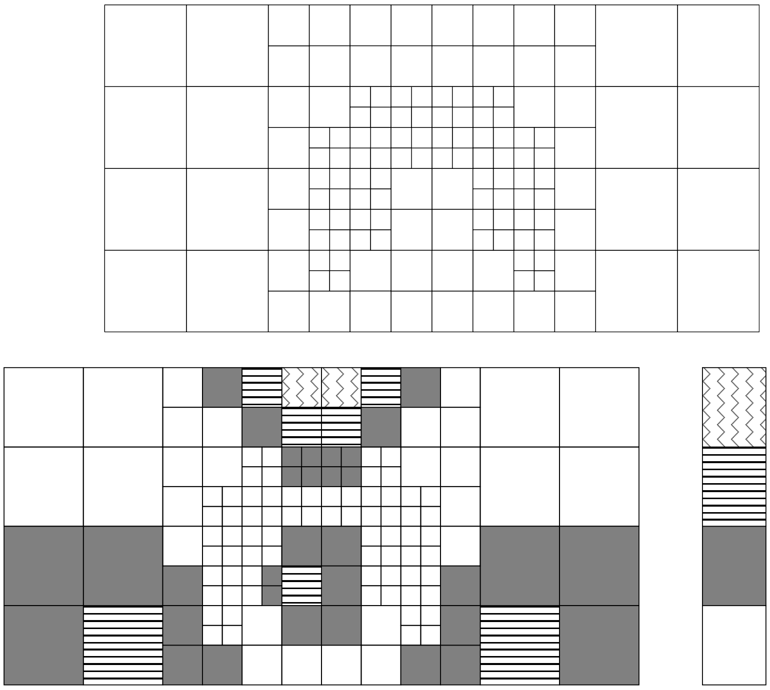


Fig. 15. Wave equation. h– and hp–meshes after 7 refinements, with 140 elements and 5486 degrees of freedom; here, $|J(\mathbf{u}) - J(\mathbf{u}_{\mathrm{DG}})| = 3.203 \times 10^{-9}$

as the corresponding distribution of the polynomial degree on this mesh and the percentage of elements with that degree. We see that some h–refinement of the primal mesh has occurred in the region of the computational domain where the two 'hills' in the dual solution $\mathbf{z}$, cf. Fig. 12, are located. In particular, the h–mesh is slightly finer in the regions where the discontinuities in the primal solution $\mathbf{u}$ intersect with these localized structures in $\mathbf{z}$. Once the h–mesh has adequately captured the structure of the primal and dual solutions, the hp–adaptive algorithm performed p–refinement to accurately resolve the Gaussians present in the dual solution.

6.4 Nonlinear hyperbolic problems

The purpose of this section is to develop the *a posteriori* error analysis of the discontinuous Galerkin finite element method for systems of nonlinear hyperbolic conservation laws.

Given a bounded open polyhedral domain Ω in $\mathbb{R}^n$, $n \geq 1$, with boundary $\partial\Omega$, we consider the following problem: find $\mathbf{u} : \Omega \to \mathbb{R}^m$, $m \geq 1$, such that

$$\operatorname{div} \mathcal{F}(\mathbf{u}) = 0 \quad \text{in } \Omega\,, \tag{50}$$

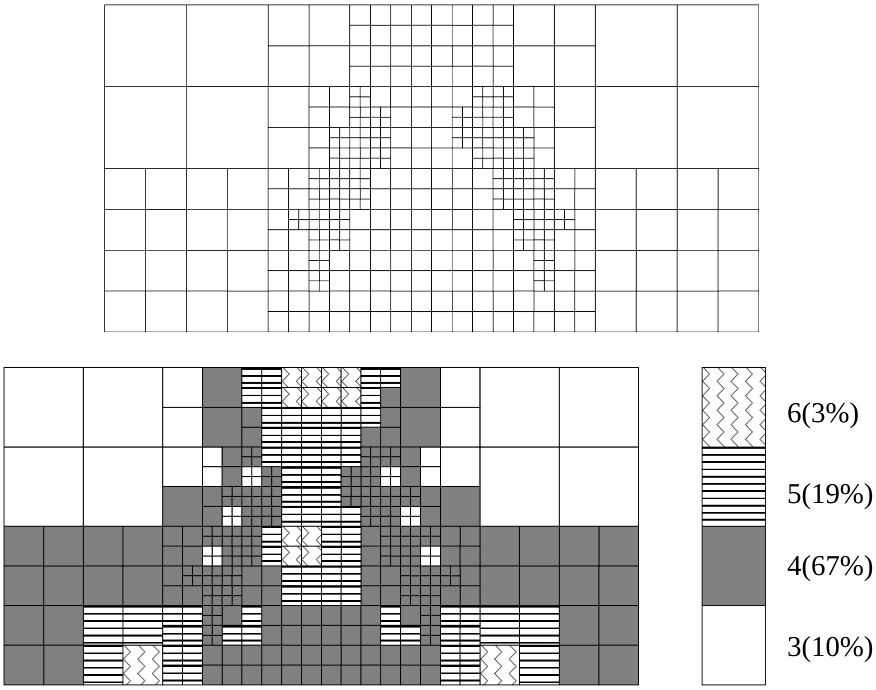


Fig. 16. Wave equation. h– and hp–meshes after 11 refinements, with 401 elements and 21616 degrees of freedom; here, $|J(\mathbf{u}) - J(\mathbf{u}_{\mathrm{DG}})| = 6.639 \times 10^{-13}$

where, $\mathcal{F} : \mathbb{R}^m \to \mathbb{R}^{m \times n}$ is continuously differentiable. We assume that the system of conservation laws (50) may be supplemented by appropriate initial/boundary conditions. For example, assuming that

$$B(\mathbf{u}, \boldsymbol{\nu}) := \sum_{i=1}^{n} \nu_i \nabla_{\mathbf{u}} \mathcal{F}_i(\mathbf{u})$$

has m real eigenvalues and a complete set of linearly independent eigenvectors for all $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n) \in \mathbb{R}^n$; then at inflow/outflow boundaries, we require that

$$B^-(\mathbf{u}, \boldsymbol{\mu})(\mathbf{u} - \mathbf{g}) = \mathbf{0}\,,$$

where $\boldsymbol{\mu}$ denotes the unit outward normal vector to $\partial\Omega$, $B^-(\mathbf{u}, \boldsymbol{\mu})$ is the negative part of $B(\mathbf{u}, \boldsymbol{\mu})$ and $\mathbf{g}$ is a (given) real–valued vector function.

To formulate the discontinuous Galerkin finite element approximation of (50), with $p_\kappa \in \mathbb{N}$ for $\kappa \in \mathcal{T}_h$ we define the finite element space

$$\begin{aligned} S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F}) = \{\mathbf{v} \in [L_2(\Omega)]^m \, : \, & \mathbf{v}|_\kappa \circ F_\kappa \in [\mathcal{Q}_{p_\kappa}]^m \text{ if } F_\kappa^{-1}(\kappa) \text{ is} \\ & \text{the open unit hypercube and} \quad \mathbf{v}|_\kappa \circ F_\kappa \in [\mathcal{P}_{p_\kappa}]^m \\ & \text{if } F_\kappa^{-1}(\kappa) \text{ is the open unit simplex; } \kappa \in \mathcal{T}_h\}\,, \end{aligned}$$

where

$$\mathcal{Q}_p = \text{span}\left\{\hat{x}^\alpha : 0 \leq \alpha_i \leq p, \ \ 1 \leq i \leq n\right\}, \quad \mathcal{P}_p = \text{span}\left\{\hat{x}^\alpha : 0 \leq |\alpha| \leq p\right\},$$

cf. (33). Given that $\mathbf{v} \in \left[H^1(\kappa)\right]^m$ for each $\kappa \in \mathcal{T}_h$, we denote by $\mathbf{v}^+$ (respectively, $\mathbf{v}^-$) the interior (respectively, exterior) trace of $\mathbf{v}$ on $\partial\kappa$.

The *hp*–DGFEM approximation of (50) is defined as follows: find $\mathbf{u}_{\mathrm{DG}}$ in $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$ such that

$$\sum_{\kappa\in\mathcal{T}_h}\left\{-\int_\kappa \mathcal{F}(\mathbf{u}_{\mathrm{DG}})\cdot\nabla\mathbf{v}\,\mathrm{d}x + \int_{\partial\kappa}\mathcal{H}(\mathbf{u}_{\mathrm{DG}}^+, \mathbf{u}_{\mathrm{DG}}^-, \boldsymbol{\mu}_\kappa)\,\mathbf{v}^+\,\mathrm{d}\sigma \right.$$
$$\left. + \int_\kappa \varepsilon\nabla\mathbf{u}_{\mathrm{DG}}\cdot\nabla\mathbf{v}\,\mathrm{d}x\right\} = 0 \qquad (51)$$

for all $\mathbf{v} \in S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$; cf. [26,27,37], for example. Here, $\mathcal{H}(\cdot,\cdot,\cdot)$ denotes a *numerical flux* function, assumed to be Lipschitz continuous, consistent and conservative. As in the linear case, the choice of the numerical flux function is completely independent of the finite element space employed; in the numerical experiments we employ the (local) Lax–Friedrichs flux function. In (51), ε denotes the coefficient of artificial viscosity defined by

$$\varepsilon|_\kappa = C_\varepsilon\left(\frac{h_\kappa}{p_\kappa}\right)^{2-\beta}\left|\,\mathrm{div}\mathcal{F}(\mathbf{u}_{\mathrm{DG}}|_\kappa)\right|, \qquad (52)$$

where C_ε is a positive constant and $0 < \beta < 1/2$; see [37]. For elements $\kappa \in \mathcal{T}_h$ whose boundary intersects $\partial\Omega$, we replace $\mathbf{u}_{\mathrm{DG}}^-$ by appropriate boundary/initial conditions on $\partial\kappa \cap \partial\Omega$.

In this section, we focus on a measurement problem concerned with controlling the error in the approximation of a given linear or nonlinear target functional $J(\cdot)$. Quantities of real or physical interest include the drag and lift coefficients of a body (such as an aircraft wing) immersed into an inviscid fluid governed by the compressible Euler equations of gas dynamics, cf. Section 4.2. Writing S to denote the surface of the body (or, rather, the contour of its two-dimensional cross-section) and $\bar{l}$ its chord length, these target functionals are defined by

$$J_{\mathrm{drag}}(\mathbf{u}) = \frac{2}{\bar{l}\bar{\rho}|\bar{\mathbf{v}}|^2}\int_S(\psi_d\cdot\boldsymbol{\mu})\,p\,\mathrm{d}s, \qquad J_{\mathrm{lift}}(\mathbf{u}) = \frac{2}{\bar{l}\bar{\rho}|\bar{\mathbf{v}}|^2}\int_S(\psi_l\cdot\boldsymbol{\mu})\,p\,\mathrm{d}s,$$

respectively. Here, $\boldsymbol{\mu}$ is the unit outward normal to S; $\bar{\mathbf{v}}$ and $\bar{\rho}$ denote the reference (or free-stream) velocity and density, respectively, and

$$\psi_d = \begin{pmatrix}\cos\alpha & -\sin\alpha\\ \sin\alpha & \cos\alpha\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix}, \qquad \psi_l = \begin{pmatrix}\cos\alpha & -\sin\alpha\\ \sin\alpha & \cos\alpha\end{pmatrix}\begin{pmatrix}0\\1\end{pmatrix},$$

where α denotes the angle of attack. We note that since the pressure p is derived from the conserved variables $(\rho, \rho u, \rho v, \rho E)$, both target functionals

$J_{\mathrm{drag}}(\cdot)$ and $J_{\mathrm{lift}}(\cdot)$ are nonlinear. Other examples of $J(\cdot)$ include the local mean value of the field or its flux through the outflow boundary of the computational domain Ω, and the point evaluation of a component of $\mathbf{u}$ in Ω.

Before we embark on our *a posteriori* error analysis, we shall introduce some notation. Assuming that the functional of interest $J(\cdot)$ is differentiable, we write $\bar{J}(\cdot;\cdot)$ to denote the mean value linearization of $J(\cdot)$ defined by

$$\begin{aligned}\bar{J}(\mathbf{u},\mathbf{u}_{\mathrm{DG}};\mathbf{u}-\mathbf{u}_{\mathrm{DG}}) &= J(\mathbf{u}) - J(\mathbf{u}_{\mathrm{DG}}) \\ &= \int_0^1 J'[\theta\mathbf{u} + (1-\theta)\mathbf{u}_{\mathrm{DG}}](\mathbf{u}-\mathbf{u}_{\mathrm{DG}})\,\mathrm{d}\theta\,, \qquad (53)\end{aligned}$$

where $J'[\mathbf{w}](\cdot)$ denotes the Gateaux derivative of $J(\cdot)$ evaluated at some $\mathbf{w}$ in V. Here, V is some suitably chosen function space such that $S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F}) \subset V$. Analogously, letting $\mathcal{N}(\mathbf{u}_{\mathrm{DG}},\mathbf{v})$ denote the left–hand side of (51), we denote by $\mathcal{M}(\mathbf{u},\mathbf{u}_{\mathrm{DG}};\cdot,\cdot)$ the mean value linearization of $\mathcal{N}(\cdot,\cdot)$ defined by

$$\begin{aligned}\mathcal{M}(\mathbf{u},\mathbf{u}_{\mathrm{DG}};\mathbf{u}-\mathbf{u}_{\mathrm{DG}},\mathbf{v}) &= \mathcal{N}(\mathbf{u},\mathbf{v}) - \mathcal{N}(\mathbf{u}_{\mathrm{DG}},\mathbf{v}) \\ &= \int_0^1 \mathcal{N}'_{\mathbf{u}}[\theta\mathbf{u} + (1-\theta)\mathbf{u}_{\mathrm{DG}}](\mathbf{u}-\mathbf{u}_{\mathrm{DG}},\mathbf{v})\,\mathrm{d}\theta \qquad (54)\end{aligned}$$

for all $\mathbf{v}$ in V. Here, $\mathcal{N}'_{\mathbf{u}}[\mathbf{w}](\cdot,\mathbf{v})$ denotes the Gateaux derivative of $\mathbf{u} \mapsto \mathcal{N}(\mathbf{u},\mathbf{v})$, for $\mathbf{v} \in V$ fixed, at some $\mathbf{w}$ in V. The linearization defined in (54) is only a *formal* calculation, in the sense that $\mathcal{N}'_{\mathbf{u}}[\mathbf{w}](\cdot,\cdot)$ may not in general exist. Instead, a suitable approximation to $\mathcal{N}'_{\mathbf{u}}[\mathbf{w}](\cdot,\cdot)$ must be determined, for example, by computing appropriate finite difference quotients of $\mathcal{N}(\cdot,\cdot)$, cf. [26]. For the proceeding analysis, we *assume* that the linearization (54) is well–defined. Under this hypothesis, we introduce the following *dual* problem: find $\mathbf{z} \in V$ such that

$$\mathcal{M}(\mathbf{u},\mathbf{u}_{\mathrm{DG}};\mathbf{w},\mathbf{z}) = \bar{J}(\mathbf{u},\mathbf{u}_{\mathrm{DG}};\mathbf{w}) \quad \forall \mathbf{w} \in V\,. \qquad (55)$$

We assume that (55) possesses a unique solution; of course, the validity of this assumption depends on both the definition of $\mathcal{M}(\mathbf{u},\mathbf{u}_{\mathrm{DG}};\cdot,\cdot)$ and the choice of the target functional under consideration, cf. [26]. For the proceeding error analysis, we must therefore *assume* that the dual problem (55) is well–posed. We then have the following result.

Theorem 3. *Let* $\mathbf{u}$ *and* $\mathbf{u}_{\mathrm{DG}}$ *denote the solutions to (50) and (51), respectively, and suppose that the dual problem (55) is well–posed. Then,*

$$J(\mathbf{u}) - J(\mathbf{u}_{\mathrm{DG}}) = \mathcal{E}_\Omega \equiv \sum_{\kappa\in\mathcal{T}_h} \eta_\kappa\,, \qquad (56)$$

where

$$\begin{aligned}\eta_\kappa = \int_\kappa \mathbf{r}_{h,p}\cdot(\mathbf{z}-\mathbf{z}_{h,p})\,\mathrm{d}x + \int_{\partial\kappa} \boldsymbol{\sigma}_{h,p}\cdot(\mathbf{z}-\mathbf{z}_{h,p})^+\,\mathrm{d}\sigma \\ - \int_\kappa \varepsilon\nabla\mathbf{u}_{\mathrm{DG}}\cdot\nabla(\mathbf{z}-\mathbf{z}_{h,p})\,\mathrm{d}x\end{aligned}$$

for all $\mathbf{z}_{h,p}$ *in* $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$. *Here,*

$$\mathbf{r}_{h,p}|_\kappa = -\mathrm{div}\mathcal{F}(\mathbf{u}_{\mathrm{DG}}) \quad \textit{and} \quad \boldsymbol{\sigma}_{h,p}|_\kappa = \mathcal{F}(\mathbf{u}_{\mathrm{DG}}) \cdot \mathbf{n}_\kappa - \mathcal{H}(\mathbf{u}_{\mathrm{DG}}^+, \mathbf{u}_{\mathrm{DG}}^-, \mathbf{n}_\kappa)$$

denote the internal and boundary finite element residuals, respectively, defined on each $\kappa \in \mathcal{T}_h$.

Proof. Choosing $\mathbf{w} = \mathbf{u} - \mathbf{u}_{\mathrm{DG}}$ in (55), recalling the linearization performed in (53), and exploiting the Galerkin orthogonality property $\mathcal{N}(\mathbf{u}, \mathbf{v}) - \mathcal{N}(\mathbf{u}_{\mathrm{DG}}, \mathbf{v}) = 0$ for all $\mathbf{v}$ in $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$, we get

$$\begin{aligned} J(\mathbf{u}) - J(\mathbf{u}_{\mathrm{DG}}) &= \bar{J}(\mathbf{u}, \mathbf{u}_{\mathrm{DG}}; \mathbf{u} - \mathbf{u}_{\mathrm{DG}}) = \mathcal{M}(\mathbf{u}, \mathbf{u}_{\mathrm{DG}}; \mathbf{u} - \mathbf{u}_{\mathrm{DG}}, \mathbf{z}) \\ &= \mathcal{M}(\mathbf{u}, \mathbf{u}_{\mathrm{DG}}; \mathbf{u} - \mathbf{u}_{\mathrm{DG}}, \mathbf{z} - \mathbf{z}_{h,p}) = -\mathcal{N}(\mathbf{u}_{\mathrm{DG}}, \mathbf{z} - \mathbf{z}_{h,p}) \end{aligned}$$

for all $\mathbf{z}_{h,p}$ in $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$. Equation (56) now follows by employing the divergence theorem.

Based on the general error representation formula derived in Theorem 3, *a posteriori* error estimates bounding the error in the computed functional $J(\cdot)$ may be deduced. Again, we shall confine ourselves to considering a Type I *a posteriori* error bound; it is a straightforward consequence of the error representation formula stated in the previous theorem.

Corollary 1. *Under the assumptions of Theorem 3, we have*

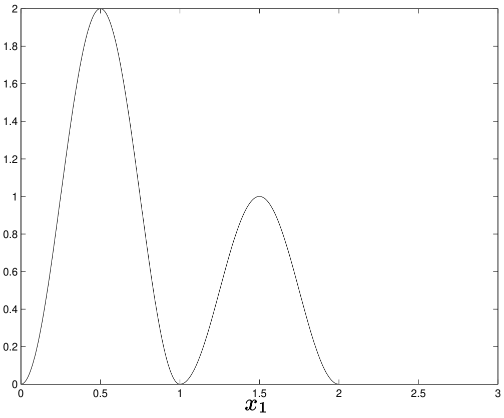
$$|J(\mathbf{u}) - J(\mathbf{u}_{\mathrm{DG}})| \leq \mathcal{E}_{|\Omega|} \equiv \sum_{\kappa \in \mathcal{T}_h} |\eta_\kappa| \,. \tag{57}$$

Proof. The error bound (57) follows from (56) by application of the triangle inequality.

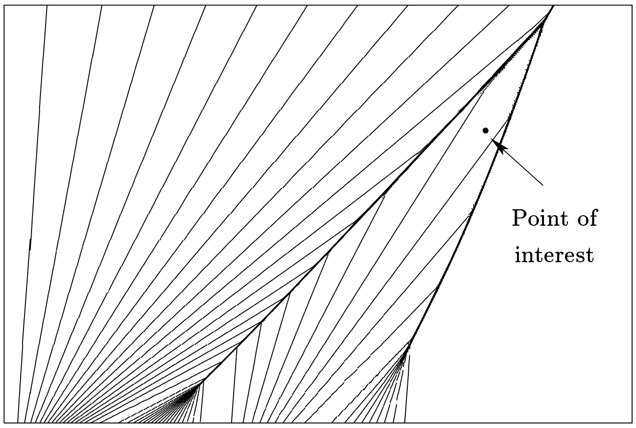
We end this section by noting that for nonlinear hyperbolic conservation laws and/or nonlinear target functionals $J(\cdot)$, both the error representation formula (56) and the Type I *a posteriori* error bound (57) depend on the unknown analytical solution to the primal and dual problems. Thus, in order to render these quantities computable, both $\mathbf{u}$ and $\mathbf{z}$ must be replaced by suitable approximations. To this end, the linearizations leading to $\mathcal{M}(\mathbf{u}, \mathbf{u}_{\mathrm{DG}}; \cdot, \cdot)$ and $\bar{J}(\mathbf{u}, \mathbf{u}_{\mathrm{DG}}; \cdot)$ are performed about $\mathbf{u}_{\mathrm{DG}}$ and the dual solution $\mathbf{z}$ is replaced by a discontinuous Galerkin approximation $\hat{\mathbf{z}}_{\mathrm{DG}}$ computed on the same mesh $\mathcal{T}_h$ as for the primal solution $\mathbf{u}_{\mathrm{DG}}$, but with a higher degree polynomial, i.e., $\hat{\mathbf{z}}_{\mathrm{DG}} \in S^{\hat{\mathbf{p}}}(\Omega, \mathcal{T}_h, \mathbf{F})$, $\hat{\mathbf{p}} = \mathbf{p} + \mathbf{p}_{\mathrm{INC}}$; in Section 6.5, we set $\mathbf{p}_{\mathrm{INC}} = \mathbf{1}$.

6.5 Numerical experiments

In this section we present some numerical examples to assess the performance of our *hp*–adaptive algorithm for the discontinuous Galerkin finite element method.



(a)



(b)

Fig. 17. Burgers' problem: (a) Initial condition; (b) Solution isolines

Burgers' equation. In this example, we consider the one–dimensional unsteady inviscid Burgers' equation for the scalar variable $\mathbf{u} \equiv u$; i.e., writing x_2 to denote time, we have

$$u_{x_2} + \left((1/2)\,u^2\right)_{x_1} = 0\,, \tag{58}$$

on the (space–time) domain $\Omega = (0,3) \times (0,2)$, subject to the initial condition

$$u(x,0) = \begin{cases} 2\,\sin^2(\pi x)\,, & 0 \le x \le 1\,, \\ \sin^2(\pi x)\,, & 1 \le x \le 2\,, \\ 0\,, & 2 \le x \le 3\,, \end{cases} \tag{59}$$

Table 10. Adaptive algorithm for Burgers' problem

Elements	DOF	$J(u-u_{\rm DG})$	$\sum_\kappa \tilde{\eta}_\kappa$	θ_1	$\sum_\kappa \lvert\tilde{\eta}_\kappa\rvert$	θ_2
24	216	0.5010e-01	0.1033	2.06	0.1913	3.82
36	324	-0.9031e-01	0.1096	-1.21	0.2058	2.28
81	736	0.1175e-01	-0.4193e-01	-3.57	0.1614	13.73
120	1150	-0.1961e-01	-0.1740e-01	0.89	0.9974e-01	5.09
207	2096	-0.9130e-02	-0.6216e-02	0.68	0.4372e-01	4.79
291	3211	0.7391e-03	0.8754e-03	1.18	0.8192e-02	11.08
336	4279	-0.3173e-03	-0.5243e-03	1.65	0.3101e-02	9.77
420	6227	-0.6514e-04	-0.4742e-04	0.73	0.8027e-03	12.32
525	9008	0.1030e-04	0.8515e-05	0.83	0.1232e-03	11.96
660	13045	-0.5401e-06	-0.1246e-05	2.31	0.3717e-04	68.82
792	18070	-0.5414e-06	-0.5697e-06	1.05	0.1095e-04	20.23
999	26020	0.1010e-06	0.5856e-07	0.58	0.1112e-05	11.01
1269	37181	0.1455e-08	0.1085e-08	0.75	0.8654e-07	59.49
1725	57850	-0.6196e-10	-0.4914e-10	0.79	0.2841e-08	45.86
2541	94649	0.3163e-11	0.2822e-11	0.89	0.9505e-10	30.05

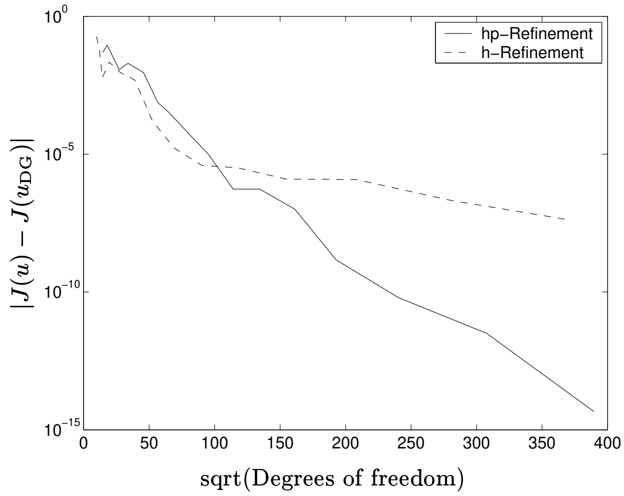
and boundary condition $u(0,x_2) = 0$, for $x_2 \in [0,2]$, cf. Fig. 17(a). The analytical solution to (58), (59) consists of two smooth 'hills' which form shock waves as time increases. These shocks will eventually merge to form a single line of discontinuity in the (x_1,x_2)–plane. Here, we select the functional of interest $J(\cdot)$ to be the value of the solution before these two shocks interact with one another. More precisely, we choose

$$J(u) = u(2.3, 1.5)\,,$$

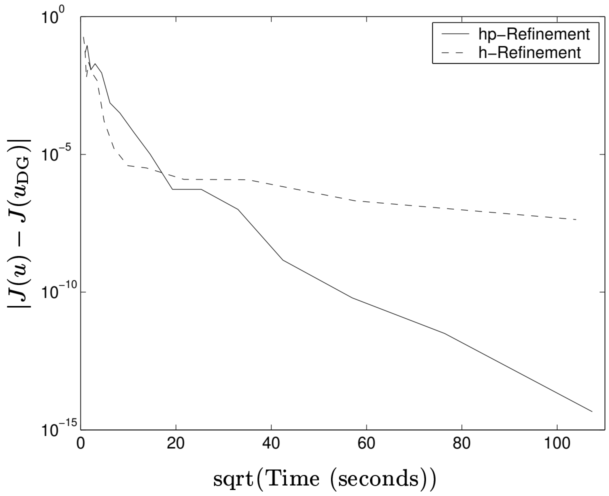
cf. Fig. 17(b); thereby $J(u) = 0.664442403975254670$.

In Table 10 we show the performance of our hp–adaptive finite element algorithm with $C_\varepsilon = 1/20$ and $\beta = 1/10$. As for the wave equation presented in Section 6.3, we see that the quality of the computed error representation formula is quite poor on coarse meshes; though, as the finite element space $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$ is enriched the effectivity index θ_1 is very close to one. Furthermore, we again observe that the Type I *a posteriori* error bound (57) is reliable and asymptotically sharp; indeed, $\tilde{\mathcal{E}}_{|\Omega|} = \sum_\kappa |\tilde{\eta}_\kappa|$ overestimates $|J(u-u_{\rm DG})|$ by just over an order of magnitude.

In Fig. 18(a) we compare the performance of the h– and hp–mesh refinement algorithms for this problem. Again, we observe exponential convergence of the error in the computed functional using hp–refinement; on the linear–



(a)



(b)

Fig. 18. Burgers' equation. Comparison between h– and hp–adaptive mesh refinement: (a) $|J(u)-J(u_{\rm DG})|$ versus number of degrees of freedom; (b) $|J(u)-J(u_{\rm DG})|$ versus computational time

log scale, the convergence line is straight. On the final mesh the true error between $J(u)$ and $J(u_{\rm DG})$ using hp–refinement is almost 5 orders of magnitude smaller than the corresponding quantity when h–refinement is employed alone. Furthermore, from Fig. 18(b) we observe that the hp–refinement algo-

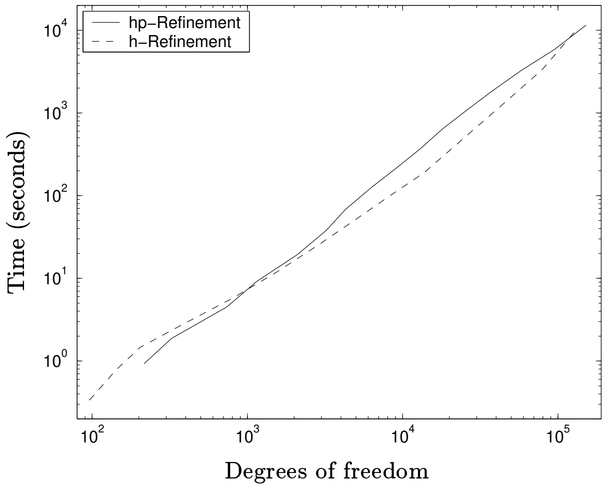


Fig. 19. Burgers' equation. Computational time versus number of degrees of freedom

rithm also outperforms the h–refinement strategy, when comparing the error in the computed target functional with respect to the computational cost. Indeed, Fig. 19 clearly shows that for the hp–DGFEM the cost per degree of freedom when hp–refinement is employed is comparable to that of using h–refinement.

Finally, in Fig. 20 we show the primal mesh after 11 adaptive hp–mesh refinements. Here, we see that the h–mesh has been refined in the region upstream of the point of interest, thereby isolating the smooth region of u from the two interacting shock waves; this renders the subsequent p–refinement in this region much more effective.

Subsonic flow around a NACA0012 airfoil. In this second example, we consider the subsonic flow around a NACA0012 airfoil; here, the upper and lower surfaces of the airfoil geometry are specified by the function $g^{\pm}$, respectively, where

$$g^{\pm}(s) = \pm 5 \times 0.12 \times (0.2969 s^{1/2} - 0.126 s - 0.3516 s^2 + 0.2843 s^3 - 0.1015 s^4),$$

cf. Fig. 21. We note that the chord length l of the airfoil is $l \approx 1.00893$; thereby, we write $\hat{g}$ to denote the rescaling of g to yield an airfoil of unit (chord) length. The computational domain Ω is now defined to be an annulus with inner boundary $\hat{g}$ and outer boundary consisting of a circle of radius 10 units. On the outer boundary, we prescribe a Mach 0.5 flow at a zero angle of attack, with farfield density $\rho = 1$ and pressure $p = 1$. The solution to this problem consists of a strictly subsonic flow, symmetric about the x–axis,

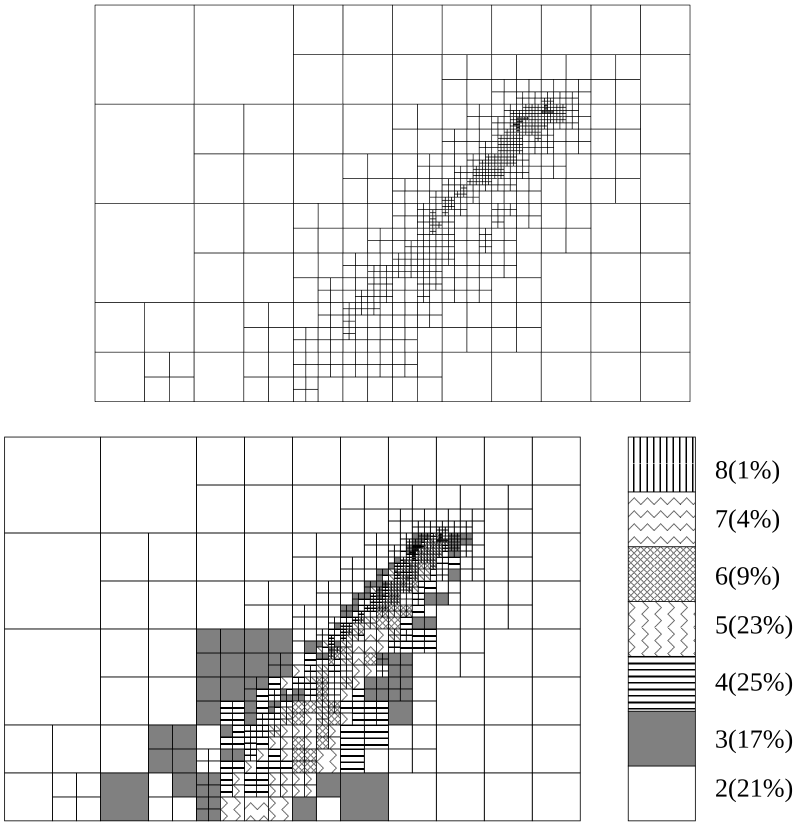


Fig. 20. Burgers' equation. h– and hp–meshes after 11 refinements, with 999 elements and 26020 degrees of freedom; here, $|J(u) - J(u_{\mathrm{DG}})| = 1.010 \times 10^{-7}$

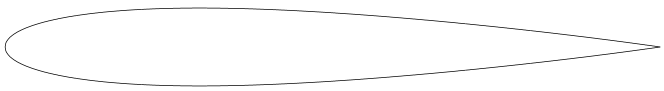


Fig. 21. Profile of the NACA0012 airfoil

cf. Fig. 22. Given that this flow is smooth, no artificial viscosity is required; thereby, we set $C_\varepsilon = 0$.

Here, we consider the evaluation of the drag coefficient on the surface of the airfoil, i.e. $J(\cdot) \equiv J_{\mathrm{drag}}(\cdot)$. Since that the angle of attack $\alpha = 0$, the true value of the target functional $J_{\mathrm{drag}}(\mathbf{u})$ is zero. In Fig. 23 we show the performance of the adaptive algorithm using both h– and hp–mesh refinement. Since the primal solution is smooth we again observe exponential convergence

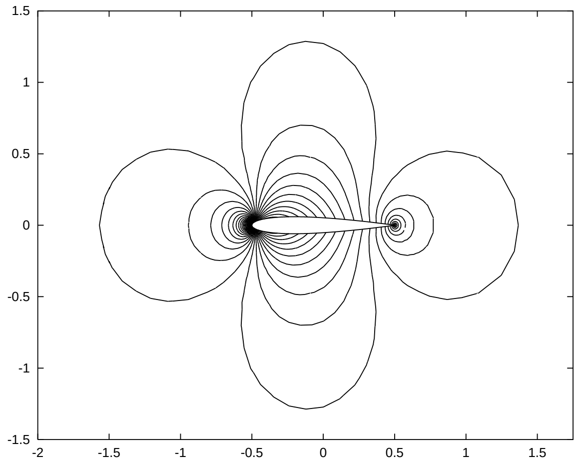


Fig. 22. Mach isolines for the subsonic NACA0012 flow

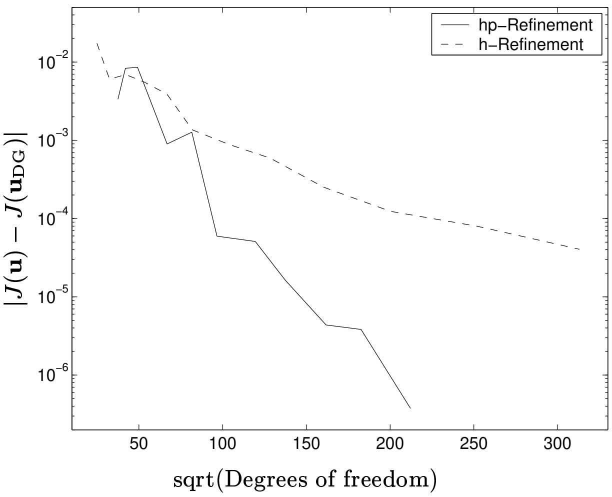


Fig. 23. Subsonic NACA0012 flow. Comparison between h– and hp–adaptive mesh refinement

of $J(\mathbf{u}_{\mathrm{DG}})$ to $J(\mathbf{u})$ when the hp–refinement is employed; on a linear–log scale the convergence line is (on average) straight. Fig. 23 also demonstrates the superiority of the adaptive hp–refinement strategy over the standard adaptive h–refinement algorithm. On the final mesh the true error between $J(\mathbf{u})$ and $J(\mathbf{u}_{\mathrm{DG}})$ using hp–refinement is over 2 orders of magnitude smaller than the corresponding quantity when h–refinement is employed alone.

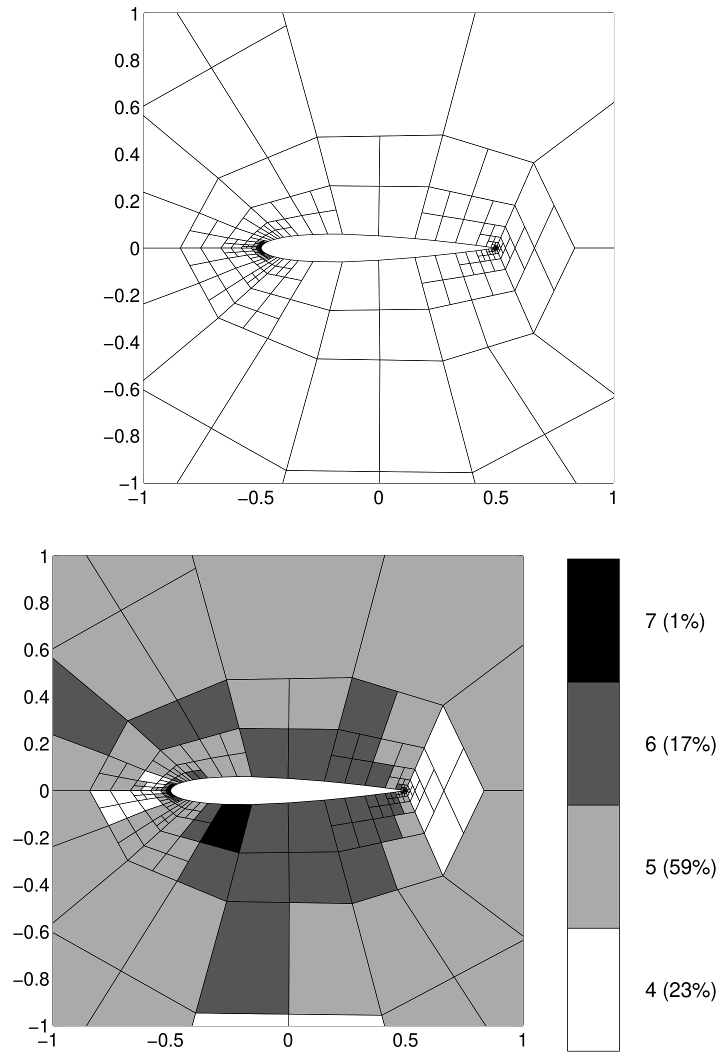


Fig. 24. Subsonic NACA0012 flow. h– and hp–meshes after 10 refinements, with 325 elements and 45008 degrees of freedom; here, $|J(\mathbf{u}) - J(\mathbf{u}_{\mathrm{DG}})| = 3.756 \times 10^{-7}$

Finally, in Fig. 24 we show the primal mesh after 10 adaptive hp–mesh refinements. Here, we see that the h–mesh has been refined in the neighborhood of the leading and trailing edges of the airfoil, cf. [27]. Moreover, elements located on the upper and lower surfaces of the airfoil have been p–enriched where the primal solution is relatively smooth.

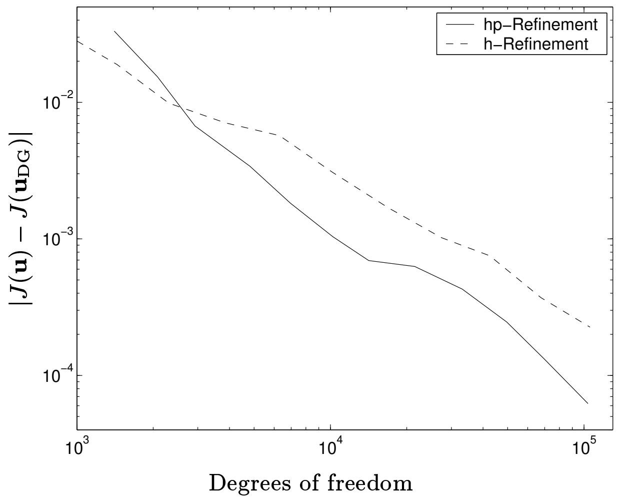


Fig. 25. Transonic NACA0012 flow. Comparison between h– and hp–adaptive mesh refinement

Transonic flow around a NACA0012 airfoil. In this final example, we again compute the flow around a NACA0012 airfoil; here, we consider a Mach 0.8 flow with angle of attack $\alpha = 1.25°$, and farfield density $\rho = 1$ and pressure $p = 1$. The structure of the solution in this example is now quite different from the subsonic flow studied in the previous section. Here, as the impinging subsonic flow passes the airfoil, it first accelerates to become supersonic, before slowing to be subsonic again; this results in the development of shocks on both the upper and lower surfaces of the airfoil, cf. [27]. Throughout this section we set the artificial viscosity parameters appearing in (52) as follows: $C_\varepsilon = 3/100$ and $\beta = 1/10$.

Here, we again consider the evaluation of the drag coefficient on the surface of the airfoil, i.e. $J(\cdot) \equiv J_{\mathrm{drag}}(\cdot)$. On the basis of a fine grid computation, the true value of the functional is given by $J(\mathbf{u}) \approx 0.0201$ (see, [27]). In Fig. 25 we show the performance of the adaptive algorithm using both h– and hp–mesh refinement. In this example both the primal and dual solutions are nonsmooth, so we no longer expect to observe exponential convergence of the error in the functional $J(\cdot)$; thus, here we plot the same quantities as in Fig. 23 against the number of degrees of freedom in $S^{\mathbf{p}}(\Omega, \mathcal{T}, \mathbf{F})$ on a log–log scale. As in our previous examples, we observe that the true error in $J(\cdot)$ is (almost) always smaller if hp–refinement is employed. Indeed, on the final mesh the true error in the drag coefficient is over 70% smaller if hp–refinement is employed as opposed to h–refinement only.

Finally, in Fig. 26 we show the primal mesh after 10 adaptive hp–mesh refinements. As in the previous example, we see that the h–mesh has been refined in the vicinity of the leading and trailing edges of the airfoil; addi-

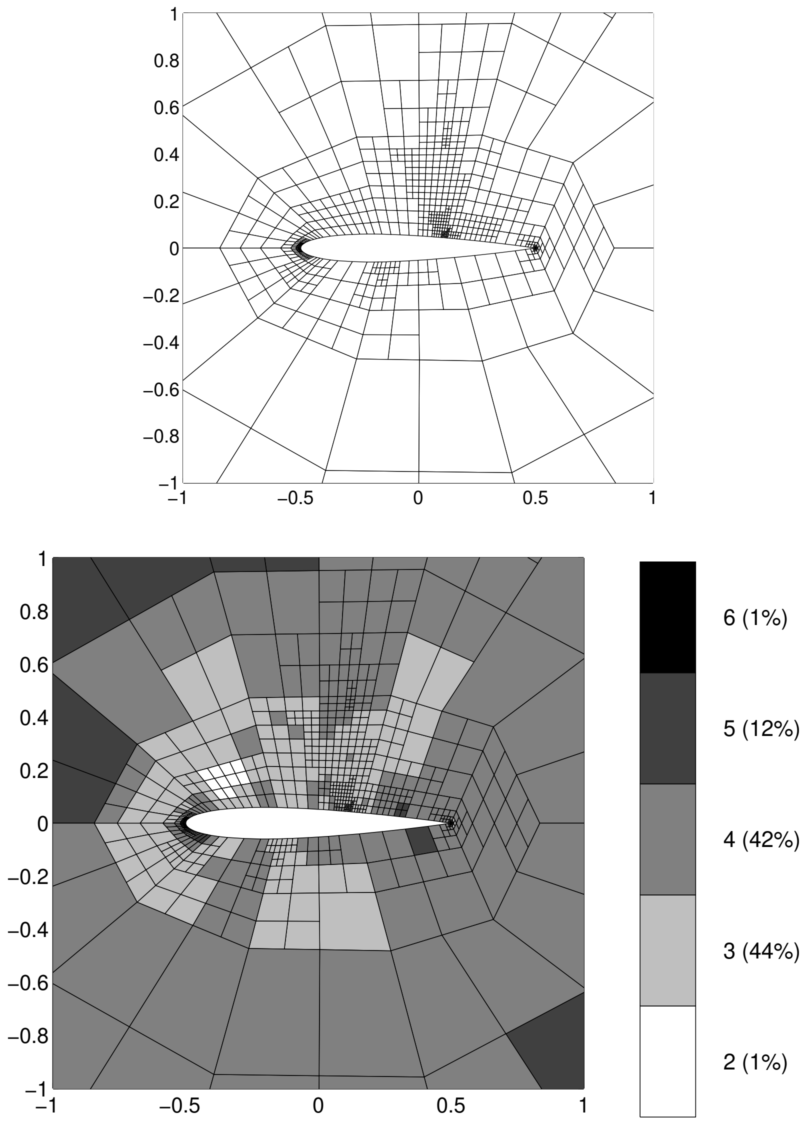


Fig. 26. Transonic NACA0012 flow. h– and hp–meshes after 10 refinements, with 783 elements and 69956 degrees of freedom; here, $|J(\mathbf{u}) - J(\mathbf{u}_{\mathrm{DG}})| = 1.311 \times 10^{-4}$

tional h–refinement has also taken place in the neighbourhood of the shocks emanating from the upper and lower surfaces of the airfoil, cf. [27]. This renders the subsequent p–refinement in the regions of the computational domain where either the primal or dual solution are locally smooth much more effective.

7 Conclusions

Using duality arguments, in this paper we developed Type I *a posteriori* error bounds for the h–version of the streamline–diffusion finite element method and the hp–version of the discontinuous Galerkin finite element method. The bounds have been implemented into adaptive finite element algorithms that are capable of delivering approximations to general output functionals of the solution for both linear and nonlinear hyperbolic problems, accurate to within a prescribed tolerance.

Unlike Type II *a posteriori* error bounds, Type I bounds explicitly involve the dual solution as local weight. Indeed, a Type I bound is always sharper than the corresponding Type II bound, although the extra sharpness does not come for free since a Type I bound requires the numerical solution of the dual problem. As the dual problem is always linear even when the primal problem is nonlinear, for nonlinear problems the additional cost of solving a single dual problem is negligible, cf. [26]. Of course if the primal problem is linear, then the computational cost is roughly doubled.

While the systematic comparative study of Type I and Type II *a posteriori* error bounds on representative model problems will require further work, it is perhaps helpful to point out some potential advantages and disadvantages of the two approaches. Although a Type II *a posteriori* error bound does not explicitly involve the dual solution as local weight, it does, nevertheless, include the strong stability constant C_{stab} of the dual problem. Analytical estimates of C_{stab} are inevitably pessimistic since they are obtained by admitting very large classes of functions as data (e.g., any possible element of $L_2(\Omega)$ or $H^1(\Omega)$ as forcing function). In addition, the analytical proof of strong stability may demand a nontrivial amount of human effort. Computational estimation of C_{stab} involves the solution of the dual problem for a large set of 'typical' data, — a task that is computationally much more demanding than solving a single dual problem for a Type I error bound. On the other hand, once a library of representative stability constants is available [15], they do not need to be recomputed; also, storing representative stability constants is clearly less memory-intensive than storing the dual solution, particularly for three-dimensional problems. Having said this, the storage of the dual solution, for use in a Type I error bound, may pay off if the resulting computational mesh is more economical than the one that stems from a Type II *a posteriori* error bound.

At any rate, on the basis of our computational experience we would argue that the need to perform a single additional computation for the dual solution involved in a Type I *a posteriori* error bound is a worthwhile price to pay for the availability of reliable error control based on a rigorous, systematic, simple and general mathematical framework. Cheap, heuristic alternatives to Type I or Type II *a posteriori* error bounds offer no control on the error; for goal-oriented computations (such as the approximation of output functionals), the optimality of the meshes that result from such heuristic devices is highly questionable.

Acknowledgements

The authors are grateful to Rémi Abgrall, Mark Ainsworth, Ivo Babuška, Tim Barth, Bernardo Cockburn, Leszek Demkowicz, Joe Flaherty, Mike Giles, Ralf Hartmann, Kathryn Harriman, Claes Johnson, Mats Larson, John Mackenzie, Peter Monk, Niles Pierce, Rolf Rannacher, Christoph Schwab, Bill Senior, Thomas Sonar, and Gerald Warnecke for helpful discussions on the subject of this paper. Part of this work was performed using the University of Leicester Mathematical Modelling Centre's supercomputer which was purchased through the EPSRC Strategic Equipment Initiative.

References

1. Adams, R.A. (1975) Sobolev Spaces. Academic Press, New York
2. Adjerid, S., Aiffa, M., and Flaherty, J.E. (1998) Computational methods for singularly perturbed systems. In Cronin, J. and O'Malley, R.E. (eds.), Singular Perturbation Concepts of Differential Equations, AMS, Providence
3. Ainsworth, M. and Senior, B. (1998) An adaptive refinement strategy for *hp*–finite element computations. Appl. Numer. Maths. **26**, 165–178
4. Ainsworth, M. and Oden, J.T. (1996) A Posteriori Error Estimation in Finite Element Analysis. Series in Computational and Applied Mathematics. Elsevier
5. Bardos, C. (1970) Problèmes aux limites pour les équations aux dérivées partielles du premier ordre à coefficients réels; théorèmes d'approximation; application à l'équation de transport. Ann. Sci. École Norm. Sup. **4**, 185–233
6. Barth, T.J. (1998) Numerical methods for gas dynamics systems on unstructured meshes. In: Kröner, D., Ohlberger, M., Rohde, C. (eds.) An Introduction to Recent Developments in Theory and Numerics of Conservation Laws. Lecture Notes in Computational Science and Engineering **5**, pp. 195–285. Springer, Berlin Heidelberg
7. Becker, R. and Rannacher, R. (1996) Weighted a posteriori error control in FE methods. Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, Universität Heidelberg, Preprint No. 1
8. Becker, R. and Rannacher, R. (2001) An optimal control approach to a-posteriori error estimation in finite element methods. Iserles, A. (ed.), Acta Numerica (to appear)
9. Bey, K.S. and Oden, J.T. (1996) *hp*-Version discontinuous Galerkin methods for hyperbolic conservation laws. Comput. Methods Appl. Mech. Engrg. **133**, 259–286
10. Braess, D. (1997) Finite Elements. Theory, Fast Solvers, and Applications in Solid Mechanics. Cambridge University Press, Cambridge
11. Cockburn, B. and Gau, H. (1995) A posteriori error estimates for general numerical methods for scalar conservation laws. Comput. Appl. Math. **14**, 37–47

12. Cockburn, B., Karniadakis, G.E., and Shu, C.-W. (2000) The development of discontinuous Galerkin methods. In: Cockburn, B., Karniadakis, G.E., Shu C.-W. (eds.) Discontinuous Galerkin Finite Element Methods. Lecture Notes in Computational Science and Engineering, **11**. Springer, Berlin Heidelberg
13. Dautray, R. and Lions, J.-L. (1993) Mathematical Analysis and Numerical Methods for Science and Technology. **6**, Evolution Problems II. Springer, Berlin Heidelberg
14. Eriksson, K., Estep, D., Hansbo, P., and Johnson, C. (1995) Introduction to adaptive methods for differential equations. In: Iserles, A. (ed.), Acta Numerica, 105–158
15. Estep, D, Larson, M., and Williams, R. (2002) Estimating the error of numerical solutions of systems of reaction-diffusion equations. Memoirs of the American Mathematical Society (to appear)
16. Friedrichs, K.O. (1958) Symmetric positive linear differential equations. Commun. Pure Appl. Math. **11**, 333–418
17. Giles, M.B., Larsson, M., Levenstam, M., and Süli, E. (1997) Adaptive error control for finite element approximations of the lift and drag coefficients in viscous flow. Numerical Analysis Group Research Report NA-97/06, University of Oxford.
`http://web.comlab.ox.ac.uk/oucl/publications/natr/na-97-06.html`
18. Giles, M.B. and Pierce, N.A. (1997) Adjoint equation in CFD: duality, boundary conditions and solution behaviour. AIAA Paper 97–1850
19. Giles, M.B. (1998) On adjoint equations for error analysis and optimal grid adaptation in CFD. In: Caughey, D. and E. Hafez, E. (eds.), Frontiers of Computational Fluid Dynamics 1998. World Scientific
20. Giles, M.B. (1999) Improved lift and drag estimates using adjoint Euler equations. AIAA Paper 99–3293
21. Giles, M.B. (2000) An introduction to the adjoint design approach and analysis. Numerical Analysis Group Research Report NA-00/04, University of Oxford.
`http://web.comlab.ox.ac.uk/oucl/publications/natr/na-00-04.html`
22. Giles, M.B and Süli, E. (2002) Adjoint methods for PDEs: a posteriori error analysis and postprocessing by duality. Iserles, A. (ed.), Acta Numerica 2002 (to appear)
23. Gui, W. and Babuška, I. (1986) The h, p and h–p versions of the finite element method in 1 Dimension. Part III. The adaptive h–p version. Numer. Math. **49**, 659–683
24. Hansbo, P. and Johnson, C. (1991) Adaptive streamline diffusion finite element methods for compressible flow using conservative variables. Comput. Methods Appl. Mech. Engrg. **87**, 267–280
25. Hartmann, R. (2000) Adaptive FE-methods for conservation equations. In: Warnecke, G. (ed.), Eighth International Conference on Hyperbolic Problems. Theory, Numerics, Applications (HYP2000) (to appear)
26. Hartmann, R. and Houston, P. (2001) Adaptive discontinuous Galerkin finite element methods for nonlinear hyperbolic conservation laws. (Submitted for publication)
27. Hartmann, R. and Houston, P. (2001) Adaptive discontinuous Galerkin finite element methods for the compressible Euler equations. (Submitted for publication)

28. Hartmann, R., Houston, P., and Süli, E. (2001) Adaptive discontinuous Galerkin finite element method for nonlinear hyperbolic problems. In M. Baines, editor, Numerical Methods for Fluid Dynamics VII, ICFD, pp. 347–353
29. Houston, P., Mackenzie, J.A., Süli, E, and Warnecke, G. (1999) A posteriori error analysis for numerical approximations of Friedrichs systems. Numer. Math. **82**, 433–470
30. Houston, P., Rannacher, R. and Süli, E. (2000) A posteriori error analysis for stabilized finite element approximations of transport problems. Comput. Meth. Appl. Mech. Engrg. **190**, 1483–1508
31. Houston, P., Schwab, Ch. and Süli, E. (2000) Stabilized *hp*–finite element methods for first–order hyperbolic problems. SIAM J. Numer. Anal. **37**, 1618–1643
32. Houston, P., Schwab, Ch., and Süli, E. (2002) Discontinuous *hp*–finite element methods for advection–diffusion–reaction problems. SIAM J. Numer. Anal. (to appear)
33. Houston, P., Senior, B. and Süli, E. (2002) Sobolev regularity estimation for *hp*–adaptive finite element methods. (Submitted for publication)
34. Houston, P. and Süli, E. (2001) Stabilized *hp*–finite element approximation of partial differential equations with non-negative characteristic form. Computing **66**, 99–119
35. Houston, P. and Süli, E. (2001) *hp*–Adaptive discontinuous Galerkin finite element methods for hyperbolic problems. SIAM J. Sci. Comp., **23**, 1225–1251
36. Hughes, T.J.R., Franca, F.P. and Mallet, M. (1986) A new finite element formulation for computational fluid dynamics: symmetric forms of the compressible Euler and Navier–Stokes equations and the second law of thermodynamics. Comp. Meth. App. Mech. Engrg. **54**, 223–234
37. Jaffre, J., Johnson, C. and Szepessy, A. (1995) Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws. Math. Models Methods Appl. Sci. **5**, 367–386
38. Johnson, C., Nävert, U. and Pitkäranta, J. (1984) Finite Element Methods for linear hyperbolic problems. Comp. Meth. Appl. Mech. Engrg. **45**, 285–312
39. Kröner, D. (1997) Numerical Schemes for Conservation Laws. Wiley-Teubner
40. Kröner, D. and Ohlberger, M. (2000) A-posteriori error estimates for upwind finite volume schemes for nonlinear conservation laws in multi dimensions. Math. Comput. **69**, 25–39
41. Lax, P.D. and Phillips, R.S. (1960) Local boundary conditions for dissipative symmetric linear differential operators. Comm. Pure Appl. Math., **13**, 427–455
42. Larson, M. and Barth, T.J. (2000) In: Cockburn, B., Karniadakis, G.E., Shu C.-W. (eds.) Discontinuous Galerkin Finite Element Methods. Lecture Notes in Computational Science and Engineering, **11**, pp. 363–368. Springer, Berlin Heidelberg
43. Mock, M.S. (1980) Systems of conservation laws of mixed type. J. Diff. Equ. **37**, 70–88
44. Monk, P. and Süli, E. (1998) The adaptive computation of far field patterns by a posteriori error estimates of linear functionals. SIAM J. Numer. Anal. **36**, 251–274
45. Oden, J.T. and Prudhomme, S. (1999) On goal-oriented error estimation for elliptic problems: Application to control of pointwise errors. Comput. Meth. Appl. Mech. Engrg. **176**, 313–331

46. Paraschivoiu, M., Peraire, J. and Patera, A. (1997) A posteriori finite element bounds for linear functional outputs of elliptic partial differential equations. Comput. Meth. Appl. Mech. Engrg. **150**, 289–312
47. Peraire, J. and Patera, A.T. (1997) Bounds for linear functional outputs of coercive partial differential equations: local indicators and adaptive refinement. In: Ladeveze, P. and Oden, J.T. (eds.), New Advances in Adaptive Computational Methods in Mechanics. Elsevier
48. Pierce, N.A. and Giles, M.B. (1998) Adjoint recovery of superconvergent functionals from approximate solutions of partial differential equations. Numerical Analysis Group Research Report NA-98/18, University of Oxford. `http://web.comlab.ox.ac.uk/oucl/publications/natr/na-98-18.html`
49. Rannacher, R. (1998) Adaptive finite element methods. In: Bulgak, H. and Zenger, C. (eds.), Proc. NATO-Summer School on Error Control and Adaptivity in Scientific Computing, Kluwer Academic Publishers, 247–278
50. Rauch, J. (1972) L_2 is a continuable initial condition for Kreiss' mixed problems. Comm. Pure Appl. Math. **25**, 265–285
51. Schwab, Ch. (1998) *p*- and *hp*-Finite Element Methods. Theory and Applications to Solid and Fluid Mechanics. Oxford University Press, Oxford
52. Sonar, T. and Süli, E. (1998) A dual graph-norm refinement indicator for finite volume approximations of the Euler equations. Numer. Math. **78**, 619–658
53. Süli, E. (1998) A posteriori error analysis and adaptivity for finite element approximations of hyperbolic problems. In: Kröner, D., Ohlberger, M., Rohde, C. (eds.) An Introduction to Recent Developments in Theory and Numerics of Conservation Laws. Lecture Notes in Computational Science and Engineering **5**, pp. 123–194. Springer, Berlin Heidelberg
54. Süli, E., Houston, P. and Schwab, Ch. (1999) *hp*–Finite element methods for hyperbolic problems. In: Whiteman, J.R. (ed.), The Mathematics of Finite Elements and Applications X. MAFELAP 1999. Elsevier, 143–162
55. Süli, E., Houston, P. and Senior, B. (2001) *hp*–Discontinuous Galerkin finite element methods for nonlinear hyperbolic problems. Int. J. Numer. Meth. Fluids (to appear)
56. Szabó, B. and Babuška, I. (1991) Finite Element Analysis. J. Wiley & Sons, New York
57. Verfürth, R. (1996) A Review of a Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques. B.G. Teubner, Stuttgart