Acta Numerica (2002), pp. 145–236
DOI: 10.1017/S096249290200003X

Adjoint methods for PDEs: *a posteriori* error analysis and postprocessing by duality

Michael B. Giles and Endre Süli
University of Oxford, Computing Laboratory,
Wolfson Building, Parks Road,
Oxford OX1 3QD, England

We give an overview of recent developments concerning the use of adjoint methods in two areas: the *a posteriori* error analysis of finite element methods for the numerical solution of partial differential equations where the quantity of interest is a functional of the solution, and superconvergent extraction of integral functionals by postprocessing.

CONTENTS

1. Introduction

Output functionals

In many scientific and engineering applications that lead to the numerical approximation of solutions to partial differential equations, the objective is merely a rough, qualitative assessment of the details of the analytical solution over the computational domain, the quantitative concern being directed towards a few *output functionals*, derived quantities of particular engineering or scientific relevance.

For example, in aeronautical engineering, a CFD calculation of the flow around a transport aircraft at cruise conditions might be performed to investigate whether there are any unexpected shocks on the pylon connecting the engine to the wing, or whether there is an unexpected boundary layer separation caused by the main shock on the wing's suction surface. However, the engineer's overall concern is the impact of such phenomena on the lift and drag on the aircraft, and the quality of the CFD calculation is judged, first and foremost, by the accuracy of the lift and drag predictions. The fine details of the flow field are much less important, and are used only in a qualitative manner to suggest ways in which the design may be modified to improve the lift or drag. This focus on a few output quantities is even clearer in design optimization, when one is trying to maximize or minimize a single objective function, possibly subject to a number of constraints.

Engineering interest in output functionals arises in many applications of CFD. Occasionally, volume integrals are of importance: one example is the infrared signature of a military aircraft, which will depend in part on a volume integral of some function of the temperature in the thermal wake behind the aircraft. However, usually it is surface integrals that are of most concern, as with lift and drag. Other examples in CFD analysis include: the mass flow through a turbomachine; the total heat flux into a turbine blade from the surrounding flow; the total production of nitrous oxides in combustion modelling; the net seepage of a pollutant into an aquifer when modelling soil contamination.

Integral quantities are important in other disciplines as well. In electrochemical simulations of the behaviour of sensors, the quantity of interest is the total current flowing into an electrode (Alden and Compton 1997). In electromagnetics, radar cross-section calculations are concerned with the scattered field emanating from an aircraft. The amplitude of the wave propagating in a particular direction can be evaluated by a convolution integral over a closed surface surrounding the aircraft (Colton and Kress 1991, Monk and Süli 1998). Similar convolution integrals are used in the analysis of multi-port electromagnetic devices, such as microwave ovens and EMR body scanners, to evaluate radiation, transmission and reflection coefficients which characterize the behaviour of the device.

In structural mechanics, one is sometimes concerned with the total force or moment exerted on a surface (Peraire and Patera 1997), but more often the focus of interest is a point functional, such as the maximum stress or temperature. Indeed, as integral quantities can be approximated with much greater accuracy than point functionals, various techniques have been developed to represent point quantities by 'equivalent' integral quantities (see, for example, Babuška and Miller (1984*a*)). In fact, the applicability of these techniques extends beyond structural engineering to other areas where the accurate evaluation of point quantities, rather than integral functionals, is of concern.

The purpose of this article is to explore the question of accurate approximation of output functionals through the use of adjoint problems and duality arguments. As a first step in this direction, we analyse the errors committed in the numerical approximation of linear functionals using an appropriately defined *adjoint* or *dual* problem; hence, we shall quantify the relationship between residual errors in the discretization and the corresponding error in the approximation of the functional. With this information, we then look at ways of improving the accuracy of the computed value of the functional, either through correcting the leading order error in the functional approximation, or through an adaptive mesh refinement algorithm that stems from a residual-based *a posteriori* error bound, aiming to produce the most accurate functional value at a given computational cost.

Partial differential operators and adjoint equations

The application of duality arguments in the theory of differential and integral equations has a long and distinguished history, including the work of Frobenius on two-point boundary value problems, the construction of a Riemann function for a second-order linear hyperbolic partial differential operator, Holmgren's theorem concerning the uniqueness of solutions to parabolic partial differential equations, and Fredholm's theory of integral equations.

The purpose of this section is to highlight, in nonrigorous terms, the intimate connection between adjoint problems and output functionals in the context of partial differential equations. Let us suppose that L is a scalar linear partial differential operator with constant coefficients, and for a sufficiently smooth function f defined over a domain $\Omega \subset \mathbb{R}^n$, let us consider the equation

$$Lu = f \qquad \text{in } \Omega,$$

subject to (unspecified) homogeneous boundary conditions on $\partial\Omega$. Assuming that the Green's function $G : (x,y) \in \Omega \times \Omega \mapsto \mathbb{R}$ of L exists, the solution u can be expressed as

$$u(x) = \int_\Omega G(x,y)\, f(y)\, \mathrm{d}y,$$

where $G(x,y)$ satisfies, for every $x \in \Omega$, the partial differential equation

$$L_y^* G = \delta(x-y), \qquad y \in \Omega,$$

subject to appropriate homogeneous boundary conditions; here L^* denotes the formal adjoint of L, and the subscript y in L_y^* indicates that the partial derivatives are taken with respect to the independent variable $y \in \Omega$.

Now, suppose that one is interested in computing $J(u)$, where $w \mapsto J(w)$ is the linear functional defined by

$$J(w) = \int_\Omega g(x)\, w(x)\, \mathrm{d}x,$$

with g a given, sufficiently smooth, weight function. Clearly, on interchanging the order of integration,

$$J(u) = \int_\Omega \int_\Omega g(x)\, G(x,y)\, f(y)\, \mathrm{d}x\, \mathrm{d}y = \int_\Omega v(y)\, f(y)\, \mathrm{d}y,$$

where we have defined v by

$$v(y) = \int_\Omega g(x)\, G(x,y)\, \mathrm{d}x.$$

Now, v obeys the adjoint partial differential equation

$$L_y^* v = g,$$

subject to appropriate homogeneous boundary conditions.

We see from this discussion that the functional value $J(u)$ may be computed without prior knowledge of u, simply by integrating the 'adjoint/dual solution' v against the forcing function f of the original problem. Indeed, the adjoint solution v may be thought of as a measure of sensitivity of the output functional J to perturbations in the data, f. These simple observations have some far-reaching consequences which will be explored in detail in Section 2.

Adjoint equations in engineering and science

The use of adjoint equations is long-established in optimal control theory (Lions 1971). In the simplest case, one has a control system in which an output $y(t)$ is related to a control input $u(t)$ through a scalar linear ordinary differential equation of the form

$$Ly = u,$$

subject to appropriate boundary conditions. The objective is to choose the control input $u(t)$ to achieve a specified state $y(T)$ at time T, while minimizing the integral of the square of the input.

Using calculus of variations, it can be shown that the optimal input must satisfy the adjoint equation

$$L^* u = 0,$$

subject to appropriate boundary conditions.

The idea of using adjoint equations for design optimization in the context of fluid dynamics was pioneered by Pironneau (1974) but, within the field of aeronautical engineering, the adjoint approach to computing design sensitivities has been primarily developed by Jameson, starting with the potential flow equations (Jameson 1988), and then the Euler equations (Jameson 1995), before proceeding to the Navier–Stokes equations (Jameson 1999, Jameson, Pierce and Martinelli 1998). An overview of recent developments in adjoint design methods for aeronautical applications is provided by Newman, Taylor, Barnwell, Newman and Hou (1999).

In studies of turbulent flow, adjoint equations have been used to investigate the active control of turbulent boundary layers to reduce drag through active re-laminarization (Bewley 2001). They have also been applied to the study of the most unstable modes which lead to the initial onset of turbulence (Airiau 2001).

In weather prediction, adjoint equations are used for a process known as *data assimilation* (Talagrand and Courtier 1997). Due to the chaotic nature of high Reynolds number fluid flow, weather prediction is very sensitive to the initial conditions specified. The idea in data assimilation is to adjust the initial conditions to improve the agreement with a limited number of subsequent measurements. As with engineering design optimization, this is essentially an optimization task, and adjoint solutions are used to find the sensitivity of the objective function, in this case the mismatch between the model and the experimental data, to changes in the initial data.

For further examples of the use of adjoint methods, see the recent special issue of the journal *Flow, Turbulence and Combustion* (Bottaro, Mauss and Henningson, eds 2001) which was devoted to a variety of applications in fluid dynamics.

Adjoint equations in numerical analysis

The use of adjoint equations and duality arguments has also penetrated the field of numerical analysis of partial differential equations. In the subject of *a priori* error estimation, these ideas can be traced back to the work of Aubin (1967), Nitsche (1968) and Oganesjan and Ruhovec (1969).

In the subject of residual-based *a posteriori* error estimation, the application of duality arguments in much more recent. The relevance of duality arguments in *a posteriori* error estimation has been highlighted in the review articles by Eriksson, Estep, Hansbo and Johnson (1996) and Becker and Rannacher (2001) (see also Becker and Rannacher (1996), Hansbo and Johnson (1991), Houston, Rannacher and Süli (2000*a*), Houston and Süli (2001*a*), Larson and Barth (2000), Melenk and Schwab (1999), Oden and Prudhomme (1999), Paraschivoiu, Peraire and Patera (1997), Peraire and Patera (1997), Rannacher (1998), Süli (1998), Süli, Houston and Schwab

(1999) and Houston and Süli (2002*b*)); concerning the use of duality arguments in post-processing and design, we refer to Giles, Larsson, Levenstam and Süli (1997), Giles and Pierce (1997, 1999), Giles (2000) and Pierce and Giles (1998), 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 linear 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 (Becker and Rannacher 1996, Becker and Rannacher 2001, Giles *et al.* 1997, Giles and Pierce 1997, Giles and Pierce 1999, Giles 2000, Larson and Barth 2000, Oden and Prudhomme 1999, Paraschivoiu *et al.* 1997, Peraire and Patera 1997, Pierce and Giles 1998).

The relevance and generality of duality-based error estimation has been powerfully argued in the work of Johnson and his collaborators; see, for example, Eriksson *et al.* (1996) 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 PDE. In the framework of Eriksson *et al.* (1996), 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.

An overview of the paper

In this article, we are fundamentally interested in the same subject as Becker and Rannacher (2001), the numerical analysis of errors in output functionals. However, while Becker and Rannacher are concerned with Galerkin finite element methods with orthogonality between the residual errors in the primal problem and the trial space for the dual problem, here we shall also discuss discretization methods, such as finite volume methods, which may lack this orthogonality property.

We begin by introducing the notion of linear primal and dual problems in an abstract weak formulation, and prove their equivalence in the Primal–Dual Equivalence Theorem. In Section 3 we show that this equivalence can be maintained in a Galerkin finite element discretization which retains orthogonality between the residual errors of one problem, and the trial space of the other. However, if one uses a Galerkin discretization with entirely different spaces for the primal and dual problems, the equivalence is lost.

Section 4 explores general discretizations, in the absence of Galerkin orthogonality. In this case it is shown how one may evaluate an adjoint error correction which, to leading order, corrects the error in the computed value for the functional. Applying this technique to smoothly reconstructed solutions from finite difference or finite element approximations, one can establish an order of convergence for the corrected functional which is, typically, twice that of the underlying approximate solution.

Section 5 shows that the reconstructed solution can also be used to obtain improved accuracy through a defect correction procedure, but even more accuracy can be achieved by using both defect correction and adjoint error correction. A key to the success of both the adjoint error correction and the defect correction is that the reconstructed solution has an error which is smooth. In Section 6 we present some preliminary analysis of how this may be achieved, given, as a starting point, an initial approximate solution with an error which is pointwise second-order convergent, but whose gradient is first-order convergent.

Section 7 is devoted to the derivation of residual-based *a posteriori* error bounds for h-version finite element approximations of linear output functionals. Specifically, we consider the approximation of the normal flux of the solution to an elliptic boundary value problem through the boundary of the domain. We highlight the significance of Type I *a posteriori* error bounds, where the solution of the adjoint problem appears as local weight to the finite element residual, and we discuss the implementation of Type I *a posteriori* error bounds into h-adaptive finite element algorithms.

In Section 8 we study the effect of mesh-dependent perturbations on the Primal–Dual Equivalence Theorem. We also consider how such perturbations affect the choice of the dual problem. In particular, we show by considering stabilized finite element approximations of a linear hyperbolic problem that if the formal adjoint of the differential operator is used to define the dual problem, then the stabilization term present in the method may lead to an *a posteriori* error bound which exhibits a rate of convergence inferior to that of the error in the output functional. We also show how the bound may be sharpened by using adjoint error correction, and how the problem may be avoided altogether by defining the adjoint problem through the use of the bilinear form of the numerical method.

In Section 9 we develop the *a posteriori* error analysis of hp-version finite element approximation of functionals of solutions to linear and nonlinear hyperbolic problems. Again, we concentrate on Type I *a posteriori* bounds, where the adjoint solution appears in the bound as local weight. We illustrate the ideas through the hp-version of the discontinuous Galerkin finite element method which admits easy and flexible implementation of adaptive local polynomial-degree variation.

We close in Section 10 with some concluding remarks, and discuss areas of further research.

2. The Primal–Dual Equivalence Theorem

Suppose that U and V are two real Hilbert spaces with norms $\|\cdot\|_U$ and $\|\cdot\|_V$, and $U_0 \subseteq U$ and $V_0 \subseteq V$ are either proper real Hilbert subspaces of U and V equipped with norms $\|\cdot\|_U$ and $\|\cdot\|_V$, respectively, or $U_0 = U$, $V_0 = V$. If U_0 is a proper subspace of U and p is a fixed element of U, we define $U_p = p + U_0$; similarly, if V_0 is a proper Hilbert subspace of V and d a fixed element in V, we let $V_d = d + V_0$. Clearly, if $p \in U_0$ then, by linearity, $U_p = U_0$; similarly, if $d \in V_0$ then $V_d = V_0$.

We consider the following variational problem, which we shall henceforth refer to as the *primal problem*.

(P) Suppose that $m : U \to \mathbb{R}$ and $\ell : V \to \mathbb{R}$ are bounded linear functionals, and let $B(\cdot\,,\cdot) : U \times V \to \mathbb{R}$ be a bounded bilinear functional. Find $J_p \in \mathbb{R}$ and $u \in U_p$ such that

$$J_p = m(u) + \ell(v) - B(u,v) \qquad \forall v \in V_d. \tag{2.1}$$

Before we embark on a detailed study of the existence and uniqueness of solutions to (P), let us make some preliminary observations.

Suppose for the moment that there exist $J_p \in \mathbb{R}$ and $u \in U_p$ satisfying (2.1). Then, in particular,

$$J_p = m(u) + \ell(d) - B(u,d). \tag{2.2}$$

On decomposing each $v \in V_d$ in (2.1) as $v = d + v_0$ where $v_0 \in V_0$, and subtracting (2.2) from (2.1), it follows that $u \in U_p$ is a solution of the problem

$$B(u, v_0) = \ell(v_0) \qquad \forall v_0 \in V_0. \tag{2.3}$$

Conversely, suppose that (2.3) has the unique solution $u \in U_p$, and define the real number $J_p \in \mathbb{R}$ by (2.2); it then follows that the pair $(J_p, u) \in \mathbb{R} \times U_p$ is the unique solution to (P). Next, we shall prove that, under suitable assumptions, both (2.1) and (2.3) have unique solutions.

Theorem 2.1. In addition to the hypotheses of (P), suppose that the bilinear form $B(\cdot\,,\cdot)$ is weakly coercive on $U_0 \times V_0$ in the following sense:

(a) there exists a constant $\gamma_0 > 0$ such that

$$\inf_{w_0 \in U_0 \setminus \{0\}} \sup_{v_0 \in V_0 \setminus \{0\}} \frac{|B(w_0, v_0)|}{\|w_0\|_U \|v_0\|_V} \geq \gamma_0;$$

(b) $\qquad \forall v_0 \in V_0 \setminus \{0\} \qquad \sup_{w_0 \in U_0} B(w_0, v_0) > 0.$

Then problem (P) has a unique solution $(J_p, u) \in \mathbb{R} \times U_p$.

Proof. Consider the problem of finding $u_0 \in U_0$ such that

$$B(u_0, v_0) = \ell(v_0) - B(d, v_0) \qquad \forall v_0 \in V_0. \tag{2.4}$$

By virtue of Babuška's generalization of the Lax–Milgram theorem (see, for example, Theorem 6.2 on p. 224 of Oden and Reddy (1983)), under the present hypotheses problem (2.4) has a unique solution $u_0 \in U_0$. Consequently, $u = d + u_0$ is the unique solution to (2.3), and the pair (J_p, u) in $\mathbb{R} \times U_p$, with J_p defined by (2.2), is the unique solution of (2.1). $\square$

Problems of the form (P) will be referred to throughout the text as *measurement problems*, since the process of computing the value

$$J_p = \mathcal{J}_p(u) = m(u) + \ell(d) - B(u, d)$$

of $\mathcal{J}_p(w)$ at $w = u$ has the physical interpretation of sampling the 'output functional' $\mathcal{J}_p(\cdot)$ at u, which can be thought of as making a certain measurement of the solution u to the variational problem (2.3). The relevance of accurate computation of output functionals in engineering applications has been highlighted in the Introduction.

In order to motivate the discussion that will follow, we consider some simple illustrative examples where the quantity of interest is an output functional.

2.1. The elliptic model problem

Suppose that Ω is a bounded open set in $\mathbb{R}^n$, $n \leq 3$, with Lipschitz-continuous boundary $\Gamma = \partial\Omega$. Given that $f \in H^{-1}(\Omega)$ and $g \in H^{1/2}(\Gamma)$ (we refer to Adams (1975) for elements from the theory of Sobolev spaces), consider Poisson's equation subject to a nonhomogeneous Dirichlet boundary condition:

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega, \\ u &= g \qquad \text{on } \Gamma. \end{aligned}$$

In this case, $U = V = H^1(\Omega)$, $U_0 = V_0 = H^1_0(\Omega)$ and

$$U_p = p + H^1_0(\Omega) = \{v \in H^1(\Omega) \, : \, \gamma_{0,\Gamma}(v) = g\},$$

where $\gamma_{0,\Gamma} \, : \, H^1(\Omega) \to H^{1/2}(\Gamma)$ is the classical trace operator and $p \in H^1(\Omega)$ is chosen so that $\gamma_{0,\Gamma}(p) = g$.

The standard weak formulation of the problem is as follows: find $u \in U_p$ such that

$$B(u, v) = \ell(v) \qquad \forall v \in V_0,$$

where

$$B(u, v) = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x, \qquad \ell(v) = \langle f, v \rangle,$$

and $\langle \cdot \, , \cdot \rangle$ denotes the duality pairing between $H^{-1}(\Omega)$ and $H^1_0(\Omega)$.

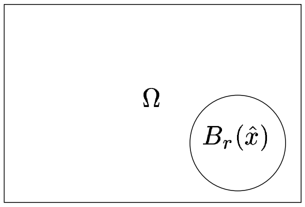


Figure 2.1. Local averaging over $B_r(\hat{x})$: a ball of radius r centred at $\hat{x} \in \Omega$

Local average. Our first example is concerned with calculating the integral average of u over an open ball $B_r(\hat{x}) \subset \Omega$ of radius r centred at $\hat{x} \in \Omega$, as illustrated in Figure 2.1:

$$\mathcal{J}_p(u) = \frac{1}{\operatorname{meas} B_r(\hat{x})} \int_{B_r(\hat{x})} u(x)\,\mathrm{d}x.$$

In this case, we take $d = 0$ and hence $V_d = V_0 = H_0^1(\Omega)$. Consider the problem (2.1) of finding $(J_p, u) \in \mathbb{R} \times U_p$ such that

$$J_p = m(u) + \ell(v) - B(u,v) \qquad \forall v \in V_d,$$

where

$$m(u) = \frac{1}{\operatorname{meas} B_r(\hat{x})} \int_{B_r(\hat{x})} u(x)\,\mathrm{d}x.$$

Hence, recalling that here $V_d = V_0$ and therefore $\ell(v) - B(u,v) = 0$ for all $v \in V_d = V_0$, we find that

$$J_p = \mathcal{J}_p(u) = m(u) = \frac{1}{\operatorname{meas} B_r(\hat{x})} \int_{B_r(\hat{x})} u(x)\,\mathrm{d}x.$$

Point value. If $u \in C(\Omega)$ it is meaningful to consider the point value

$$\mathcal{J}_p(u) = u(\hat{x})$$

of u at $\hat{x} \in \Omega$; to do so, we again take $d = 0$, $V_d = V_0 = H_0^1(\Omega)$, and seek $(J_p, u) \in \mathbb{R} \times U_p$ such that

$$J_p = m(u) + \ell(v) - B(u,v) \qquad \forall v \in V_d,$$

where

$$m(u) = u(\hat{x}).$$

As $V_d = V_0$, it directly follows that $J_p = u(\hat{x})$. Since $m(u) : u \mapsto u(\hat{x})$ is not a bounded linear functional on the space $H_0^1(\Omega)$, $\Omega \subset \mathbb{R}^n$, for $n \geq 2$, this example does not directly fit into the theoretical setting described here in two or more space dimensions, although $m(u)$ can be approximated by

$$m_r(u) = \frac{1}{\operatorname{meas} B_r(\hat{x})} \int_{B_r(\hat{x})} u(x)\,\mathrm{d}x,$$

with $0 < r \ll 1$; alternatively, since $m : u \mapsto u(\hat{x})$ is a bounded linear functional on $W^{1,p}(\Omega)$ for $p > n$, by extending the present Hilbertian theoretical framework to reflexive Banach spaces, the case of $m(u) = u(\hat{x})$ could be covered directly without having to resort to approximations of the type $m(u) \approx m_r(u)$.

Normal flux. Let us assume that $f \in L^2(\Omega)$ and consider the weighted normal flux of u over Γ, with weight function $\psi \in H^{1/2}(\Gamma)$, defined by

$$\mathcal{J}_p(u) = \int_\Gamma \frac{\partial u}{\partial \nu} \psi \,\mathrm{d}s, \tag{2.5}$$

where $\boldsymbol{\nu}$ is the unit outward normal vector to Γ. Strictly speaking, the integral should be thought of as the duality pairing between $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$. In this case, let

$$V_d = H^1_{-\psi,\Gamma}(\Omega) = \{v \in H^1(\Omega) \,:\, \gamma_{0,\Gamma}(v) = -\psi\},$$

and define $m : U \to \mathbb{R}$ as the trivial linear functional which maps every element of $U = H^1(\Omega)$ into $0 \in \mathbb{R}$. We consider the problem of finding $(J_p, u) \in \mathbb{R} \times U_p$ such that

$$J_p = \ell(v) - B(u, v) \qquad \forall v \in V_d. \tag{2.6}$$

A simple calculation based on Green's second identity shows that

$$J_p = \mathcal{J}_p(u) = \int_\Gamma \frac{\partial u}{\partial \nu} \psi \,\mathrm{d}s. \tag{2.7}$$

The relevance of rewriting the normal flux (2.5) in the form (2.6) will be explained in Section 4. It will be shown that the equality (2.7) is not preserved under discretization; indeed, we shall see that it is *not* the discretization of (2.5) but that of (2.6) that yields the more accurate approximation to (2.5).

2.2. The hyperbolic model problem

Our second model problem is a boundary value problem for a first-order hyperbolic equation. Suppose that $\Omega = (0,1)^n$ and let Γ denote the union of open faces of Ω. Let $\mathbf{b} = (b_1, \ldots, b_n)^{\mathrm{T}}$ belong to $[C^1(\bar{\Omega})]^n$, with each b_i, $i = 1, \ldots, n$, positive on $\bar{\Omega}$; suppose further that $c \in C(\bar{\Omega})$, $f \in L^2(\Omega)$ and $g \in L^2(\Gamma_-)$, where

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

is the *inflow boundary* of Ω and $\boldsymbol{\nu}(x)$ signifies the unit outward normal vector to Γ at $x \in \Gamma$. Consider the transport problem

$$\begin{aligned} \mathbf{b}\cdot\nabla u + cu &= f \qquad \text{in } \ \Omega,\\ u &= g \qquad \text{on } \ \Gamma_-. \end{aligned} \tag{2.8}$$

Under our hypotheses, Γ is noncharacteristic (*i.e.*, the vector field $\mathbf{b}$ is, everywhere on Γ, transversal to Γ). We adopt the following (standard) hypothesis: there exists a positive constant γ such that

$$c(x) - \frac{1}{2}\,\nabla\cdot\mathbf{b}(x) \geq \gamma \qquad \forall x \in \bar{\Omega}. \tag{2.9}$$

In order to deduce the correct weak formulation of (2.8), suppose for the moment that the boundary value problem has a solution u in $H^1(\Omega)$, and let $V_d = V_0 = V = H^1(\Omega)$. On multiplying the partial differential equation in (2.8) by $v \in V$ and integrating by parts, we find that

$$-(u, \nabla\cdot(\mathbf{b}v)) + (cu, v) + \langle u, v\rangle_{\Gamma_+} = (f, v) + \langle g, v\rangle_{\Gamma_-} \qquad \forall v \in V, \tag{2.10}$$

where $(\cdot\,,\cdot)$ denotes the L^2 inner product over Ω,

$$\Gamma_+ = \{x \in \Gamma \,:\, \mathbf{b}\cdot\boldsymbol{\nu} > 0\},$$

and

$$\langle w, v\rangle_{\Gamma_\pm} = \int_{\Gamma_\pm} |\mathbf{b}\cdot\boldsymbol{\nu}| wv\,\mathrm{d}s.$$

We consider the inner product $(\cdot\,,\cdot)_U$ defined by

$$(w, v)_U = (w, v) + \langle w, v\rangle_{\Gamma_+},$$

let U denote the closure of V in $L^2(\Omega)$ with respect to the norm $\|\cdot\|_U$ defined by

$$\|w\|_U = (w, w)_U^{1/2},$$

and put $U_p = U_0 = U$. Clearly, U is a Hilbert space. For $w \in U$ and $v \in V$, we now consider the bilinear form $B(\cdot\,,\cdot) \,:\, U \times V \to \mathbb{R}$ defined by

$$B(w, v) = -(w, \nabla\cdot(\mathbf{b}v)) + (cw, v) + \langle w, v\rangle_{\Gamma_+}$$

and for $v \in V$ we introduce the linear functional $\ell \,:\, V \to \mathbb{R}$ by

$$\ell(v) = (f, v) + \langle g, v\rangle_{\Gamma_-}.$$

We shall say that $u \in U$ is a weak solution to the boundary value problem (2.8) if

$$B(u, v) = \ell(v) \qquad \forall v \in V. \tag{2.11}$$

Theorem 2.2. Assuming (2.9), for each $f \in L^2(\Omega)$ and $g \in L^2(\Gamma_-)$ there exists a unique $u \in U$ satisfying (2.11).

For a survey of well-posedness results for linear hyperbolic boundary value problems, we refer to Bardos (1970) and Dautray and Lions (1993).

On choosing $v \in C_0^\infty(\Omega)$ in (2.11), we deduce that $\mathbf{b} \cdot \nabla u + cu = f$ in the sense of distributions on Ω; as $f - cu \in L^2(\Omega)$, it then follows that any weak solution u of (2.11) in U satisfies $\mathbf{b} \cdot \nabla u \in L^2(\Omega)$. Thus, since $\mathbf{b}$ is transversal to Γ at each point of Γ, we conclude that $\gamma_{0,\Gamma_0}(u) \in L^2(\Gamma_0)$ for any open subset $\Gamma_0 \subset \Gamma$. This will be important in our third example below,where we consider the normal flux of u through $\Gamma_0 = \Gamma_+$.

Local average. For this hyperbolic model problem an example of a quantity of physical interest is the integral average of u over an open ball $B_r(\hat{x})$:

$$\mathcal{J}_p(u) = \frac{1}{\operatorname{meas} B_r(\hat{x})} \int_{B_r(\hat{x})} u(x)\,\mathrm{d}x.$$

We set $p = 0$ and $U_p = U_0 = U$ (with U defined above by completion of $H^1(\Omega)$ in the norm $\|\cdot\|_U$), $d = 0$ and $V_d = V_0 = V = H^1(\Omega)$, and seek $(J_p, u) \in \mathbb{R} \times U_p$ such that

$$J_p = m(u) + \ell(v) - B(u, v) \qquad \forall v \in V_d,$$

where

$$m(u) = \frac{1}{\operatorname{meas} B_r(\hat{x})} \int_{B_r(\hat{x})} u(x)\,\mathrm{d}x \qquad \forall v \in V_d.$$

Clearly,

$$J_p = \mathcal{J}_p(u) = m(u) = \frac{1}{\operatorname{meas} B_r(\hat{x})} \int_{B_r(\hat{x})} u(x)\,\mathrm{d}x.$$

Point value. In general, weak solutions to (2.11) exhibit discontinuities across characteristic hypersurfaces; however, if u is continuous in an open neighbourhood of a point $\hat{x} \in \Omega \cup \Gamma_+$, then it is meaningful to consider the point value

$$\mathcal{J}_p(u) = u(\hat{x})$$

of u at $\hat{x}$. Again, we put $d = 0$, let $V_d = V_0 = V = H^1(\Omega)$, and seek $(J_p, u) \in \mathbb{R} \times U_p$ such that

$$J_p = m(u) + \ell(v) - B(u, v) \qquad \forall v \in V_d,$$

where

$$m(u) = u(\hat{x}).$$

Clearly,

$$J_p = \mathcal{J}_p(u) = u(\hat{x}).$$

As $H^1(\Omega)$ is a proper subspace of U and, as we have already seen in the elliptic case, the functional $m : u \mapsto u(\hat{x})$ is not bounded on $H^1(\Omega)$, except

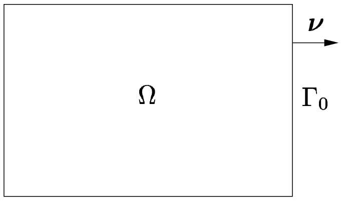


Figure 2.2. Outflow normal flux over a part Γ_0 of the outflow boundary Γ_+ of Ω; $\boldsymbol{\nu}$ denotes the unit outward normal vector to Γ

for $n = 1$, m will not be a bounded linear functional on the space $U \supset H^1(\Omega)$ either, and will need to be approximated through local averaging, as in the elliptic case, to fit into the present theoretical framework.

Outflow normal flux. If the quantity of interest is the weighted normal flux of u over an open subset Γ_0 of Γ_+, as illustrated in Figure 2.2, given by

$$\mathcal{J}_p(u) = \int_{\Gamma_0} |\mathbf{b}\cdot\boldsymbol{\nu}|u\psi\,\mathrm{d}s = \langle u, \psi\rangle_{\Gamma_0},$$

where $\boldsymbol{\nu}$ is the unit outward normal vector to Γ_0 and $\psi \in L^2(\Gamma_0)$ is a fixed weight-function, we let $d = 0$, $V_d = V_0 = V = H^1(\Omega)$ and seek $(J_p, u) \in \mathbb{R} \times U_p$ such that

$$J_p = m(u) + \ell(v) - B(u, v) \qquad \forall v \in V_d,$$

where

$$m(u) = \int_{\Gamma_0} |\mathbf{b}\cdot\boldsymbol{\nu}|u\psi\,\mathrm{d}s \qquad \forall v \in V_d = \langle u, \psi\rangle_{\Gamma_0}.$$

Trivially,

$$J_p = \mathcal{J}_p(u) = \int_{\Gamma_0} |\mathbf{b}\cdot\boldsymbol{\nu}|u\psi\,\mathrm{d}s.$$

After this overview of measurement problems, we introduce the concept of duality, and show that there is an alternative route to obtaining J_p which does not require knowledge of u.

2.3. The dual problem

We begin by associating with the measurement problem (P) the following *dual problem*:

(D) Find $J_d \in \mathbb{R}$ and $z \in V_d$ such that

$$J_d = m(w) + \ell(z) - B(w, z) \qquad \forall w \in U_p. \tag{2.12}$$

In order to ascertain the well-posedness of (D), we recall the following result, which is a straightforward consequence of Proposition A.2 in Melenk and Schwab (1999).

Proposition 2.3. The bounded bilinear form $B : U \times V \to \mathbb{R}$ is weakly coercive on $U_0 \times V_0$ in the sense of (a) and (b) of Theorem 2.1 if and only if $B(\cdot\,,\cdot)$ is adjoint-weakly coercive in the following sense:

(a) there exists a constant $\tilde{\gamma}_0 > 0$ such that

$$\inf_{v_0 \in V_0 \setminus \{0\}} \sup_{w_0 \in U_0 \setminus \{0\}} \frac{|B(w_0, v_0)|}{\|w_0\|_U \|v_0\|_V} \geq \tilde{\gamma}_0;$$

(b) $\qquad \forall w_0 \in U_0 \setminus \{0\} \qquad \sup_{v_0 \in V_0} B(w_0, v_0) > 0.$

Now we are ready to make a statement about the well-posedness of (D).

Theorem 2.4. In addition to the hypotheses of (P), suppose that the bilinear form $B(\cdot\,,\cdot)$ is weakly coercive on $U_0 \times V_0$ in the sense of (a) and (b) of Theorem 2.1. Then there exists a unique pair $(J_d, z) \in \mathbb{R} \times V_d$ that satisfies (D).

Proof. As $B(\cdot\,,\cdot) : U_0 \times V_0 \to \mathbb{R}$ is weakly coercive on $U_0 \times V_0$, by Proposition 2.3 it is adjoint-weakly coercive on $U_0 \times V_0$. Therefore the adjoint bilinear form $B' : V_0 \times U_0 \to \mathbb{R}$ defined by

$$B'(v, w) = B(w, v)$$

is weakly coercive on $V_0 \times U_0$. By Theorem 2.1, the following problem has a unique solution $z_0 \in V_0$: find $z_0 \in V_0$ such that

$$B'(z_0, w_0) = m(w_0) - B(w_0, d) \qquad \forall w_0 \in U_0.$$

Equivalently, there exists a unique $z_0 \in V_0$ such that

$$B(w_0, z_0) = m(w_0) - B(w_0, d) \qquad \forall w_0 \in U_0.$$

Consequently, $z = d + z_0$ is the unique solution to the following problem: find $z \in V_d$ such that

$$B(w_0, z) = m(w_0) \qquad \forall w_0 \in U_0. \tag{2.13}$$

Problem (2.13) will be referred to as the *dual* to problem (2.3). Let us define

$$J_d = m(p) + \ell(z) - B(p, z). \tag{2.14}$$

On writing any $w \in U_p$ as $w = p + w_0$ with $w_0 \in U_0$, we deduce from (2.13) and (2.14) that

$$J_d = m(w) + \ell(z) - B(w, z) \qquad \forall w \in U_p.$$

Hence the pair $(J_d, z) \in \mathbb{R} \times V_d$, with J_d defined by (2.14), is the unique solution of problem (D). □

Our next result encapsulates a rather elementary relationship between the primal and dual problems.

Theorem 2.5. (Primal–Dual Equivalence Theorem) Let u and z denote the solutions to the primal problem (P) and the dual problem (D), respectively; then,

$$J_p = \mathcal{J}_p(u) = \mathcal{J}_d(z) = J_d.$$

Proof. On inserting $w = u$ into (D) and $v = z$ into (P) the required identity trivially follows. □

Despite its simplicity, the practical consequences of this result are far-reaching. For, suppose that instead of a single linear functional $\ell : V \to \mathbb{R}$, we have been given N linear functionals $\ell^{(j)} : V \to \mathbb{R}$, $j = 1, \ldots, N$, where $N \gg 1$, and assume that the task is to find $J_p^{(j)} \in \mathbb{R}$ such that $(J_p^{(j)}, u^{(j)}) \in \mathbb{R} \times U_p$ satisfies

$$J_p^{(j)} = m(u^{(j)}) + \ell^{(j)}(v) - B(u^{(j)}, v) \qquad \forall v \in V_d.$$

Note that, in this case, only the numbers $J_p^{(j)}$, $j = 1, \ldots, N$, are required, while $u^{(j)}$, $j = 1, \ldots, N$, are of no interest. The most direct approach to obtaining $J_p^{(j)}$, $j = 1, \ldots, N$, would be to solve each of the following problems:

Find $u^{(j)} \in U_p$ such that

$$B(u^{(j)}, v) = \ell^{(j)}(v) \qquad \forall v \in V_0, \tag{2.15}$$

for $j = 1, \ldots, N$, and then compute $J_p^{(j)}$ via

$$J_p^{(j)} = m(u^{(j)}) + \ell^{(j)}(d) - B(u^{(j)}, d), \qquad j = 1, \ldots, N.$$

Theorem 2.5, however, offers a more attractive alternative. This consists of first solving the following (single) dual problem:

Find $z \in V_d$ such that

$$B(w, z) = m(w) \qquad \forall w \in V_0, \tag{2.16}$$

computing, for each $j \in \{1, \ldots, N\}$,

$$J_d^{(j)} = m(p) + \ell^{(j)}(z) - B(p, z),$$

and exploiting the fact that, by the Primal–Dual Equivalence Theorem,

$$J_p^{(j)} = J_d^{(j)}, \qquad j = 1, \ldots, N.$$

Obviously, the latter approach involves less effort since the complexity of evaluating $\ell^{(j)}(z)$ is typically substantially smaller than that of determining $u^{(j)}$. Of course, in practice neither (P) nor (D) can be solved exactly, and one has to resort to numerical approximations. We shall show, however, in the next section that the Primal–Dual Equivalence Theorem can be preserved under discretization.

3. Galerkin approximations and duality

3.1. The Discrete Primal–Dual Equivalence Theorem

Suppose that $\{U_0^h\}_{h>0}$ and $\{V_0^h\}_{h>0}$ are two families of finite-dimensional subspaces of U_0 and V_0, respectively, parametrized by $h \in (0,1]$. When U_0 is a proper Hilbert subspace of U, we assign to $p \in U$ the affine variety $U_p^h = p + U_0^h \subset U_p \subset U$; similarly, when V_0 is a proper Hilbert subspace of V, we assign to $d \in V$ the affine variety $V_d^h = d + V_0^h \subset V_d$.

We consider the following finite-dimensional variational problem, which we shall henceforth refer to as the *discrete primal problem.*

(P^h) Suppose that $m : U \to \mathbb{R}$ and $\ell : V \to \mathbb{R}$ are bounded linear functionals and $B(\cdot\,,\cdot) : U \times V \to \mathbb{R}$ is a bounded bilinear functional. Find $J_p^h \in \mathbb{R}$ and $u^h \in U_p^h$ such that

$$J_p^h = m(u^h) + \ell(v^h) - B(u^h, v^h) \qquad \forall v^h \in V_d^h. \tag{3.1}$$

As in Section 2, it is easily seen that

$$J_p^h = m(u^h) + \ell(d) - B(u^h, d),$$

where $u^h \in U_p^h$ solves

$$B(u^h, v_0^h) = \ell(v_0^h) \qquad \forall v_0^h \in V_0^h. \tag{3.2}$$

Thus the existence of a unique solution to (P^h) is equivalent to the requirement that (3.2) have a unique solution u^h in U_p^h. The latter can be ensured, in a similar manner as for the continuous problem in the previous section, by assuming that the bilinear functional $B(\cdot\,,\cdot)$ is weakly coercive on $U_0^h \times V_0^h$ (with U_0^h and V_0^h equipped with the induced norms $\|\cdot\|_U$, $\|\cdot\|_V$). It is important to note, however, that weak coercivity of $B(\cdot\,,\cdot)$ on $U_0^h \times V_0^h$ is an independent assumption, generally *not* implied by weak coercivity of $B(\cdot\,,\cdot)$ on $U_0 \times V_0$.

Let $\{U_0^H\}_{H>0}$ and $\{V_0^H\}_{H>0}$ be two families of finite-dimensional subspaces of U_0 and V_0, respectively, parametrized by $H \in (0,1]$, typically different from the families $\{U_0^h\}_{h>0}$ and $\{V_0^h\}_{h>0}$. We assign to $p \in U$ the affine variety $U_p^H = p + U_0^H \subset U_p \subset U$; similarly, we assign to $d \in V$ the affine variety $V_d^H = d + V_0^H \subset V_d \subset V$. We now define the *discrete dual problem* as follows:

(D^H) Find $J_d^H \in \mathbb{R}$ and $z^H \in V_d^H$ such that

$$J_d^H = m(w^H) + \ell(z^H) - B(w^H, z^H) \qquad \forall w^H \in U_p^H.$$

In complete analogy with (P^h),

$$J_d^H = m(p) + \ell(z^H) - B(p, z^H),$$

where $z^H \in V_d^H$ solves

$$B(w_0^H, z^H) = m(w_0^H) \qquad \forall w_0^H \in U_0^H. \tag{3.3}$$

Thus, the existence of a unique solution to (D^H) is equivalent to the requirement that (3.3) have a unique solution z^H in V_d^H; the latter can be ensured by requiring that $B(\cdot\,,\cdot)$ is adjoint-weakly coercive on $U_0^H \times V_0^H$, or equivalently (*cf.* Proposition 2.3), that $B(\cdot\,,\cdot)$ is weakly coercive on $U_0^H \times V_0^H$.

Note that there is an interchange in the identity of the test and trial spaces. In the primal problem, U_p^h is the trial space and V_0^h is the test space, whereas in the dual problem it is V_d^H which is the trial space, and U_0^h is the test space.

Next we present representation formulae for the error between J_p, J_d and their approximations J_p^h, J_d^H, respectively.

Theorem 3.1. (Error representation formula) Let $(J_p, u) \in \mathbb{R} \times U_p$ and $(J_d, z) \in \mathbb{R} \times V_d$ denote the solutions to (P) and (D), respectively, and let $(J_p^h, u^h) \in \mathbb{R} \times U_p^h$ and $(J_d^H, z^H) \in \mathbb{R} \times V_d^H$ be the solutions to (P^h) and (D^H), respectively. Then,

$$J_p - J_p^h = B(u - u^h, z - z^h) \qquad \forall z^h \in V_d^h, \tag{3.4}$$

$$J_d - J_d^H = B(u - u^H, z - z^H) \qquad \forall u^H \in U_p^H. \tag{3.5}$$

Proof. Since $V_d^h \subset V_d$, we have from (P) that

$$J_p = m(u) + \ell(v^h) - B(u, v^h) \qquad \forall v^h \in V_d^h.$$

Recalling from (P^h) that

$$J_p^h = m(u^h) + \ell(v^h) - B(u^h, v^h) \qquad \forall v^h \in V_d^h$$

and subtracting, we find that

$$J_p - J_p^h = m(u - u^h) - B(u - u^h, v^h) \qquad \forall v^h \in V_d^h. \tag{3.6}$$

On the other hand, as $u - u^h \in U_0$, we deduce from (2.13) that

$$B(u - u^h, z) = m(u - u^h),$$

which we can use to eliminate $m(u - u^h)$ from (3.6) to deduce (3.4). The proof of (3.5) is analogous. □

The initial hypothesis stated in (P) that $B(\cdot\,,\cdot)$ is a bounded bilinear functional on $U \times V$ implies the existence of a positive constant γ_1 such that

$$|B(w, v)| \le \gamma_1 \|w\|_U \, \|v\|_V \qquad \forall w \in U \quad \forall v \in V.$$

Thus we deduce the following result.

Corollary 3.2. (*A priori* error bound) Let $(J_p, u) \in \mathbb{R}\times U_p$ and $(J_d, z) \in \mathbb{R} \times V_d$ denote the solutions to (P) and (D), respectively, and let $(J_p^h, u^h) \in \mathbb{R}\times U_p^h$ and $(J_d^H, z^H) \in \mathbb{R}\times V_d^H$ be the solutions to (P^h) and (D^H), respectively. Then,

$$|J_p - J_p^h| \leq \gamma_1 \|u - u^h\|_U \inf_{v^h \in V_d^h} \|z - v^h\|_V, \tag{3.7}$$

$$|J_d - J_d^H| \leq \gamma_1 \|z - z^H\|_V \inf_{w^H \in U_p^H} \|u - w^H\|_U. \tag{3.8}$$

This abstract superconvergence result expresses the fact that the rate of convergence of J_p^h to J_p as $h \to 0$ (respectively, J_d^H to J_d as $H \to 0$) is higher than that of u^h to u in the norm of U as $h \to 0$ (respectively, z^H to z in the norm of V as $H \to 0$).

Our next result is the discrete counterpart of the Primal–Dual Equivalence Theorem.

Theorem 3.3. (Discrete Primal–Dual Equivalence Theorem) Let us suppose that $(J_p^h, u^h) \in \mathbb{R}\times U_p^h$ and $(J_d^H, z^H) \in \mathbb{R}\times V_d^H$ denote the solutions to the primal problem (P^h) and the dual problem (D^H), respectively; then,

$$J_d^H = J_p^h + \rho^{hH},$$

where

$$\rho^{hH} = B(u - u^h, z - z^h) - B(u - u^H, z - z^H),$$

for any $u^H \in U_p^H$ and any $z^h \in V_d^h$.

Proof. The result is a direct consequence of the previous theorem, on subtracting (3.4) from (3.5), and recalling from the Primal–Dual Equivalence Theorem that $J_p = J_d$. □

To conclude this section, let us note, in particular, that if

$$\{U_p^h\}_{h>0} = \{U_p^H\}_{H>0} \quad \text{and} \quad \{V_d^h\}_{h>0} = \{V_d^H\}_{H>0},$$

then $\rho^{hH} = 0$ and there is exact equivalence between the primal and dual formulations. When the families are not the same, in general, the error term ρ^{hH} is not equal to 0, but may be made arbitrarily small by sending the discretization parameters h and H to 0. Indeed,

$$|\rho^{hH}| \leq \gamma_1 \left\{ \|u - u^h\|_U \inf_{v^h \in V_d^h} \|z - v^h\|_V + \|z - z^H\|_V \inf_{w^H \in U_p^H} \|u - w^H\|_U \right\},$$

so ρ^{hH} will converge to 0, as $h, H \to 0$, whenever the four terms on the right-hand side converge to 0 with h and H. If the bounded bilinear form $B(\cdot\,,\cdot)$

is weakly coercive on the appropriate spaces, then this follows from

$$\lim_{h\to 0}\inf_{w^h\in U_p^h}\|u-w^h\|_U=0,\qquad \lim_{h\to 0}\inf_{v^h\in V_d^h}\|z-v^h\|_V=0,$$
$$\lim_{H\to 0}\inf_{w^H\in U_p^H}\|u-w^H\|_U=0,\qquad \lim_{H\to 0}\inf_{v^H\in V_d^H}\|z-v^H\|_V=0,$$

which, in turn, are the standard approximability hypotheses for abstract Galerkin methods.

3.2. Error representation in terms of residuals

We present a further application of the error representation formulae (3.4) and (3.5), concerned with improving the approximation to the output functional by correcting its value. The discussion in this section is an abstract version of the linear theory presented in Pierce and Giles (2000), Giles (2001), Giles and Pierce (2001, 2002) applied to Galerkin approximations. The extension to non-Galerkin approximations is discussed in the following section.

We begin by noting that, for any $u^h\in U_p^h$, and any $v_0\in V_0$, it follows from (2.3) that

$$\begin{aligned}B(u-u^h,v_0)&=B(u,v_0)-B(u^h,v_0)\\&=\ell(v_0)-B(u^h,v_0).\end{aligned}$$

Since $R_p(u^h):v\mapsto \ell(v)-B(u^h,v)$ is a bounded linear functional on V, we can write

$$B(u-u^h,v_0)=\langle R_p(u^h),v_0\rangle\qquad \forall v_0\in V_0, \tag{3.9}$$

where $\langle\cdot\,,\cdot\rangle$ is the duality pairing between V', the dual space of V, and V.

Recalling from (3.4) that

$$J_p-J_p^h=B(u-u^h,z-z^h)\qquad \forall z^h\in V_d^h,$$

where u^h is the second component of the solution $(J_p^h,u^h)\in\mathbb{R}\times U_p^h$ to the primal problem (P^h), it follows that

$$\begin{aligned}J_p-J_p^h&=\langle R_p(u^h),z-z^h\rangle\\&=\langle R_p(u^h),z^H-z^h\rangle+\langle R_p(u^h),z-z^H\rangle\qquad \forall z^h\in V_d^h,\end{aligned}\tag{3.10}$$

where $z^H\in V^H$ is the second component of the solution $(J_d^H,z^H)\in\mathbb{R}\times V_d^H$ to the dual problem (D^H). Taking $z^h=d$, and defining $z_0^H=z^H-d$, we obtain

$$J_p-J_p^h=\langle R_p(u^h),z_0^H\rangle+\langle R_p(u^h),z-z^H\rangle. \tag{3.11}$$

The first term on the right-hand side of (3.11) is computable from u^h, z^H and the data, and can be moved across to the left to yield

$$\begin{aligned} J_p - J_p^{hH} &= \langle R_p(u^h), z - z^H \rangle \\ &= B(u - u^h, z - z^H), \end{aligned} \tag{3.12}$$

where we define J_p^{hH} as

$$\begin{aligned} J_p^{hH} &= J_p^h + \langle R_p(u^h), z_0^H \rangle \\ &= m(u^h) + l(d) - B(u^h, d) + l(z_0^H) - B(u - u^h, z_0^H) \\ &= m(u^h) + l(z^H) - B(u^h, z^H). \end{aligned} \tag{3.13}$$

We shall refer to J_p^{hH} as the corrected functional value.

First, we note that, since

$$\langle R_p(u^h), z_0^h \rangle = \ell(z_0^h) - B(u^h, z_0^h) = 0 \qquad \forall z_0^h \in V_0^h, \tag{3.14}$$

we have that

$$\begin{aligned} |\langle R_p(u^h), z_0 \rangle| &= \inf_{z_0^h \in V_0^h} |\langle R_p(u^h), z_0 - z_0^h \rangle| \\ &= \inf_{z^h \in V_d^h} |\langle R_p(u^h), z - z^h \rangle|. \end{aligned}$$

Hence, on comparing (3.10) with (3.12), we deduce that, if

$$|\langle R_p(u^h), z_0 - z_0^H \rangle| \ll |\langle R_p(u^h), z_0 \rangle|, \tag{3.15}$$

or, equivalently,

$$|\langle R_p(u^h), z - z^H \rangle| \ll \inf_{z^h \in V_d^h} |\langle R_p(u^h), z - z^h \rangle| \tag{3.16}$$

for sufficiently small h and H, then the corrected value J_p^{hH} will represent a more accurate approximation to J_p than J_p^h does.

Clearly, if $z^H \in V_d^h$, then (3.15) does not hold. Thus, to ensure the validity of (3.15) it is necessary to assume that V_d^h and V_d^H differ. Incidentally, this requirement is also reasonable from the computational point of view: since (P) and (D) are driven by different data, their respective solutions u and z will, in general, exhibit different features, and there is no reason to presume that the family of test spaces $\{V_d^h\}_{h>0}$ for the discretization of the primal problem (P) will also be an appropriate choice as a family of trial spaces for the discretization of the dual problem (D).

In the context of h-version finite element methods (Strang and Fix 1973, Ciarlet 1978, Brenner and Scott 1994, Braess 1997) several possible strategies for the selection of V_d^H can be devised. Suppose, for example, that V^h is a finite element space on a subdivision $\mathcal{T}_h$, of granularity h, of the computational domain Ω consisting of (continuous or discontinuous) piecewise polynomials of degree k, and V^H is a finite element space on a possibly different subdivision $\mathcal{T}_H$, of granularity H, of Ω consisting of (continuous or discontinuous) piecewise polynomials of degree K. Below, we list a number of approaches for choosing V_d^H.

(a) Choose $\mathcal{T}_H = \mathcal{T}_h$ and $K > k$, *e.g.*, $K = k + 1$. Thus the numerical solution of the dual problem is performed on the same mesh as for the primal problem, but higher-degree piecewise polynomials are used than for the primal. This approach may be inefficient since the computational cost of solving the dual problem could be considerably larger than that of solving the primal.

(b) Choose $K = k$ and $\mathcal{T}_H = \mathcal{T}_{\lambda h}$ where $\lambda \in (0, 1)$, *e.g.*, $\lambda = 1/2$. Here, the finite element space for the dual is based on a supermesh obtained by global refinement of the primal mesh, and involves piecewise polynomials of the same degree as for the primal. Similarly to (a), the computational cost of solving the dual will be larger than that of solving the primal, although one may benefit from the fact that $K = k$, and therefore the numerical algorithm for the dual is essentially the same as for the primal, albeit on a finer mesh.

(c) Choose $K = k$, and select $\mathcal{T}_H$ adaptively, based on an *a posteriori* error bound for the dual problem. Thereby the mesh $\mathcal{T}_H$ for the dual may be completely different from $\mathcal{T}_h$. In this approach, the dual problem is solved on its own mesh whose choice is governed only by the data for the dual, and not by the choice of the primal mesh. This is perfectly reasonable, since the solutions to the primal and dual problems will in general exhibit completely different behaviour and it is unreasonable to expect that a mesh which is adequate for one will also be appropriate for the other. Of course, a practical drawback is that the adaptive design of the dual mesh $\mathcal{T}_H$ requires additional effort. Further, one needs to transfer information between the different mesh families $\{\mathcal{T}_H\}$ and $\{\mathcal{T}_h\}$ to evaluate the correction term $\langle R(u^h), z_0^H \rangle$.

In the case of *hp*-version finite element methods, which admit variation of both the local mesh size and the local polynomial degree, a further alternative is available:

(d) Choose both $\mathcal{T}_H$ and the local polynomial degree for the dual finite element space adaptively, based on an *a posteriori* error bound. This approach admits even more flexibility in the choice of the dual finite element space V_d^H than (c) – of course, with the added computational cost involved in *hp*-adaptivity in the solution of the dual problem.

Although the discussion in this section was restricted to Galerkin methods, and finite element methods in particular, the idea of error correction is much more general. In order to give a flavour of the scope of the technique, in the next section we consider the question of error correction for a general class of discretization methods for differential equations.

4. Error correction for general discretizations

Now we present a slightly more general version of the argument above, which does not assume that the numerical approximations to u and z stem from a Galerkin type method: it suffices that the approximation u^h to the primal solution $u \in U_p$ is chosen from U_p^h and the approximation z^H to the dual solution $z \in V_d$ is selected from V_d^H; exactly how u^h and z^H are defined is irrelevant. For example, one may suppose that u^h and z^H have been obtained by piecewise polynomial interpolation of finite difference or finite volume approximations to u and z on meshes of size h and H, respectively.

Starting with the equivalence of J_p and J_d, we have from (2.12) that

$$\begin{aligned} J_p &= m(u^h) + \ell(z) - B(u^h, z) \\ &= m(u^h) + \ell(z^H) - B(u^h, z^H) + \ell(z - z^H) - B(u^h, z - z^H). \end{aligned}$$

As

$$\ell(v) = B(u, v) \qquad \forall v \in V_0,$$

on choosing $v = z - z^H \in V_0$ we obtain

$$J_p = m(u^h) + \ell(z^H) - B(u^h, z^H) + B(u - u^h, z - z^H).$$

If we define J_p^h as

$$J_p^h = m(u^h) + l(d) - B(u^h, d),$$

and again define J_p^{hH} as in (3.13) to be

$$J_p^{hH} = J_p^h + \langle R_p(u^h), z_0^H \rangle = m(u^h) + l(z^H) - B(u^h, z^H),$$

we deduce that

$$J_p - J_p^{hH} = B(u - u^h, z - z^H), \tag{4.1}$$

whereas

$$J_p - J_p^h = B(u - u^h, z - z^H) + \langle R_p(u^h), z_0^H \rangle. \tag{4.2}$$

The key point is that if z^H is a very good approximation to z so that

$$|B(u - u^h, z - z^H)| \ll |\langle R_p(u^h), z_0^H \rangle|,$$

then J_p^{hH} will be a much more accurate approximation to J_p than J_p^h.

Next we shall present an experimental illustration of this abstract result. In the example u^h and z^H are computed by means of a finite difference method, in tandem with spline interpolation to construct piecewise polynomial functions from the set of nodal values delivered by the difference scheme.

4.1. *Example 1: Elliptic problem in 1D*

The example concerns the second-order ordinary differential equation

$$\mathcal{L}u \equiv -u'' = f(x), \qquad x \in (0,1),$$

subject to homogeneous Dirichlet boundary conditions

$$u(0) = 0, \qquad u(1) = 0.$$

Let us suppose that the boundary value problem has been solved on a uniform grid,

$$\{x_j = jh \, : \, j = 0, \dots, N\},$$

with spacing $h = 1/N$, $N \geq 2$, using the second-order finite difference scheme

$$-\frac{U_{j+1} - 2U_j + U_{j-1}}{h^2} = f(x_j), \qquad j = 1, \dots, N-1,$$
$$U_0 = 0, \quad U_N = 0.$$

Here U_j denotes the approximation to $u(x_j)$, $j = 0, \dots, N$. We then define u^h by natural cubic spline interpolation through the values U_j, $j = 0, \dots, N$, with end conditions $(u^h)''(0) = -f(0)$, $(u^h)''(1) = -f(1)$.

Let us suppose that the quantity of interest is

$$J_p = \mathcal{J}_p(u) = m(u) = \int_0^1 u(x)g(x)\,\mathrm{d}x,$$

where $g \in L^2(0,1)$ is a given weight function. It follows that the corresponding dual problem is then

$$\mathcal{L}^* z \equiv -\, z'' = g(x), \qquad x \in (0,1),$$
$$z(0) = 0, \quad z(1) = 0.$$

The numerical approximation $z^H = z_0^H$ to z is defined analogously to u^h, with the mesh size H for the dual finite difference scheme taken to be *equal* to h, for simplicity.

With $d = 0$, the uncorrected approximation J_p^h is given by

$$J_p^h = m(u^h)$$

whereas the corrected approximation J_p^{hH} is given by

$$\begin{aligned} J_p^{hH} &= m(u^h) + \ell(z^H) - B(u^h, z^H) \\ &= m(u^h) + \langle R(u^h), z^H \rangle, \end{aligned}$$

where

$$B(w,v) = \int_0^1 w'(x)\, v'(x)\, \mathrm{d}x, \qquad \ell(v) = \int_0^1 f(x)\, v(x)\, \mathrm{d}x,$$

for $w \in U_p = U_0 = H_0^1(0,1)$ and $v \in V_0 = V_d = H_0^1(0,1)$. Now, letting $(\cdot\,,\cdot)$ denote the inner product of $L^2(0,1)$,

$$\begin{aligned} \langle R(u^h), v\rangle &= \ell(v) - B(u^h, v) \\ &= (f, v) - B(u^h, v) \\ &= \int_0^1 f(x)\, v(x)\, \mathrm{d}x - \int_0^1 (u^h)'(x)\, v'(x)\, \mathrm{d}x \\ &= \int_0^1 \Big\{ f(x) + (u^h)''(x) \Big\}\, v(x)\, \mathrm{d}x \qquad \forall v \in H_0^1(0,1), \end{aligned}$$

where we have made use of the fact that u^h is a cubic spline, and therefore $\mathcal{L}u^h$ is continuous on $[0,1]$. Hence,

$$R(u^h) = f + (u^h)'' = f - \mathcal{L}u^h \in L^2(0,1)$$

and

$$\langle R(u^h), z^H\rangle = (R(u^h), z^H),$$

so that

$$J_p^{hH} = m(u^h) + (R(u^h), z^H).$$

A numerical experiment. The aim of the numerical experiment which we shall now perform is to show that the addition of the 'adjoint correction term' $(R(u^h), z^H)$ to $m(u^h)$ is important, in that J_p^{hH} is a more accurate approximation to $m(u)$ than $J_p^h = m(u^h)$ is.

In the numerical experiment, we took

$$f(x) = -x^3(1-x)^3, \qquad g(x) = -\sin(\pi x).$$

Figure 4.1 depicts the residual

$$R(u^h) = f - \mathcal{L}u^h$$

for $h = \frac{1}{32}$, as well as the values at the three Gaussian quadrature points on each subinterval $[x_{j-1}, x_j]$, $j = 1, \ldots, N$, which have been used in the numerical integration of the inner product $(R(u^h), z_0^H) = (R(u^h), z^H)$, with $H = h$. Since u^h is a cubic spline, $\mathcal{L}u^h$ is continuous and piecewise linear. The best piecewise linear approximation to f has an approximation error

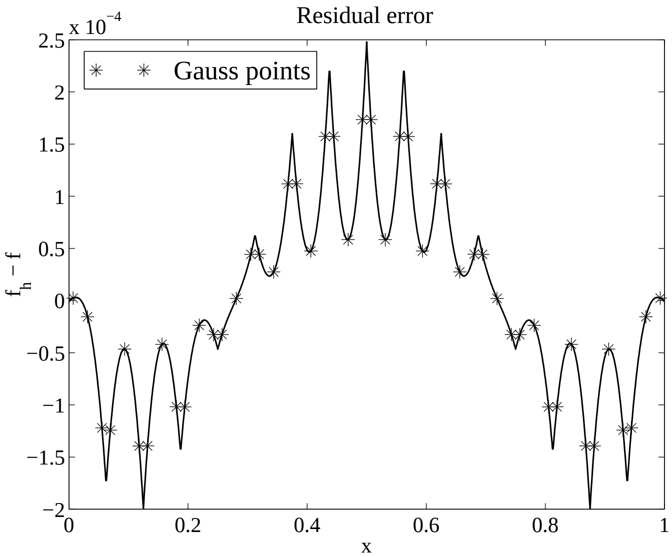


Figure 4.1. Residual $R(u^h)$ for the 1D Poisson equation

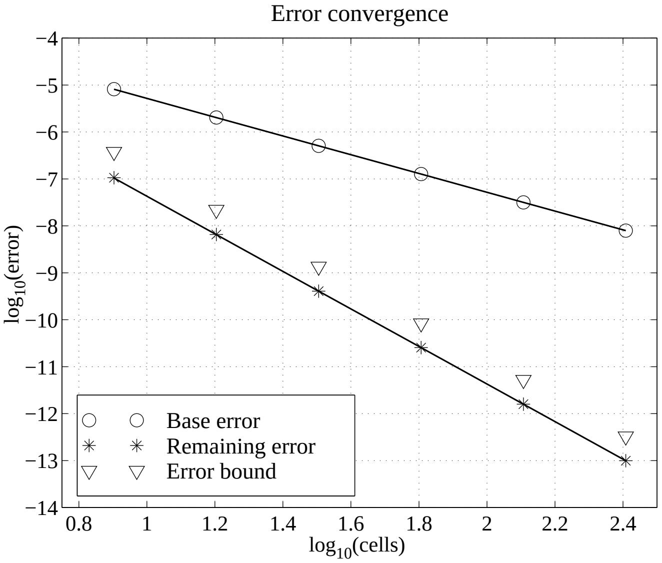


Figure 4.2. Errors in approximating J_p for the 1D Poisson equation

whose dominant term is quadratic on each subinterval; this explains the scalloped shape of $R(u^h)$ in Figure 4.1.

Figure 4.2 is the log-log plot of the three error curves corresponding to:

(a) the 'base error', $|m(u) - m(u^h)|$;

(b) the error $|m(u) - J_p^{hH}|$ resulting after the inclusion of the adjoint correction term $(R(u^h), z^H)$ via $J_p^{hH} = m(u^h) + (R(u^H), z^H)$; and

(c) the error bound $\|\mathcal{L}^{-1}\|\ \|f - \mathcal{L}u^h\|\ \|g - \mathcal{L}z^H\|$ which bounds the magnitude of (b). Here $\mathcal{L}^{-1}$ denotes the inverse of the differential operator $\mathcal{L} : H^2(0,1) \cap H_0^1(0,1) \to L^2(0,1)$.

The superimposed lines have slopes -2 and -4, confirming that the base approximation $m(u^h)$ to $m(u)$ is second-order accurate, while the error in the corrected approximation J_p^{hH} and the error bound are both fourth-order. We note in passing that, on a grid with 16 cells, which might be a reasonable choice for practical computations, the error in the corrected functional value J_p^{hH} is over 200 times smaller than the uncorrected error.

To conclude this experiment, let us explain why the corrected functional value J_p^{hH} converges to the analytical value $J_p = \mathcal{J}_p(u) = m(u)$ as $\mathcal{O}(h^4)$ when $h \to 0$. According to (4.1),

$$J_p - J_p^{hH} = B(u - u^h, z - z^H), \tag{4.3}$$

where, in the present experiment, u^h is the cubic spline interpolant based on the finite difference approximation U_j, $j = 0, \dots, N$, to the nodal values $u(x_j)$, $j = 0, \dots, N$, on a uniform mesh of size h; z^H is defined analogously, on a mesh of size $H = h$. Since U approximates u with $\mathcal{O}(h^2)$ error in the discrete L^2-norm based on the internal nodes, and the first-order central difference quotient of U approximates u' with $\mathcal{O}(h^2)$ error in the same norm, it follows that

$$\|u - u^h\|_U = \|u' - (u^h)'\|_{L^2(0,1)} = \mathcal{O}(h^2).$$

Analogously, with $H = h$,

$$\|z - z^H\|_V = \|z' - (z^H)'\|_{L^2(0,1)} = \mathcal{O}(h^2).$$

Thus we deduce from (4.3) that

$$|J_p - J_p^{hH}| = \mathcal{O}(h^4),$$

as required. For further details we refer to the paper of Giles and Pierce (2001).

4.2. Example 2: Elliptic problem in 2D

This example concerns the 2D Laplace equation

$$\begin{aligned} -\Delta u &= 0 \qquad \text{in } \Omega, \\ u &= g \qquad \text{on } \Gamma = \partial\Omega. \end{aligned}$$

The output of interest is a weighted integral of the normal flux

$$\mathcal{J}_p(u) = \int_\Gamma \frac{\partial u}{\partial \nu}\, \psi \,\mathrm{d}s.$$

As described in Section 2.1, the associated primal problem uses

$$B(u,v) = \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x, \qquad \ell(u) = 0, \qquad m(v) = 0.$$

The definition of J_p^h required the selection of an appropriate d. If, for example, u^h is twice continuously differentiable, we may integrate by parts to deduce that

$$-B(u^h, d) = (\Delta u^h, d) + \left\langle \frac{\partial u^h}{\partial \nu}, \psi \right\rangle_\Gamma .$$

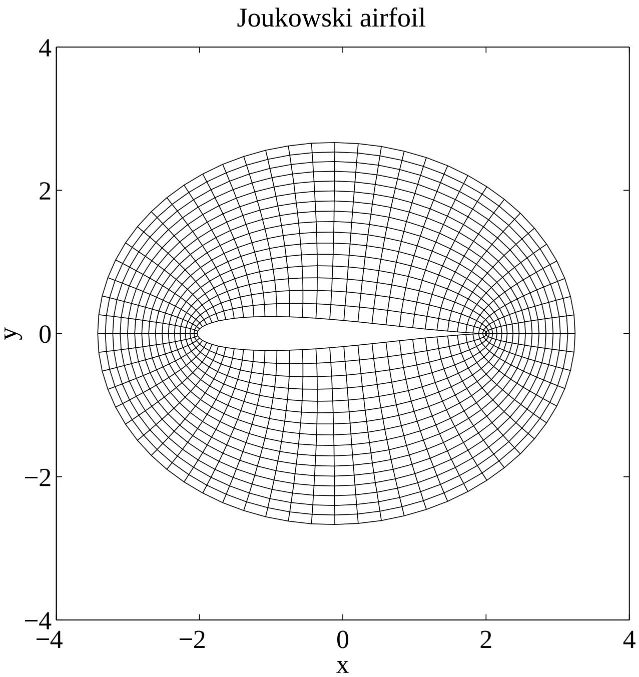


Figure 4.3. The computational grid for a 2D Laplace problem

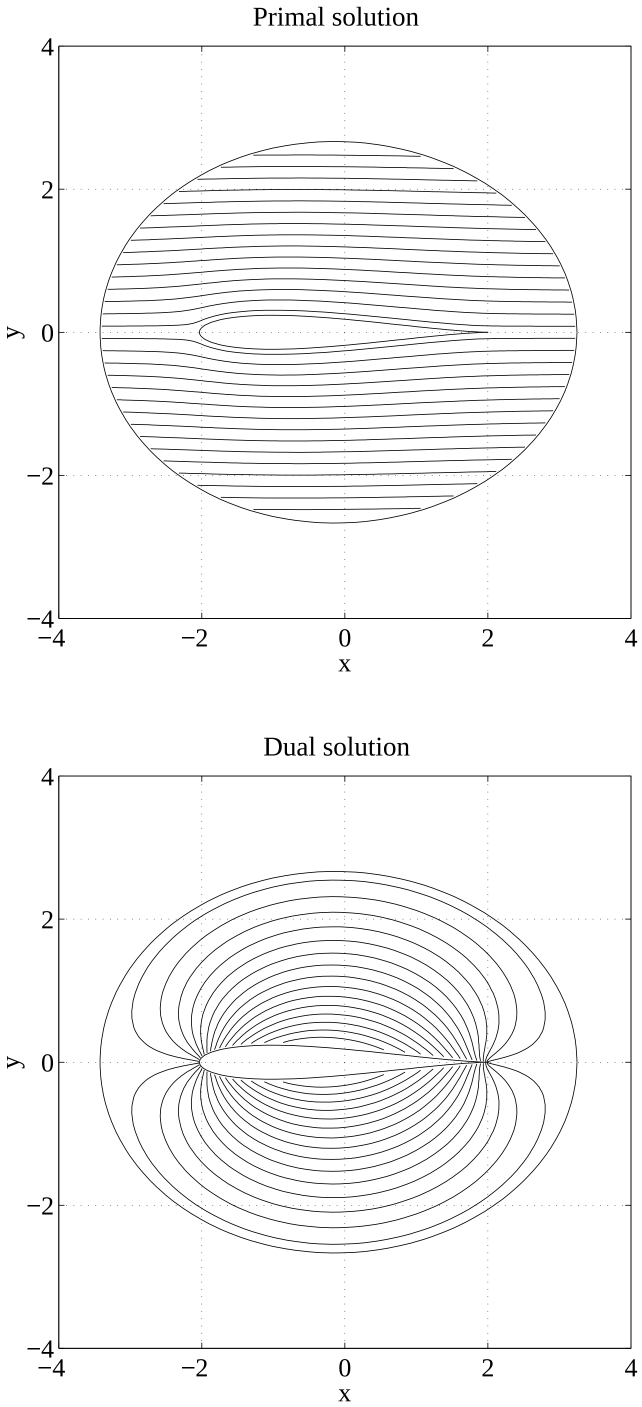


Figure 4.4. The approximate solutions u^h and z^H for a 2D Laplace problem

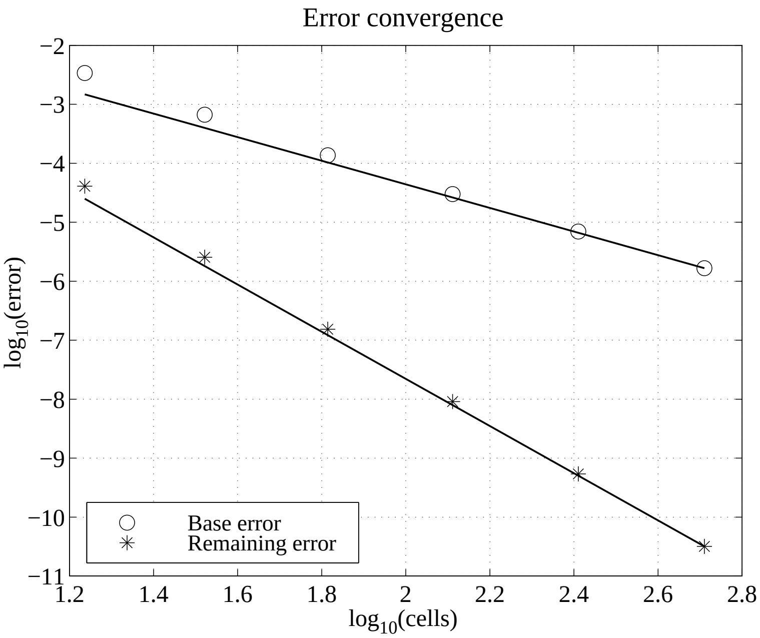


Figure 4.5. Error convergence for a 2D Laplace problem

Thus, we obtain $J_p^h = \mathcal{J}_p(u^h)$ if we choose d to equal 0 on the interior of Ω, and $-\psi$ on Γ. Strictly speaking, this violates the condition that $d \in H^1(\Omega)$, and so $(\Delta u^h, d)$ should be considered to be the limit of an appropriate sequence $(\Delta u^h, d_n)$ where $d_n \in H^1(\Omega)$, $n = 1, 2, \dots$.

The corrected functional is then given by

$$J_p^{hH} = -B(u^h, z^H) = J_p^h + \langle R_p(u^h), z_0^H \rangle = J_p^h + (-\Delta u^h, z^H).$$

The numerical results are obtained for a test problem for which the analytic solution has been constructed by a conformal mapping. An initial Galerkin finite element approximation u^{FE} to u is obtained using bilinear shape functions on the computational grid shown in Figure 4.3. The new approximate solution u^h is then obtained by bicubic spline interpolation through the values of u^{FE} at the grid nodes. The approximate dual solution z^H is obtained similarly using the same computational grid.

The errors in the functional are shown in Figure 4.5. The superimposed lines of slope -2 and -4 show that the base value for the functional, J_p^h, is again second-order accurate whereas the corrected value, J_p^{hH}, is fourth-order accurate. This improvement in the order of accuracy is achieved despite the presence of the singularity in the dual solution.

5. Linear defect error correction

Adjoint error correction is not the only means of improving the accuracy of numerical calculations. In this section, based on Giles (2001), we look at the use of defect correction (Barrett, Moore and Morton 1988, Koren 1988, Skeel 1981, Stetter 1978), and show that it can be extremely effective in reducing the errors in a model 1D Helmholtz problem; the combination of defect and adjoint error correction is even more accurate.

The primary motivation for this investigation is the need for high-order accuracy for aeroacoustic and electromagnetic calculations. In steady CFD calculations, grid adaptation can be used to provide high grid resolution in the limited areas that require it. However, using standard second-order accurate methods, the wave-like nature of aeroacoustic and electromagnetic solutions would lead to grid refinement throughout the computational domain in order to reduce the wave dispersion and dissipation to acceptable levels. The preferable alternative is to use higher-order methods, allowing one to use fewer points per wavelength, which can lead to a very substantial reduction in the total number of grid points for three-dimensional calculations. The difficulty with this is that one often wants to use unstructured grids because of their geometric flexibility, and the construction of higher order approximations on unstructured grids is complicated and computationally expensive.

5.1. General approach for Galerkin approximations

We start with a problem whose weak formulation is to find $u \in U_p$ such that

$$B(u,v) = \ell(v) \qquad \forall v_0 \in V_0,$$

and suppose that we have a Galerkin discretization which defines $u^h \in U_p^h$ to be the solution of

$$B(u^h, v^h) = \ell(v_0^h) \qquad \forall v_0^h \in V_0^h.$$

Next, suppose that we have a method for defining a reconstructed approximation u_R^h from u^h. The purpose of this reconstruction, as with the reconstruction in the last numerical example in the previous section, is to maintain the order of accuracy in the L^2-norm, and improve it in the H^1-norm.

Writing $u = u_R^h + e$ gives

$$B(e,v) = \ell(v) - B(u_R^h, v) \qquad \forall v_0 \in V_0.$$

The error $e \in U_0$ can then be approximated by $e^h \in U_0^h$, which is the solution of

$$B(e^h, v^h) = \ell(v_0^h) - B(u_R^h, v^h) \qquad \forall v_0^h \in V_0^h.$$

An improved value for u^h_R can then be defined by reconstruction from $u^h + e^h$ or, equivalently, by adding the reconstructed error e^h_R to u^h_R. The entire process may then be repeated to further improve the accuracy. This follows the procedure described by Barrett *et al.*, who also showed that it can converge to a solution of an appropriately defined Petrov–Galerkin discretization (Barrett *et al.* 1988).

5.2. 1D Helmholtz problem

The model problem to be solved is the 1D Helmholtz equation

$$-u'' - \pi^2 u = 0, \quad x \in \Omega = (0, 10),$$

subject to the Dirichlet boundary condition $u = 1$ at $x = 0$ and the radiation boundary condition $u' - \mathrm{i}\pi u = 0$ at $x = 10$. The analytic solution is $u = \exp(\mathrm{i}\pi x)$ and the domain contains precisely five wavelengths. The output functional of interest is the value $u(10)$ at the right-hand boundary. This can be viewed as a model of a far-field boundary integral giving the radiated acoustic energy in aeroacoustics, or the radar cross-section in electromagnetics (Monk and Süli 1998).

Multiplying by $\bar{v}$ (the complex conjugate of v) and integrating by parts yields the weak form: find $u \in U_p$ such that

$$B(u, v) = \ell(v) \qquad v \in V_0,$$

where

$$B(u, v) = (u', v') - \pi^2 (u, v) - \mathrm{i}\pi u(10)\bar{v}(10) \qquad \ell(v) = 0,$$

and

$$U_p = \{u \in H^1(\Omega) : u(0) = 1\},$$
$$V_0 = \{v \in H^1(\Omega) : v(0) = 0\}.$$

Note that the inner product $(\cdot\,, \cdot)$ is now a complex inner product

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

With this change, the theory presented before for real-valued functions extends naturally to complex-valued functions.

Using a piecewise linear Galerkin discretization, $u^h \in U^h_p$ is defined to be the solution of

$$B(u^h, v^h) = \ell(v^h_0) \qquad \forall v^h_0 \in V^h_0.$$

It is well established that this discretization is second-order convergent in $L^2(\Omega)$, producing dispersion but no dissipation on a uniform grid.

The reconstructed solution u^h_R is defined by cubic spline interpolation of the nodal values $u^h(x_j)$. The choice of end conditions for the cubic spline is very important. A natural cubic spline would have $(u_R^h)''=0$ at both ends, but this would introduce small but significant errors at each end since $u''\neq 0$ for the analytic solution. Instead, at $x=10$ we require the splined solution to satisfy the analytic boundary condition by imposing $(u_R^h)'-i\pi u_R^h=0$. At $x=0$, the analytic boundary condition is already imposed through having the correct value for the end-point $U(0)$. Therefore, here we require that $(u_R^h)''+\pi^2 u_R^h=0$, so the splined solution satisfies the original ordinary differential equation at the boundary.

The error e and its Galerkin approximation e^h are defined according to the general approach described above. The reconstruction e_R^h is again obtained by cubic spline interpolation, with the same end conditions.

5.3. Adjoint error correction

The output functional of interest is

$$\mathcal{J}(u)=u(10),$$

so we define

$$m(u)=u(10), \qquad \ell(v)=0,$$

to obtain

$$J_p=m(u)+\ell(v)-B(u,v) \qquad \forall v\in V_0.$$

The Galerkin approximation z^h to the dual solution is the piecewise linear solution $z^h\in V_0^h$ for which

$$B(w^h,z^h)=m(w^h) \qquad \forall w^h\in U_0^h.$$

Defect correction can also be applied to the dual solution.

5.4. Numerical results

Numerical results have been obtained for grids with 4, 8, 16, 32, 64 and 128 points per wavelength. To test the ability to cope with irregular grids, the coordinates for the grid with N intervals are defined as

$$x_0=0, \quad x_N=10, \quad x_j=\frac{10}{N}\left(j+\sigma_j\right), \quad 0<j<N,$$

where σ_j is a uniformly distributed random variable in the range $[-0.3,0.3]$.

Figure 5.1 shows the L^2 norm of the error in the reconstructed cubic spline solution before and after defect correction. Without defect correction, the error is second-order, while with defect correction it is fourth-order. Note that a second application of defect correction makes a significant reduction in the error even though it remains fourth-order. This is because one application of the defect correction procedure gives a correction that is second-order in magnitude, with a corresponding error that is second-order in relative

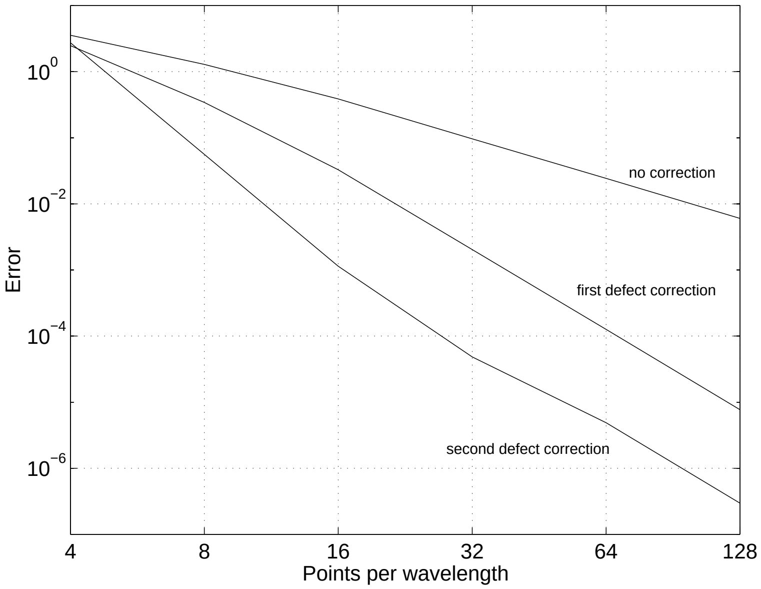


Figure 5.1. L^2 error in the numerical approximation to $u(x)$

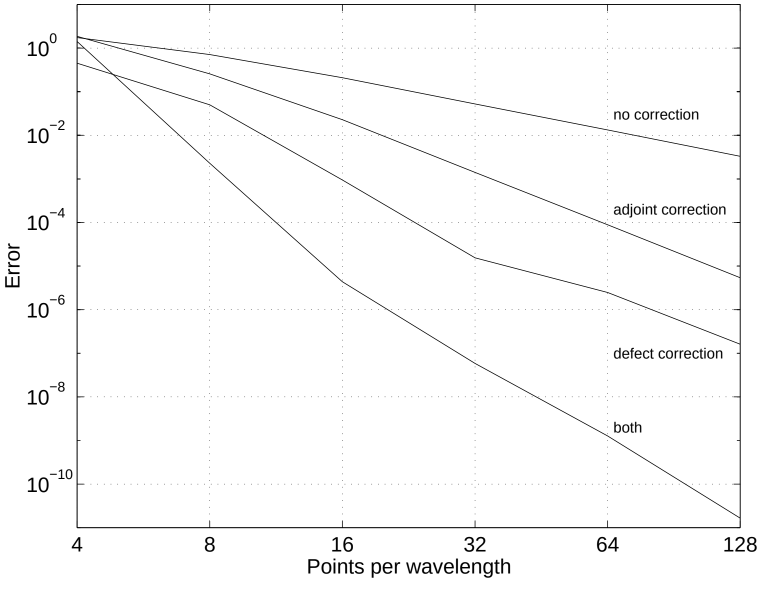


Figure 5.2. Error in the numerical approximation to $u(10)$

magnitude and therefore fourth-order in absolute magnitude. It is this error that is corrected by a second application of the defect correction procedure.

Figure 5.2 shows the error in the numerical value for the output functional $u(10)$. Without any correction, the error is second-order. Using either defect correction or adjoint error correction on their own increases the order of accuracy to fourth-order, but using them both increases the accuracy to sixth-order. Note that the calculation with 8 points per wavelength plus both defect and adjoint error correction gives an error which is approximately 2×10^{-3}. This is more accurate than the calculation with 128 points per wavelength and no corrections, and comparable to the results using 14 points and defect correction, or 30 points with adjoint error correction.

In 3D, the computational cost is proportional to the cube of the number of points per wavelength, so this indicates the potentially huge savings offered by the combination of defect and adjoint error correction. The cost of computing the corrections is five times the cost of the original calculation, due to the additional two calculations for the defect correction, and the one adjoint calculation plus its two defect corrections. In practice, the second defect correction for the primal and adjoint calculations make negligible difference to the value obtained after the adjoint error correction, so these can be omitted, reducing the cost of the corrections to just three times the cost of the original calculation.

6. Reconstruction with biharmonic smoothing

Up to now, we have relied on using cubic spline reconstruction in one space dimension, or its tensor-product version on multidimensional Cartesian product meshes. Here, we consider an alternative reconstruction technique that is applicable on more general nonuniform triangulations.

Suppose that we have a partial differential equation with analytic solution $u \in H^5(\Omega)$, $\Omega = (0,1)^n$, with periodic boundary conditions.

Let u^h be a numerical approximation to u with

$$\|u - u^h\|_{L^2(\Omega)} = \mathcal{O}(h^2).$$

Nothing is assumed about the accuracy of ∇u^h, but in practice, if u^h has come from a piecewise linear finite element approximation, then it will be only a first-order accurate approximation to ∇u in the $L^2(\Omega)$-norm.

We now define a new approximate solution $\tilde{u}$ by

$$h^s \, \Delta^2 \tilde{u} + \tilde{u} = u^h \tag{6.1}$$

subject to periodic boundary conditions on $\Omega = (0,1)^n$. The purpose of the biharmonic term is to smooth the solution to improve the order of accuracy of the derivative $\nabla \tilde{u}$.

6.1. *General analysis*

It follows immediately that

$$h^s\,\Delta^2(\tilde{u}-u)+(\tilde{u}-u)=(u^h-u)-h^s\,\Delta^2 u.$$

The analysis proceeds by splitting the error into two components

$$\tilde{u}-u=e+f,$$

with

$$h^s\,\Delta^2 e+e=u^h-u, \tag{6.2}$$

and

$$h^s\,\Delta^2 f+f=-h^s\,\Delta^2 u, \tag{6.3}$$

where e and f satisfy periodic boundary conditions.

Considering equation (6.3) first, multiplying by f and integrating by parts gives

$$h^s\,\|\Delta f\|^2_{L^2(\Omega)}+\|f\|^2_{L^2(\Omega)}=-h^s\,(\Delta^2 u,f)\ \le\ \frac{1}{2}\big(\|h^s\,\Delta^2 u\|^2_{L^2(\Omega)}+\|f\|^2_{L^2(\Omega)}\big),$$

and hence

$$\|f\|_{L^2(\Omega)}\ \le\ h^s\,\|\Delta^2 u\|_{L^2(\Omega)}.$$

Furthermore, taking the gradient of equation (6.3), multiplying both sides by ∇f, and integrating by parts yields similarly that

$$\|\nabla f\|_{L^2(\Omega)}\ \le\ h^s\,\|\Delta^2\nabla u\|_{L^2(\Omega)}.$$

Thus,

$$\|f\|_{H^1(\Omega)}=\mathcal{O}(h^s).$$

Turning now to equation (6.2), multiplying by e and integrating by parts yields

$$h^s\,\|\Delta e\|^2_{L^2(\Omega)}+\|e\|^2_{L^2(\Omega)}\ =\ (u^h-u,e)\ \le\ \frac{1}{2}\big(\|u^h-u\|^2_{L^2(\Omega)}+\|e\|^2_{L^2(\Omega)}\big),$$

and hence

$$\|e\|_{L^2(\Omega)}\ \le\ \|u^h-u\|_{L^2(\Omega)},$$

and

$$h^{s/2}\,\|\Delta e\|_{L^2(\Omega)}\ \le\ \|u^h-u\|_{L^2(\Omega)}.$$

Hence, by the Gagliargo–Nirenberg inequality, we deduce that

$$h^{s/4}\,\|\nabla e\|_{L^2(\Omega)}\ \le\ \|u^h-u\|_{L^2(\Omega)},$$

and thus

$$\|\nabla e\|_{H^1(\Omega)}\ \le\ \mathcal{O}(h^{2-s/4}).$$

Combining the bounds for e and f yields the final result that

$$\|\tilde{u}-u\|_{H^1(\Omega)} \;\leq\; \mathcal{O}(h^p), \quad p=\min(s,2-s/4).$$

The power p is maximized when $s=\frac{8}{5}$, giving $p=\frac{8}{5}$. On the other hand, choosing $s=2$ gives $p=\frac{3}{2}$.

6.2. Fourier analysis

Because we assume periodic boundary conditions, the error component e can be calculated through Fourier analysis. Writing

$$u^h-u=\sum_{\boldsymbol{n}\in\mathbb{Z}^n} a_{\boldsymbol{n}}\exp(2\pi i\ \boldsymbol{n}\cdot\boldsymbol{x}),$$

with $a_{-\boldsymbol{n}}=\bar{a}_{\boldsymbol{n}}$, it follows that

$$e=\sum_{\boldsymbol{n}\in\mathbb{Z}^n} G_{\boldsymbol{n}}a_{\boldsymbol{n}}\exp(2\pi i\ \boldsymbol{n}\cdot\boldsymbol{x}),$$

where

$$G_{\boldsymbol{n}}=\frac{1}{1+16\pi^4h^s|\boldsymbol{n}|^4}.$$

Now,

$$\|u^h-u\|^2_{L^2(\Omega)}=\sum_{\boldsymbol{n}\in\mathbb{Z}^n}|a_{\boldsymbol{n}}|^2=\mathcal{O}(h^4),$$

and

$$\|e\|^2_{H^1(\Omega)}=\sum_{\boldsymbol{n}\in\mathbb{Z}^n}H_{\boldsymbol{n}}|a_{\boldsymbol{n}}|^2,$$

where

$$H_{\boldsymbol{n}}=\left(1+4\pi^2|\boldsymbol{n}|^2\right)G^2_{\boldsymbol{n}}=\frac{1+4\pi^2|\boldsymbol{n}|^2}{(1+16\pi^4h^s|\boldsymbol{n}|^4)^2}.$$

When $|\boldsymbol{n}|=\mathcal{O}(1)$, $H_{\boldsymbol{n}}=\mathcal{O}(1)$, and when $|\boldsymbol{n}|=\mathcal{O}(h^{-1})$, $H_{\boldsymbol{n}}=\mathcal{O}(h^{6-2s})$. Hence, provided $s\leq 3$ and most of the 'energy' of u^h-u is contained in the lowest and highest wave numbers, then

$$\|e\|_{H^1(\Omega)}=\mathcal{O}(h^2).$$

However, $H_{\boldsymbol{n}}$ is a maximum when $|\boldsymbol{n}|=\mathcal{O}(h^{-s/4})$, in which case $H_{\boldsymbol{n}}=\mathcal{O}(h^{-s/2})$. If all of the 'energy' of u^h-u is at these wavelengths, then

$$\|e\|_{H^1(\Omega)}=\mathcal{O}(h^{2-s/4}),$$

which corresponds to the bound obtained from the general analysis.

Thus, the general analysis gives error bounds which are tight with respect to the order of accuracy, but this order is only achieved if most of the initial solution error is in a certain intermediate wave number range. In practice, it seems more likely that the initial solution error will lie in the lowest and highest wave numbers, in which case we will obtain

$$\|\tilde{u} - u\|_{H^1(\Omega)} = \mathcal{O}(h^2),$$

if we use $s = 2$.

7. *A posteriori* error estimation by duality

The purpose of this section is to develop another application of duality. Here we shall be concerned with the derivation of *a posteriori* bounds on the error in an output functional of the solution to a differential equation. For recent surveys of the subject of *a posteriori* estimation, see Eriksson, Estep, Hansbo and Johnson (1995), Süli (1998), Becker and Rannacher (2001), and the monographs of Ainsworth and Oden (2000) and Verfürth (1996). In order to motivate the key ideas, it is helpful to begin with a specific example, following Giles *et al.* (1997).

7.1. Elliptic model problem: approximation of the normal flux

Let Ω be a bounded domain in $\mathbb{R}^n$ with Lipschitz-continuous boundary Γ. Given that ψ is an element of $H^{1/2}(\Gamma)$, we let $H^1_{-\psi}(\Omega)$ denote the space of all v in $H^1(\Omega)$ whose trace, $\gamma_{0,\Gamma}(v)$, on Γ is equal to $-\psi$.

We consider the boundary value problem

$$-\nabla \cdot \boldsymbol{\sigma}(u) = f \qquad \text{in } \Omega, \qquad u = 0 \quad \text{on } \Gamma, \tag{7.1}$$

where $f \in L^2(\Omega)$ and $\boldsymbol{\sigma}(u) = A\nabla u$, with A an $n \times n$ matrix-function, uniformly positive definite on $\bar{\Omega}$, with continuous real-valued entries defined on $\bar{\Omega}$. This problem has a unique weak solution $u \in H^1_0(\Omega)$, satisfying

$$B(u, v) = (f, v) \qquad \forall v \in H^1_0(\Omega). \tag{7.2}$$

Here and below $(\cdot\,,\cdot)$ denotes the inner product in $L^2(\Omega)$ and

$$B(v, w) = (\boldsymbol{\sigma}(v), \nabla w) = (\nabla v, A^{\mathrm{T}}\nabla w),$$

for $v, w \in H^1(\Omega)$, where A^{T} is the transpose of A.

Let us suppose that the quantity of interest in the outward normal flux through Γ defined by

$$N(u) = \int_\Gamma \boldsymbol{\nu} \cdot \boldsymbol{\sigma}(u)\,\mathrm{d}s,$$

where $\boldsymbol{\nu}$ denotes the unit outward normal vector to Γ. In order to compute

$N(u)$, for $u \in H_0^1(\Omega)$ denoting the weak solution to problem (7.1) and $\psi \in H^{1/2}(\Gamma)$, we consider the slightly more general problem of computing the weighted normal flux through the boundary, defined by

$$\mathcal{J}_p(u) = N_\psi(u) = \int_\Gamma \boldsymbol{\nu} \cdot \boldsymbol{\sigma}(u)\, \psi \,\mathrm{d}s. \tag{7.3}$$

We note that, since $\boldsymbol{\sigma}(u) \in [L^2(\Omega)]^n$ and $\nabla \cdot \boldsymbol{\sigma}(u) \in L^2(\Omega)$, according to the Trace Theorem for the function space $H(\mathrm{div}, \Omega)$ (see Theorem 2.2 in Girault and Raviart (1986)), the normal stress $\boldsymbol{\nu} \cdot \boldsymbol{\sigma}(u)|_\Gamma$ is correctly defined as an element of $H^{-1/2}(\Gamma)$, and $N_\psi(u)$ is meaningful, provided that the integral over Γ is interpreted as a duality pairing between $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$. Moreover, applying a generalization of Green's identity (see Theorem 2.2 in Girault and Raviart (1986)), we deduce that, for any $v \in H^1_{-\psi}(\Omega)$,

$$\begin{aligned} N_\psi(u) &= (f, v) - (\boldsymbol{\sigma}(u), \nabla v) \\ &= (f, v) - B(u, v). \end{aligned} \tag{7.4}$$

Because of (7.2), the value of the expression $(f, v) - B(u, v)$ on the right-hand side of (7.4) is independent of the choice of $v \in H^1_{-\psi}(\Omega)$. Thus, (7.4) can be interpreted as an equivalent (and correct) definition of the weighted normal flux (7.3) of u across Γ.

Next, we construct our finite element approximation to $N_\psi(u)$. We consider a family of finite-dimensional Galerkin trial spaces U^h and test spaces $V^h = U^h$ contained in $H^1(\Omega)$, which consist of continuous piecewise polynomials of degree k defined on a family of regular subdivisions $\mathcal{T}_h$ of Ω into open n-dimensional simplices κ. We denote the diameter of a simplex $\kappa \in \mathcal{T}_h$ by h_κ and assume that the family $\{\mathcal{T}_h\}$ is shape-regular, that is, there exists a positive constant c such that $\mathrm{meas}(\kappa) \geq c\, h_\kappa^n$ for all $\kappa \in \mathcal{T}_h$ and all $\mathcal{T}_h$, where $\mathrm{meas}(\kappa)$ is the n-dimensional volume of κ. Further, for each function $\psi \in H^{1/2}(\Gamma)$ such that $-\psi = v|_\Gamma$ for some $v \in U^h$, we let $U^h_{-\psi} \subset U^h$ be the space of all $w \in U^h$ with $w|_\Gamma = -\psi$. In particular, U_0^h is the space of all $v \in U^h$ which vanish on Γ. Clearly, $U_0^h \subset H_0^1(\Omega)$.

The finite element approximation of (7.1) is defined as follows: find $u^h \in U_0^h$ such that

$$B(u^h, v^h) = (f, v^h) \qquad \text{for all } v^h \in U_0^h. \tag{7.5}$$

Motivated by the identity (7.4), we define the approximation $N_\psi^h(u^h)$ to $N_\psi(u)$ as follows:

$$N_\psi^h(u^h) = (f, v^h) - B(u^h, v^h), \quad v^h \in U^h_{-\psi}. \tag{7.6}$$

We note that, because of (7.5), $N^h_\psi(u^h)$ is independent of the choice of $v^h \in U^h_{-\psi}$. Furthermore, we observe that, in general,

$$N^h_\psi(u^h) \neq \int_\Gamma \boldsymbol{\nu} \cdot \boldsymbol{\sigma}(u^h)\,\psi\,\mathrm{d}s = N_\psi(u^h),$$

in contrast with identity (7.4) satisfied by the analytical solution u. This raises the question as to which of $N_\psi(u^h)$ and $N^h_\psi(u^h)$ is the more accurate approximation to $N_\psi(u)$. As we shall see, the answer to this question is: $N^h_\psi(u^h)$. Indeed, the error estimate in Theorem 7.2 below shows that, for sufficiently smooth data and continuous piecewise polynomial finite elements of degree k, the order of convergence of $N^h_\psi(u^h)$ to $N_\psi(u)$ is $\mathcal{O}(h^{2k})$. In general, this high order of convergence cannot be achieved by using the 'naive' approximation $N_\psi(u^h) = \int_\Gamma \boldsymbol{\nu} \cdot \boldsymbol{\sigma}(u^h)\,\mathrm{d}s$.

In order to derive a representation formula for the error $N_\psi(u) - N^h_\psi(u^h)$ in the boundary flux, we introduce the following dual problem in variational form: find $z \in H^1_{-\psi}(\Omega)$ such that

$$B(v, z) = 0 \qquad \text{for all } v \in H^1_0(\Omega). \tag{7.7}$$

Consider the global error $e = u - u^h$. Upon setting $v = e$ in (7.7) we obtain

$$0 = B(e, z) = B(e, z - \pi^h z) + B(e, \pi^h z), \tag{7.8}$$

where we made use of the fact that the error e is zero on the boundary Γ; here $\pi^h : H^1_{-\psi}(\Omega) \to U^h_{-\psi}$ is a linear operator satisfying the approximation property (7.10) below. Since the definitions of $N_\psi(u)$ and $N^h_\psi(u^h)$ are independent of the choice of $v \in H^1_{-\psi}(\Omega)$ and $v^h \in U^h_{-\psi}$, respectively, we deduce that

$$\begin{aligned} B(e, \pi^h z) &= \big((f, \pi^h z) - B(u^h, \pi^h z)\big) - \big((f, \pi^h z) - B(u, \pi^h z)\big) \\ &= N^h_\psi(u^h) - N_\psi(u). \end{aligned}$$

On substituting this into (7.8) we arrive at the error representation formula

$$\begin{aligned} N_\psi(u) - N^h_\psi(u^h) &= B(e, z - \pi^h z) \\ &= B(u, z - \pi^h z) - B(u^h, z - \pi^h z) \\ &= (f, z - \pi^h z) - B(u^h, z - \pi^h z) \end{aligned} \tag{7.9}$$

where, to obtain the last equality, we made use of the fact that u obeys (7.2) and $z - \pi^h z$ belongs to $H^1_0(\Omega)$.

Our aim is to investigate the problem of approximating $N_\psi(u)$ from two points of view. First, we analyse the convergence rate of the approximation through an *a priori* error analysis following Babuška and Miller (1984*b*) and Barrett and Elliott (1987); we shall then perform an *a posteriori* error analysis and highlight the relevance of the *a posteriori* error bound for adaptive mesh refinement.

***A priori* error analysis.** For the purposes of the *a priori* error analysis we assume that there exists a linear operator $\pi^h : H^1_{-\psi}(\Omega) \to U^h_{-\psi}$ and a positive constant c such that

$$|v - \pi^h v|_{H^1(\Omega)} \leq c h^{s-1} |v|_{H^s(\Omega)}, \quad 1 \leq s \leq k+1, \tag{7.10}$$

for all $v \in H^s(\Omega)$, $1 \leq s \leq k+1$, and all $h = \max_{\kappa \in \mathcal{T}_h} h_\kappa$ (≤ 1, say).

The next theorem is a direct consequence of inequality (3.7) from Corollary 3.2.

Theorem 7.1. Assume that (7.10) holds, $u \in H^s(\Omega) \cap H^1_0(\Omega)$, $s \geq 1$, and $z \in H^t(\Omega) \cap H^1_0(\Omega)$, $t \geq 1$, where u and z are the solutions to (7.2) and (7.7), respectively. Then,

$$|N_\psi(u) - N^h_\psi(u^h)| \leq c h^{\sigma+\tau-2} |u|_{H^\sigma(\Omega)} \, |z|_{H^\tau(\Omega)}, \tag{7.11}$$

where $1 \leq \sigma \leq \min(s, k+1)$, $1 \leq \tau \leq \min(t, k+1)$, c is a constant, and $h = \max_{\kappa \in \mathcal{T}_h} h_\kappa$.

Proof. It follows from the error representation formula (7.9) that

$$|N_\psi(u) - N^h_\psi(u^h)| \leq c |e|_{H^1(\Omega)} \, |z - \pi^h z|_{H^1(\Omega)}.$$

A standard energy-norm error estimate gives

$$|e|_{H^1(\Omega)} \leq c h^{\sigma-1} |u|_{H^\sigma(\Omega)}, \quad 1 \leq \sigma \leq \min(s, k+1).$$

Further, using the approximation property (7.10), we obtain

$$|z - \pi^h z|_{H^1(\Omega)} \leq c h^{\tau-1} |z|_{H^\tau(\Omega)}, \quad 1 \leq \tau \leq \min(t, k+1),$$

Consequently,

$$|N_\psi(u) - N^h_\psi(u^h)| \leq c h^{\sigma+\tau-2} |u|_{H^\sigma(\Omega)} |z|_{H^\tau(\Omega)},$$

for $1 \leq \tau \leq \min(t, k+1)$, $1 \leq \sigma \leq \min(s, k+1)$. □

A numerical experiment. We include a numerical experiment to illustrate the point that $N^h_\psi(u^h)$ is a more accurate approximation to $N_\psi(u)$ than $N(u^h)$ is. Let us consider Laplace's equation in cylindrical polar coordinates given by

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{\partial^2 u}{\partial z^2} = 0, \tag{7.12}$$

with boundary conditions

$$\begin{aligned} u &= 0, && r \le 1,\ z = 0, \\ \frac{\partial u}{\partial n} &= 0, && r > 1,\ z = 0, \\ & && r = 0,\ z \ge 0, \\ u &= 1, && r,\ z \to \infty, \end{aligned} \tag{7.13}$$

and suppose that the quantity of interest is the weighted normal flux through a part of the boundary so that

$$N_\psi(u) = \int_\Gamma \frac{\partial u}{\partial \nu} \psi r \, \mathrm{d}s, \tag{7.14}$$

where

$$\psi = \begin{cases} -\frac{\pi}{2} & 0 \le r \le 1, \quad z = 0, \\ 0 & \text{elsewhere on } \Gamma. \end{cases} \tag{7.15}$$

It can be shown that the exact solution is $N_\psi(u) = 1$. As described above, we may rewrite $N_\psi(u)$ as

$$N_\psi(u) = -\int_\Omega \nabla u \cdot \nabla v \, r \mathrm{d}r \, \mathrm{d}z \tag{7.16}$$

for any $v \in H^1_{-\psi}(\Omega)$ (where $H^1_{-\psi}(\Omega)$ is defined as before, except that the measure is now $r\mathrm{d}r\,\mathrm{d}z$ rather than $\mathrm{d}x\,\mathrm{d}y$), and we denote the corresponding numerical approximation to (7.16) by $N^h_\psi(u^h)$. To show that $N^h_\psi(u^h)$ is a more accurate approximation to $N_\psi(u)$ than $N_\psi(u^h)$, we consider the convergence of these two quantities to $N_\psi(u) = 1$ on a sequence of regular meshes of mesh size h using a piecewise linear finite element method to compute u^h. The results are shown in Table 7.1, from which it is clear that the approximation $N^h_\psi(u^h)$ converges to $N_\psi(u)$ at over twice the rate at which $N_\psi(u^h)$ converges to $N_\psi(u)$; note that $u \in H^{3/2-\varepsilon}(\Omega)$, $\varepsilon > 0$, thus leading to approximately first-order convergence by virtue of our *a priori* error estimate from the last theorem.

Table 7.1. The numerical approximations $N^h_\psi(u^h)$ and $N_\psi(u^h)$ to $N_\psi(u)$

h	$N_\psi(u^h)$	|error|	order	$N^h_\psi(u^h)$	|error|	order
0.5	0.451	0.549		1.270	0.270	
0.25	0.551	0.449	0.29	1.129	0.129	1.06
0.125	0.644	0.356	0.33	1.064	0.064	1.02
0.0625	0.725	0.275	0.37	1.032	0.032	1.01
0.03125	0.793	0.207	0.41	1.016	0.016	1.00

***A posteriori* error analysis.** We adopt the following local approximation property. There exists a linear operator $\pi^h : H^1_{-\psi}(\Omega) \to U^h_{-\psi}$ and a positive constant c such that

$$\|v - \pi^h v\|_{L^2(\kappa)} + h_\kappa |v - \pi^h v|_{H^1(\kappa)} \le c h^s_{\hat\kappa} |v|_{H^s(\hat\kappa)}, \quad 1 \le s \le k+1, \quad (7.17)$$

for all $v \in H^s(\hat\kappa)$ and each $\kappa \in \mathcal{T}_h$; here $\hat\kappa$ denotes the union of all elements (including κ itself) in the partition $\mathcal{T}_h$ whose closure has non-empty intersection with the closure of κ, and

$$h_{\hat\kappa} = \max_{\sigma \in \mathcal{T}_h;\, \sigma \subset \hat\kappa} h_\sigma.$$

For the proof of existence of the 'quasi-interpolation' operator π^h we refer to Brenner and Scott (1994), for example.

In addition, we make the following assumption concerning the regularity of the dual problem: there exists a real number $t \ge 1$ such that, for every τ, $1 \le \tau \le t$, there is a positive constant C_τ, independent of ψ, such that the solution z to the dual problem (7.7) satisfies the estimate

$$|z|_{H^\tau(\Omega)} \le C_\tau \|\psi\|_{H^{\tau-1/2}(\Gamma)} \quad (7.18)$$

whenever $\psi \in H^{\tau-1/2}(\Gamma)$. For instance, this bound holds when $\Gamma \in C^{\tau-1,1}$ and the entries of A belong to $C^{[\tau]}(\Omega)$; see Gilbarg and Trudinger (1983).

For each triangle $\kappa \in \mathcal{T}_h$ and u^h denoting the solution to (7.5) in U^h_0, we introduce the *residual term* $\mathcal{R}_\kappa(u^h)$ by

$$\mathcal{R}_\kappa(u^h) = \|\nabla \cdot \boldsymbol{\sigma}(u^h) + f\|_{L^2(\kappa)} + h_{\hat\kappa}^{-1/2} \|[\mathbf{n} \cdot \boldsymbol{\sigma}(u^h)]/2\|_{L^2(\partial\kappa \setminus \Gamma)}, \quad (7.19)$$

where $[w]$ is the jump in w across the faces of elements in the partition and $\mathbf{n}$ is the unit outward normal vector to $\partial\kappa$.

Theorem 7.2. Suppose that (7.10) and (7.18) hold and that the weight function ψ belongs to $H^{t-1/2}(\Gamma)$, $t \ge 1$; then we have that

$$|N_\psi(u) - N^h_\psi(u^h)| \le c \sum_{\kappa \in \mathcal{T}_h} \mathcal{R}_\kappa(u^h) \min_{\{\tau\,:\,1 \le \tau \le \min(t,k+1)\}} h^\tau_{\hat\kappa} \omega_{\kappa,\tau}, \quad (7.20)$$

where c is a constant, and the local weight $\omega_{\kappa,\tau}$ is defined by

$$\omega_{\kappa,\tau} = |z|_{H^\tau(\hat\kappa)},$$

and z is the weak solution of (7.7).

Proof. The starting point of the proof is the third line of the error representation formula (7.9); we integrate by parts triangle-by-triangle using

Green's identity to deduce that

$$\begin{aligned} N_\psi(u) - N_\psi^h(u^h) &= (f, z - \pi^h z) - (\boldsymbol{\sigma}(u^h), \nabla(z - \pi^h z)) \qquad (7.21)\\ &= \sum_{\kappa \in \mathcal{T}_h} \int_\kappa (f + \nabla \cdot \boldsymbol{\sigma}(u^h))(z - \pi^h z)\,\mathrm{d}x \\ &\quad - \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa \setminus \Gamma} ([\mathbf{n} \cdot \boldsymbol{\sigma}(u^h)]/2)(z - \pi^h z)\,\mathrm{d}s \\ &= \mathrm{I} + \mathrm{II}, \end{aligned}$$

where we made use of the fact that z, the weak solution of the dual problem (7.7), belongs to $C^{0,\alpha}(\Omega)$, for some α in $(0,1)$ (see Theorem 5.24 in Gilbarg and Trudinger (1983)), so that the jump $[z - \pi^h z](x) = [z](x) = 0$ at any point x of an internal face $\partial\kappa \setminus \Gamma$ for each element κ in the partition.

Next we estimate expressions I and II. In I, we apply the Cauchy–Schwarz inequality and the approximation property (7.17); hence,

$$|\mathrm{I}| \leq c \sum_{\kappa \in \mathcal{T}_h} \|f + \nabla \cdot \sigma(u^h)\|_{L^2(\kappa)} \min_{\{\tau\,:\,1 \leq \tau \leq \min(t, k+1)\}} h_{\hat{\kappa}}^{\tau} |z|_{H^\tau(\hat{\kappa})}.$$

Now, we consider II. We begin by recalling that the multiplicative trace inequality

$$\|w\|_{L^2(\partial\kappa)}^2 \leq c\|w\|_{L^2(\kappa)} \left(h_\kappa^{-1}\|w\|_{L^2(\kappa)} + |w|_{H^1(\kappa)}\right) \qquad \forall w \in H^1(\kappa),\ \kappa \in \mathcal{T}_h \qquad (7.22)$$

(see Brenner and Scott (1994)), followed by application of approximation property (7.17), yields

$$\|z - \pi^h z\|_{L^2(\partial\kappa)} \leq c h_{\hat{\kappa}}^{\tau - 1/2} |z|_{H^\tau(\hat{\kappa})}.$$

Thus,

$$|\mathrm{II}| \leq c \sum_{\kappa \in \mathcal{T}_h} \|[\mathbf{n} \cdot \boldsymbol{\sigma}(u^h)]\|_{L^2(\partial\kappa \setminus \Gamma)} \min_{\{\tau\,:\,1 \leq \tau \leq \min(t, k+1)\}} h_{\hat{\kappa}}^{\tau - 1/2} |z|_{H^\tau(\hat{\kappa})}.$$

Substituting the bounds on I and II into (7.21) and recalling the definition of the residual term (7.19), we deduce (7.20). □

Since the data for the dual problem (7.7) is generated by a known function, ψ, we may calculate the weight $\omega_{\tau,\alpha}$ by approximating the solution of the dual problem numerically. The right-hand side of the *a posteriori* error estimate can thus be used for quantitative error estimation and local mesh adaptation.

Suppose, for example, that given a positive tolerance `TOL`, the aim of the computation is to find $N_\psi^h(u^h)$ such that

$$|N_\psi(u) - N_\psi^h(u^h)| \leq \mathtt{TOL}. \qquad (7.23)$$

In order to achieve (7.23), by virtue of (7.20) it suffices to ensure that

$$c \sum_{\kappa \in \mathcal{T}_h} \mathcal{R}_\kappa(u^h) \min_{\{\tau\,:\,1 \le \tau \le \min(t,k+1)\}} h_\kappa^\tau \omega_{\kappa,\tau} \le \mathtt{TOL}. \tag{7.24}$$

This inequality can now be used as a *stopping criterion* in an adaptive mesh refinement algorithm. A second ingredient of an adaptive algorithm is a local *refinement criterion*; assuming that N denotes the number of elements in $\mathcal{T}_h$, a possible refinement criterion might involve checking, on each element $\kappa \in \mathcal{T}_h$, whether

$$c\mathcal{R}_\kappa(u^h) \min_{\{\tau\,:\,1 \le \tau \le \min(t,k+1)\}} h_\kappa^\tau \omega_{\kappa,\tau} \le \frac{\mathtt{TOL}}{N}. \tag{7.25}$$

If (7.25) is satisfied on an element κ, then κ is accepted as being of adequate size; if, on the other hand, (7.25) is violated then κ is refined. After (7.25) has been checked on each element κ in $\mathcal{T}^h$ and a new, finer, subdivision $\mathcal{T}_{h'}$ has been generated, a new solution $u^{h'}$ is computed on $\mathcal{T}_{h'}$, thus completing a single step of the adaptive algorithm. Adaptation proceeds until the stopping criterion (7.24) is satisfied. The adaptive algorithm then terminates and delivers $N_\psi^h(u^h)$, accurate to within the specified tolerance TOL, as required by (7.23).

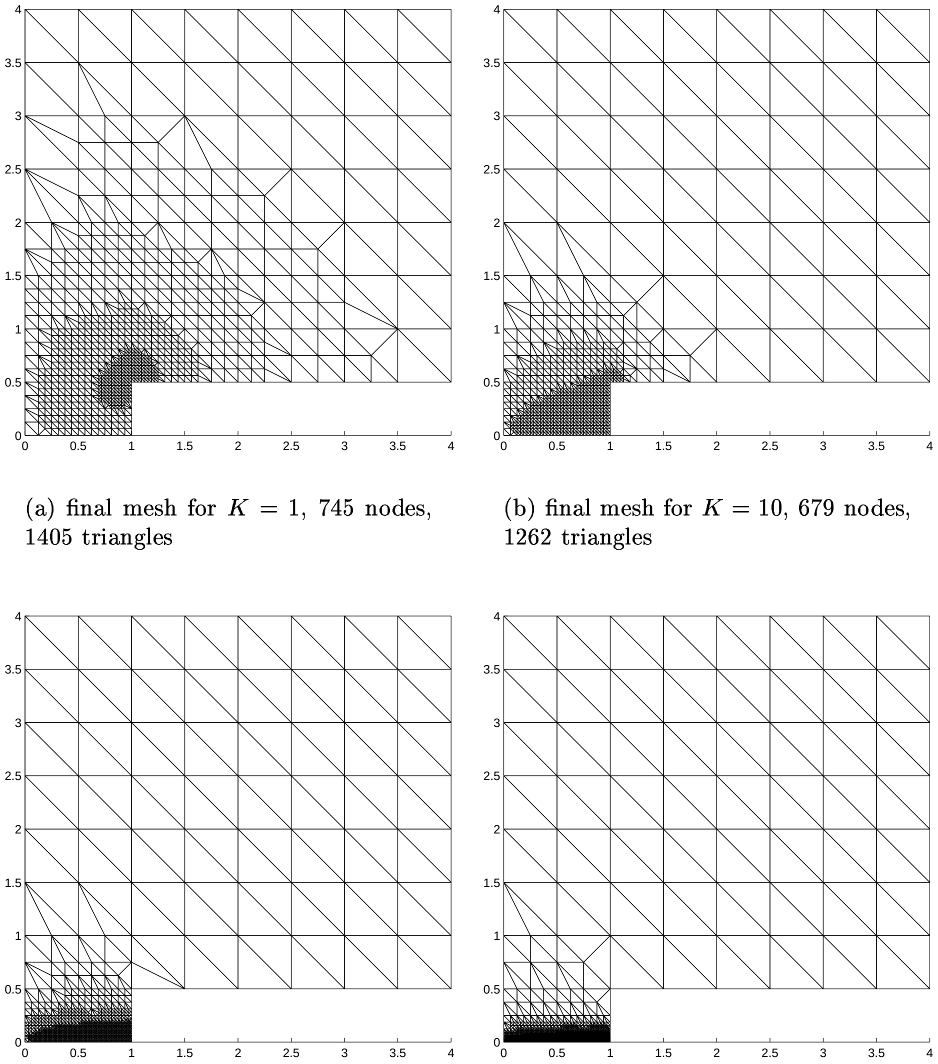
For extensions of the theory discussed in this section to superconvergent lift and drag computations for the Stokes and Navier–Stokes equations, we refer to Giles *et al.* (1997). The technique of postprocessing presented here is based on early ideas of Wheeler (1973), Babuška and Miller (1984*a*, 1984*b*, 1984*c*); see also Barrett and Elliott (1987) for a rigorous error analysis in the presence of variational crimes.

A numerical experiment. The purpose of this experiment is to illustrate the performance of the *a posteriori* error bound (7.20) and to compare it with some heuristic mesh refinement criteria. Let us consider a reaction–diffusion equation in cylindrical polar coordinates,

$$-\nabla^2 u + Ku = 0,$$

in an L-shaped domain with boundary conditions

$$\begin{aligned} u = 1, &\quad r \le 1,\ z = 0, \\ u \to 0, &\quad r,\ z \to \infty, \\ \frac{\partial u}{\partial \nu} = 0, &\quad r = 0,\ z > 0, \\ &\quad r = 1,\ 0 \le z \le 0.5, \\ &\quad r > 1,\ z = 0.5, \end{aligned}$$



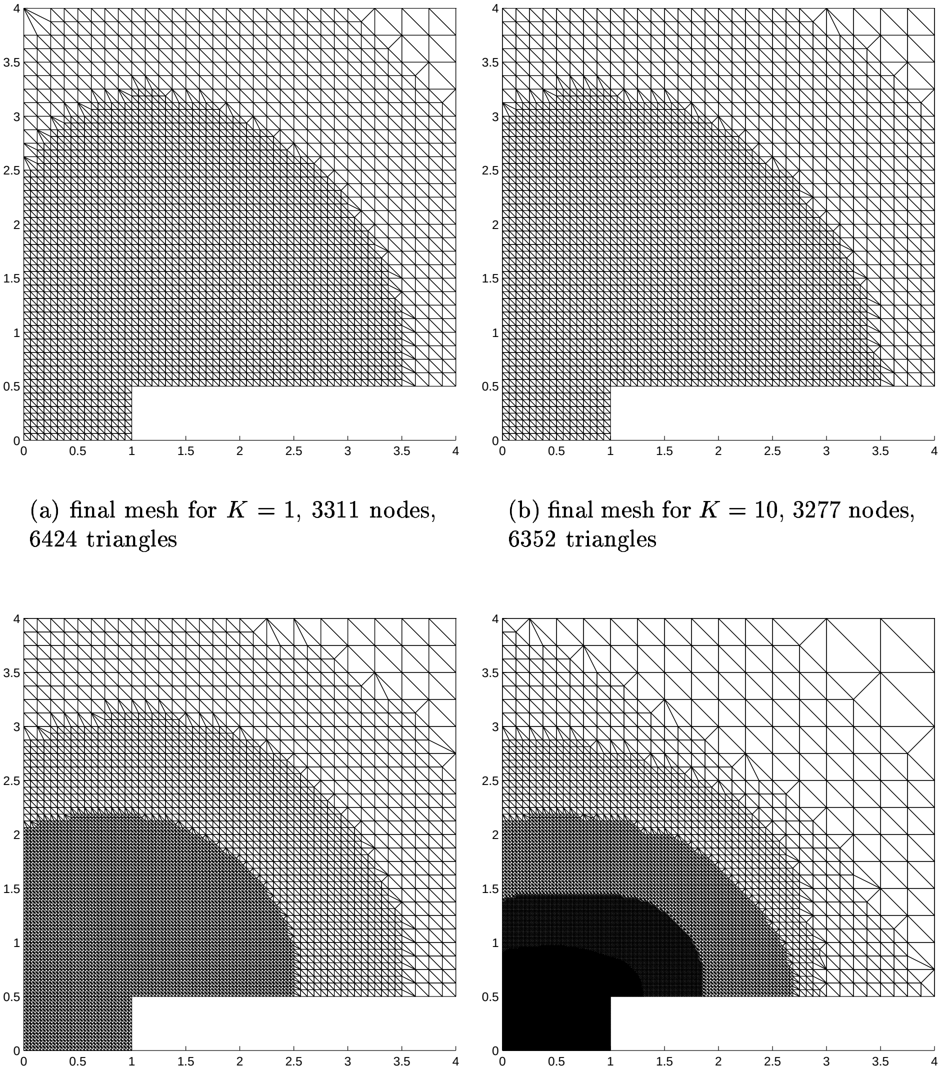
(a) final mesh for $K = 1$, 745 nodes, 1405 triangles

(b) final mesh for $K = 10$, 679 nodes, 1262 triangles

(c) final mesh for $K = 100$, 972 nodes, 1822 triangles

(d) final mesh for $K = 1000$, 3904 nodes, 7494 triangles

Figure 7.1. The final meshes for calculating the linear functional using a refinement indicator based on (7.24)



(a) final mesh for $K = 1$, 3311 nodes, 6424 triangles

(b) final mesh for $K = 10$, 3277 nodes, 6352 triangles

(c) final mesh for $K = 100$, 6463 nodes, 12636 triangles

(d) final mesh for $K = 1000$, 24753 nodes, 48941 triangles

Figure 7.2. The final meshes produced using the empirical error indicator $\|\nabla u^h\|_{L^2(\kappa)}$

and suppose we wish to evaluate the linear functional

$$N_\psi(u) = \int_\Gamma \frac{\partial u}{\partial \nu} \psi \, r \, \mathrm{d}s,$$

where

$$\psi = \begin{cases} \frac{\pi}{2} & 0 \le r \le 1, \quad z = 0, \\ 0 & \text{elsewhere on } \Gamma. \end{cases}$$

We consider solving this problem using an adaptive finite element algorithm. Firstly we use an inequality of the form of (7.24) above as a stopping criterion and to guide mesh refinement. The final meshes (chosen to ensure accuracy of the numerical solution to within 1% of the exact linear functional) are shown in Figure 7.1. Secondly, we consider an empirical refinement indicator, namely $\|\nabla u^h\|_{L^2(\kappa)}$. Instead of attempting to equidistribute the error bound we now use the *fixed fraction method* for mesh refinement. In this method we refine and re-coarsen a fixed proportion of the elements at each refinement level. The elements to be refined are those with the largest refinement indicators while those to be re-coarsened have the smallest refinement indicators. In our experiments we chose to refine one-third of the elements and re-coarsen one-tenth. The algorithm was terminated when the error in the computed linear functional was less than the adaptive tolerance TOL chosen for the first approach. The results of this are shown in Figure 7.2. Comparison of this with Figure 7.1 shows that this second approach, based on the empirical refinement indicator, requires at least four times as many nodes as the first to achieve the same accuracy in the computed linear functional. Also, the meshes produced are far less sensitive to the governing parameter K.

We conclude with some remarks. In order to be able to implement the bound (7.20) both the constant c and the dual solution z have to be precomputed. The computation of c in turn involves knowledge of the constants from (7.17) and (7.22); constructive proofs of (7.17) and (7.22) which supply actual values of these constants are available: see, for example, Carstensen (2000). However, it is clear that, while the *a posteriori* error bound (7.20) is reliable (*i.e.*, it consistently overestimates the actual error $|N_\psi(u) - N_\psi^h(u^h)|$), the factor of overestimation may be excessive. Thus, in the next section, we shall develop a minimalistic framework of *Type I a posteriori* error estimation which does not require knowledge of these constants. The guiding principle in the derivation of a Type I error bound is to perform only the absolute minimum in the way of upper bounds on the error representation formula so as to ensure sharpness of the resulting estimate.

7.2. Abstract Type I a posteriori *error estimation*

Let us return to the abstract framework of Section 3.1, and recall that the error between $J_p = \mathcal{J}_p(u)$ and its Galerkin approximation J_p^h is expressed by the error representation formula (3.10), which states that

$$J_p - J_p^h = \langle R_p(u^h), z - z^h \rangle \qquad \forall z^h \in V_d^h, \tag{7.26}$$

where $\langle \cdot\,, \cdot \rangle$ is the duality pairing between V', the dual of the real Hilbert space V, and V; $R_p(u^h) : V \to \mathbb{R}$ denotes the linear functional, called the *residual*, defined by

$$R_p(u^h) \,:\, w \mapsto \ell(w) - B(u^h, w).$$

Now, writing

$$\mathcal{E}_\Omega(u^h; w) \equiv \langle R_p(u^h), w \rangle, \qquad w \in V,$$

and recalling (3.11), we have that

$$\begin{aligned} J_p - J_p^h &= \langle R_p(u^h), z^H - z^h \rangle + \langle R_p(u^h), z - z^H \rangle \\ &\equiv \mathcal{E}_\Omega(u^h; z^H - z^h) + \mathcal{E}_\Omega(u^h; z - z^H) \quad \forall z^h \in V_d^h, \end{aligned} \tag{7.27}$$

where $z^H \in V_d^H$ is the numerical solution of the dual problem (2.13), defined by

$$B(w_0^H, z^H) = m(w_0^H) \qquad \forall w_0^H \in U_0^H.$$

Let us suppose that the aim of the computation is to ensure that, for a given tolerance $\mathtt{TOL} > 0$,

$$|J_p - J_p^h| \leq \mathtt{TOL}. \tag{7.28}$$

Clearly, (7.28) is equivalent to requiring that

$$|\mathcal{E}_\Omega(u^h; z - z^h)| \leq \mathtt{TOL} \qquad \forall z^h \in V_d^h. \tag{7.29}$$

Hence, a sufficient condition for (7.29) is that

$$|\mathcal{E}_\Omega(u^h; z^H - z^h)| + |\mathcal{E}_\Omega(u^h; z - z^H)| \leq \mathtt{TOL} \qquad \forall z^h \in V_d^h. \tag{7.30}$$

In order to ensure that (7.30) holds, we select $\theta \in (0, 1)$, and demand that

$$\mathcal{E}_{\mathrm{P}} \equiv |\mathcal{E}_\Omega(u^h; z^H - z^h)| \leq (1 - \theta)\,\mathtt{TOL} \qquad \forall z^h \in V_d^h, \tag{7.31}$$

$$\mathcal{E}_{\mathrm{D}} \equiv |\mathcal{E}_\Omega(u^h; z - z^H)| \leq \theta\,\mathtt{TOL}. \tag{7.32}$$

Let us discuss each of these two inequalities in detail.

(7.31) The residual $R_p(u^h)$ appearing in $\mathcal{E}_\Omega(u^h; z^H - z^h)$ is computable, since it involves only the numerical solution $u^h \in U_p^h$ and the data (*i.e.*, the functional ℓ, in this case). Further, $z^H \in V_d^H$ is the numerical solution of the dual problem (2.13), defined by

$$B(w_0^H, z^H) = m(w_0^H) \qquad \forall w_0^H \in U_0^H. \tag{7.33}$$

As z^h in (7.31) is an arbitrary element of V_d^h, we need to fix it. It is worth noting at this point that, due to the Galerkin orthogonality property (3.14),

$$\langle R_p(u^h), z^H - z^h\rangle = \langle R_p(u^h), z_0^H - z_0^h\rangle = \langle R_p(u^h), z_0^H\rangle,$$

and therefore the choice of z^h does not influence the value of $\mathcal{E}_\Omega(u^h; z^H - z^h)$. Still, this does not mean that any z^h (such as $z^h = 0$, for example) will be a useful choice. Of course, if $\mathcal{E}_\Omega(u^h; z^H - z^h) = \langle R_p(u^h), z_0^H\rangle$ were all we cared about, the choice of z^h would be immaterial; however, it has to be borne in mind that, in addition to a *stopping criterion* such as (7.31), our adaptive algorithm will also require a *local refinement criterion*. A local refinement criterion can be obtained by localization of the term $\langle R_p(u^h), z^H - z^h\rangle$. By this, we mean the following. Let us suppose that u^h has been computed using a Galerkin finite element method over a subdivision $\mathcal{T}_h$ of the computational domain Ω into finite elements κ. We assume the existence of the following decomposition:

$$\mathcal{E}_\Omega(u^h; w) \equiv \langle R_p(u^h), w\rangle = \sum_{\kappa\in\mathcal{T}_h} \eta_\kappa(u^h|_\kappa, w|_\kappa).$$

Then,

$$\langle R_p(u^h), z^H - z^h\rangle = \sum_{\kappa\in\mathcal{T}_h} \eta_\kappa(u^h|_\kappa, (z^H - z^h)|_\kappa),$$

and therefore,

$$\begin{aligned}|\mathcal{E}_\Omega(u^h; z^H - z^h)| &\le \sum_{\kappa\in\mathcal{T}_h} |\tilde{\eta}_\kappa| \\ &\equiv \mathcal{E}_{|\Omega|}(u^h; z^H - z^h),\end{aligned} \tag{7.34}$$

where

$$\tilde{\eta}_\kappa = \eta_\kappa(u^h|_\kappa, (z^H - z^h)|_\kappa).$$

Now, $\mathcal{E}_{|\Omega|}(u^h; z^H - z^h)$ is referred to as the *localization* of the expression $|\mathcal{E}_\Omega(u^h; z^H - z^h)|$. While the left-hand side of (7.34) is completely independent of $z^h \in V_d^h$, the right-hand side of this inequality is strongly dependent on z^h because Galerkin orthogonality (*cf.* (3.14)) is a nonlocal property. Ideally, in order to minimize the degree of overestimation in (7.34) which may result from an unfortunate choice of z^h, we would like to choose $z^h \in V_d^h$ so that the right-hand side of (7.34) is as close as possible to the left-hand side. This, of course, is a practically unrealistic demand as it would lead to a complicated optimization problem.

A more reasonable choice from the practical point of view is $z^h = \pi^h z^H$, where $\pi^h : V_d \to V_d^h$ is a finite element interpolation or quasi-interpolation operator; this particular choice of z^h is motivated by the expectation that $\mathcal{E}_{|\Omega|}(u^h; z^H - \pi^h z^H)$ exhibits the same asymptotic behaviour as the expression $\mathcal{E}_\Omega(u^h; z^H - \pi^h z^H) = \mathcal{E}_\Omega(u^h; z_0^H)$ in the limit of $h \to 0$. While this expectation is certainly fulfilled in most situations, it is by no means so in general because, in the presence of global superconvergence effects, $\mathcal{E}_\Omega(u^h; z^H - \pi^h z^H)$ may exhibit a higher rate of convergence than $\mathcal{E}_{|\Omega|}(u^h; z^H - \pi^h z^H)$, as $h \to 0$. However, such global cancellation effects and the related mismatch in the asymptotic behaviour will be largely absent on locally refined unstructured computational meshes, such as those that arise in the course of adaptive mesh refinement.

At any rate, once a choice of z^h has been made, the expression $\mathcal{E}_\mathrm{P}$ is *computable*, and condition (7.31) can be checked. Indeed, a sufficient condition for (7.31), based on localization, is that

$$\mathcal{E}_\mathrm{P}^\mathrm{loc} \equiv \mathcal{E}_{|\Omega|}(u^h; z^H - \pi^h z^H) \leq (1-\theta)\mathtt{TOL}. \tag{7.35}$$

A bound of this kind, which explicitly involves the numerical approximation z^H to the dual solution z, will be referred to as a *Type I a posteriori error bound*.

Next we shall discuss the computation of the dual solution z^H involved in (7.35); we shall see that the correct choice of z^H is closely related to the validity of (7.32).

(7.32) Unlike $\mathcal{E}_\mathrm{P}^\mathrm{loc}$, the term $\mathcal{E}_\mathrm{D}$ involves the (unknown) analytical dual solution z. We note, however, that (7.32) can be restated as a *dual measurement problem* concerned with finding a solution $z^H \in V_d^H$ to (7.33) such that the error in the output functional $\mathcal{E}_\Omega(u^h; \cdot)$ satisfies

$$|\mathcal{E}_\Omega(u^h; z) - \mathcal{E}_\Omega(u^h; z^H)| \leq \theta\, \mathtt{TOL}. \tag{7.36}$$

A further important difference between the terms $\mathcal{E}_\mathrm{P}^\mathrm{loc}$ and $\mathcal{E}_\mathrm{D}$ is that (due to the localization) in $\mathcal{E}_\mathrm{P}^\mathrm{loc}$ the absolute value signs appear under the summation over the elements $\kappa \in \mathcal{T}_h$, while in $\mathcal{E}_\mathrm{D}$ the absolute value sign is outside the sum. Thus, we expect $\mathcal{E}_\mathrm{D}$ to be smaller than $\mathcal{E}_\mathrm{P}^\mathrm{loc}$.

Motivated by these observations, we select $\theta \in (0,1)$ such that $0 < \theta \ll 1$; we then aim to compute $u^h \in U_p^h$ such that

$$\mathcal{E}_\mathrm{P}^\mathrm{loc} \equiv \mathcal{E}_{|\Omega|}(u^h; z^H - \pi^h z^H) \approx (1-\theta)\, \mathtt{TOL}$$

and $z^H \in V_d^H$ such that

$$\mathcal{E}_\mathrm{D} \equiv |\mathcal{E}_\Omega(u^h; z) - \mathcal{E}_\Omega(u^h; z^H)| \leq \theta\, \mathtt{TOL}. \tag{7.37}$$

Together, these will imply that $\mathcal{E}_\mathrm{D} \ll \mathcal{E}_\mathrm{P}^\mathrm{loc}$.

The dual measurement problem (7.37) is very similar to the problem (7.28) that we had set out to solve, except that (7.37) concerns the dual solution z while (7.28) involves the primal solution u. To ensure the validity of (7.37) we could derive an *a posteriori* error bound on $|\mathcal{E}_\Omega(u^h; z) - \mathcal{E}_\Omega(u^h; z^H)|$; the corresponding error representation formula would involve the residual associated with the numerical solution of the dual problem (2.13) and the analytical solution of the dual to the dual problem. As the use of a Type I *a posteriori* error bound on $\mathcal{E}_{\mathrm{D}}$ based on such an error representation formula would necessitate the numerical solution of the dual to the dual problem (which is not a particularly appealing prospect), one can instead use a cruder Type II *a posteriori* error bound on $\mathcal{E}_{\mathrm{D}}$ which, in the spirit of Eriksson *et al.* (1995, 1996), eliminates the dual solution from the *a posteriori* error estimate through bounding its norms above by a *stability constant*. This then terminates the potentially infinite sequence of mutually dual problems which would otherwise arise. The crudeness of the Type II bound on $\mathcal{E}_{\mathrm{D}}$ is of no particular concern here: from the practical point of view there appears to be little advantage in performing reliable error control on $\mathcal{E}_{\mathrm{D}}$; our aim, when using the Type II bound on $\mathcal{E}_{\mathrm{D}}$, is merely to generate an adequate sequence of finite element approximations z^H to the dual solution z which we can then use to compute $\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}}$.

Indeed, the numerical experiments in Houston and Süli (2001*a*) indicate (*cf.* also Hartmann (2001) and the remarks in Section 3.2 about the selection of the space V_d^H) that, with a reasonable choice of the dual finite element space V_d^H, $\mathcal{E}_{\mathrm{D}}$ is typically an order of magnitude smaller than $\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}}$. Therefore, as an alternative to the costly exercise of computing z^H through rigorous error control for the dual measurement problem (7.37), we may simply absorb $\mathcal{E}_{\mathrm{D}}$ into $\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}}$, and replace (7.35) by

$$\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}} \leq \mathtt{TOL}, \tag{7.38}$$

without compromising the reliability of the adaptive algorithm.

Of course, for the Type I error bound (7.35) to be an accurate approximation of (7.30) 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 V_d^H is sufficiently different from the primal finite element space V_d^h; for example, if V_d^H is chosen to coincide with V_d^h then $z^H = z^h$ and thereby $\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}} = 0$, so $0 < \mathcal{E}_{\mathrm{D}} \ll \mathcal{E}_{\mathrm{P}}^{\mathrm{loc}}$ cannot hold. We refer to the comments in Section 3.2 for further details on this issue.

In order to construct a working adaptive algorithm, in addition to an *a posteriori* error bound we also need a mesh refinement criterion and a mesh modification strategy. Some of the possible approaches are reviewed in the next section.

7.3. Mesh refinement criteria and mesh modification strategies

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:

$$|\tilde{\eta}_\kappa| \leq \frac{\mathtt{TOL}}{N}, \tag{7.39}$$

where

$$\tilde{\eta}_\kappa = \eta_\kappa(u^h|_\kappa, (z^H - z^h)|_\kappa),$$

as in (7.34), where N is the number of elements in $\mathcal{T}_h$. If inequality (7.39) 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 λ, $0 < \lambda \ll 1$, and marking elements κ with

$$|\tilde{\eta}_\kappa| \le \lambda \frac{\mathtt{TOL}}{N}$$

for derefinement. It is assumed that the hierarchy of meshes is generated from a coarse *background mesh*, supplied by the user, beyond which no derefinement can occur.

Fixed fraction criterion. Of course, other refinement criteria are also possible. For example, the *fixed fraction strategy* involves choosing two numbers φ_{ref} and φ_{deref} in the interval $(0, 100)$ with $\varphi_{\text{deref}} + \varphi_{\text{ref}} < 100$, ordering the *local refinement indicators* $|\tilde{\eta}_\kappa|$, $\kappa \in \mathcal{T}_h$, according to their size, and then refining those elements κ which correspond to $\varphi_{\text{ref}}\%$ of the largest entries in the ordered sequence (the top 20%, say), and derefining those elements κ which correspond to the $\varphi_{\text{deref}}\%$ 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* (see, *e.g.*, Giles (1998), Rannacher (1998) and Becker and Rannacher (2001)) aims to design a subdivision $\mathcal{T}_h$ of the computational domain $\Omega \subset \mathbb{R}^n$ (or, equivalently, a mesh function $h(x)$ defined on Ω) for the primal problem so that the number N of elements in the subdivision $\mathcal{T}_h$ is minimized, subject to the constraint that

$$\mathcal{E}_{|\Omega|}(u^h; z - \pi^h z) \approx \mathtt{TOL}.$$

Assuming that the computational domain $\Omega \subset \mathbb{R}^n$ has been subdivided into elements $\kappa \in \mathcal{T}_h$, we can write

$$N = \sum_{\kappa \in \mathcal{T}_h} 1 = \sum_\kappa \int_\kappa \frac{1}{\operatorname{meas}(\kappa)}\, \mathrm{d}x \approx \sum_\kappa \int_\kappa \frac{\mathrm{d}x}{h^n(x)}.$$

On the other hand,

$$\mathcal{E}_{|\Omega|}(u^h; z - \pi^h z) = \sum_{\kappa \in \mathcal{T}_h} |\eta_\kappa|,$$

where

$$\eta_\kappa = \eta_\kappa(u^h; z - \pi^h z).$$

Let us suppose that $|\eta_\kappa|$ can be expressed as

$$|\eta_\kappa| = \int_\kappa A(x) h^k(x)\,\mathrm{d}x \tag{7.40}$$

for some positive real number k, where $A(x) = \mathcal{O}(1)$ as $h \to 0$. Then,

$$\mathcal{E}_{|\Omega|}(u^h; z - \pi^h z) = \int_\Omega A(x) h^k(x)\,\mathrm{d}x.$$

Thus, to find an 'optimal' h, we need to solve the following constrained optimization problem:

$$\int_\Omega \frac{\mathrm{d}x}{h^n(x)} \to \min \quad \text{subject to} \quad \int_\Omega A(x) h^k(x)\,\mathrm{d}x - \mathtt{TOL} = 0.$$

Let us consider the Lagrangian

$$\mathcal{L}(\lambda, h) = \int_\Omega \frac{\mathrm{d}x}{h^n(x)} + \lambda\left(\int_\Omega A(x) h^k(x)\,\mathrm{d}x - \mathtt{TOL}\right),$$

where $\lambda \in \mathbb{R}$ is a Lagrange multiplier. An elementary calculation shows that the Gateaux derivative of $\mathcal{L}$ in the 'direction' $\hat{h}$ is

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial h}(\lambda, h; \hat{h}) &= \lim_{\epsilon \to O} \frac{\mathcal{L}(\lambda, h + \epsilon \hat{h}) - \mathcal{L}(\lambda, h)}{\epsilon} \\ &= \int_\Omega \left\{k\lambda A(x) h^{k-1}(x) - n h^{-n-1}(x)\right\} \hat{h}(x)\,\mathrm{d}x.\end{aligned}$$

Now, from the requirement that, at a stationary point $(\lambda^{\mathrm{opt}}, h^{\mathrm{opt}})$,

$$\frac{\partial \mathcal{L}}{\partial h}(\lambda^{\mathrm{opt}}, h^{\mathrm{opt}}; \hat{h}) = 0$$

for all $\hat{h}$, we deduce that

$$h^{\mathrm{opt}}(x) = \left(\frac{n}{k\lambda A(x)}\right)^{\frac{1}{k+n}}. \tag{7.41}$$

Substituting this into the constraint

$$\int_\Omega A(x) h^k(x)\,\mathrm{d}x = \mathtt{TOL},$$

we deduce that

$$\left(\frac{n}{k\lambda}\right)^{\frac{k}{k+n}} W = \mathtt{TOL}, \tag{7.42}$$

where

$$W = \int_\Omega A^{\frac{n}{k+n}}(x)\,\mathrm{d}x. \tag{7.43}$$

Eliminating λ from (7.41) using (7.42), we obtain

$$h^{\mathrm{opt}}(x) = \left(\frac{\mathtt{TOL}}{W}\right)^{\frac{1}{k}} A^{-\frac{1}{k+n}}(x), \qquad x \in \kappa, \quad \kappa \in \mathcal{T}_h,$$

where W is defined by (7.43) and A is defined (elementwise) by (7.40); of course, in practice the dual solution z involved in A is replaced by its finite element approximation z^H (*i.e.*, $\tilde{\eta}_\kappa$ is used instead of η_κ). An application of the optimized mesh criterion will be given in the next section.

Any of these criteria can be coupled with a suitable mesh modification algorithm. For example, in two space dimensions a red–green refinement strategy may be used. 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 only 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 U^h is contained in $C(\bar{\Omega})$. In certain nonconforming methods, such as the discontinuous Galerkin finite element method (*cf.* Cockburn, Karniadakis and Shu (2000) and Section 9), it is not assumed that U^h is contained in $C(\bar{\Omega})$, so the existence of hanging nodes in the mesh is perfectly acceptable.

A numerical experiment. The purpose of this numerical experiment is to illustrate the sharpness of a Type I error bound. We consider the reaction–diffusion equation

$$-\nabla^2 u + u = f(x,y) \qquad \text{in } \Omega = (0,1) \times (0,1)$$

with boundary conditions

$$\begin{aligned} u &= 0, & y &= 0, \\ \frac{\partial u}{\partial \nu} &= 0, & x &= 0, \\ & & x &= 1, \\ u &= x^2(1-x)^2, & y &= 1. \end{aligned}$$

Table 7.2. Reliability of a Type I *a posteriori* error bound

h	$N_{\psi}^{h}(u^h)$	|error|	error bound	effectivity index
1/4	-2.082×10^{-2}	5.417×10^{-3}	9.258×10^{-3}	1.709
1/8	-1.693×10^{-2}	1.528×10^{-3}	2.903×10^{-3}	1.900
1/16	-1.579×10^{-2}	3.958×10^{-4}	7.897×10^{-4}	1.995
1/32	-1.550×10^{-2}	9.984×10^{-5}	2.001×10^{-4}	2.005
1/64	-1.542×10^{-2}	2.502×10^{-5}	5.019×10^{-5}	2.006

Here $f(x,y)$ is chosen so that the exact solution to the problem is $u(x,y) = yx^2(1-x)^2$. We consider the numerical approximation of the linear functional

$$N_{\psi}(u) = \int_{\Gamma} \psi \frac{\partial u}{\partial \nu} \mathrm{d}s$$

where

$$\psi = \begin{cases} -\cos(2\pi x) & y = 0, \\ 0 & \text{elsewhere on } \Gamma, \end{cases}$$

which has exact value $-3/(2\pi^4)$. As described above we may derive a Type I *a posteriori* error bound:

$$|N_{\psi}(u) - N_{\psi}^{h}(u^h)| \leq \sum_{\kappa} \left| \int_{\kappa} (u^h - f)(v^h - z)\mathrm{d}x + \frac{1}{2}\int_{\partial\kappa} \left[\frac{\partial u^h}{\partial \nu}\right] |v^h - z| \, \mathrm{d}s \right| .$$

Table 7.2 demonstrates that this really does provide an upper bound on the error in the computed linear functional. Here we have taken a sequence of regular meshes and computed the numerical approximation $N_{\psi}^{h}(u^h)$ to the linear functional, the actual error, the error bound and the effectivity index, which is the ratio of the error bound to the actual error and thus measures the extent to which the error bound overestimates the error. In the computations the dual solution z appearing in the inequality above has been replaced by an approximation $\tilde{z}$ computed on the same mesh as the primal approximation u^h, but with a piecewise quadratic finite element space.

8. Mesh-dependent perturbations and duality

In many instances, the bilinear functional $B(\cdot\,,\cdot)$ and the linear functional $\ell(\cdot)$ that appear in the statement of (2.1) have to be replaced by numerical approximations $B_h(\cdot\,,\cdot)$ and $\ell_h(\cdot)$, respectively. For example, numerical quadrature or numerical approximation of a curved computational domain Ω by a polyhedral domain Ω_h may lead to such perturbations of $B(\cdot\,,\cdot)$ and $\ell(\cdot)$. We note in this respect that many finite volume methods can be restated as Petrov–Galerkin finite element methods of the form (3.1) with numerical quadrature.

8.1. Error correction and primal–dual equivalence

Let us suppose again that $\{U_0^h\}_{h>0}$ and $\{V_0^h\}_{h>0}$ are two families of finite-dimensional subspaces of U_0 and V_0, respectively, parametrized by $h \in (0,1]$. When U_0 is a proper Hilbert subspace of U, we assign to $p \in U$ the affine variety $U_p^h = p + U_0^h \subset U_p \subset U$; similarly, when V_0 is a proper Hilbert subspace of V, we assign to $d \in V$ the affine variety $V_d^h = d + V_0^h \subset V_d$.

We consider the following *discrete primal problem.*

($\hat{\mathrm{P}}^h$) Suppose that $m : U \to \mathbb{R}$ and $\ell_h : V_d^h \to \mathbb{R}$ are linear functionals and $B_h(\cdot\,,\cdot) : U_p^h \times V_d^h \to \mathbb{R}$ is a bilinear functional. Find $J_p^h \in \mathbb{R}$ and $u^h \in U_p^h$ such that

$$J_p^h = m(u^h) + \ell_h(v^h) - B_h(u^h, v^h) \qquad \forall v^h \in V_d^h. \tag{8.1}$$

It is also possible to include the case when $m(\cdot)$ has been approximated by a linear functional $m_h(\cdot)$, but for the sake of brevity we shall not discuss this here since this extension can be handled similarly.

In analogy with ($\hat{\mathrm{P}}^h$), we define the *discrete dual problem* as follows. Suppose that $\{U_0^H\}_{H>0}$ and $\{V_0^H\}_{H>0}$ are two families of finite-dimensional subspaces of U_0 and V_0, respectively, parametrized by $H \in (0,1]$, typically different from the families $\{U_0^h\}_{h>0}$ and $\{V_0^h\}_{h>0}$. We assign to $p \in U$ the affine variety $U_p^H = p + U_0^H \subset U_p \subset U$; similarly, we assign to $d \in V$ the affine variety $V_d^H = d + V_0^H \subset V_d \subset V$.

($\hat{\mathrm{D}}^H$) Suppose that $m_H : U_p^H \to \mathbb{R}$ and $\ell : V_d \to \mathbb{R}$ are linear functionals, and $B_H(\cdot\,,\cdot) : U_p^H \times V_d^H \to \mathbb{R}$ is a bilinear functional. Find $J_d^H \in \mathbb{R}$ and $z^H \in V_d^H$ such that

$$J_d^H = m_H(w^H) + \ell(z^H) - B_H(w^H, z^H) \qquad \forall w^H \in U_p^H.$$

Again, one may also include the case when $\ell(\cdot)$ has been approximated by a linear functional $\ell_H(\cdot)$; for the sake of brevity, we shall refrain from discussing this.

Next we present representation formulae for the error between J_p, J_d and their respective approximations J_p^h, J_d^H. In particular, we shall see that, when J_p and J_d are appropriately corrected by terms which stem from perturbing the bilinear functional $B(\cdot\,,\cdot)$ and the linear functionals $m(\cdot)$ and $\ell(\cdot)$, we recover error representation formulae analogous to (3.4) and (3.5).

Theorem 8.1. (Error representation formula) Let $(J_p, u) \in \mathbb{R} \times U_p$ and $(J_d, z) \in \mathbb{R} \times V_d$ denote the solutions to (P) and (D), respectively, and let $(J_p^h, u^h) \in \mathbb{R} \times U_p^h$ and $(J_d^H, z^H) \in \mathbb{R} \times V_d^H$ be the solutions to ($\hat{\mathrm{P}}^h$)

and $(\hat{\mathrm{D}}^H)$, respectively. Let us define

$$\hat{J}_p^h = J_p^h + \big[(\ell - \ell_h)(z^h) - (B - B_h)(u^h, z^h)\big], \tag{8.2}$$

$$\hat{J}_d^H = J_d^H + \big[(m - m_H)(u^H) - (B - B_H)(u^H, z^H)\big]. \tag{8.3}$$

Then,

$$J_p - \hat{J}_p^h = B(u - u^h, z - z^h) \qquad \forall z^h \in V_d^h, \tag{8.4}$$

$$J_d - \hat{J}_d^H = B(u - u^H, z - z^H) \qquad \forall u^H \in U_p^H. \tag{8.5}$$

Proof. Since $V_d^h \subset V_d$, we have from (P) that

$$J_p = m(u) + \ell(v^h) - B(u, v^h) \qquad \forall v^h \in V_d^h.$$

Recalling from $(\hat{\mathrm{P}}^h)$ that

$$J_p^h = m(u^h) + \ell_h(v^h) - B_h(u^h, v^h) \qquad \forall v^h \in V_d^h$$

and subtracting, we find that, for any $v^h \in V_d^h$,

$$\begin{aligned} J_p - J_p^h &= \ell(v^h) - \ell_h(v^h) + m(u - u^h) - \big[B(u, v^h) - B_h(u^h, v^h)\big] \\ &= \ell(v^h) - \ell_h(v^h) + m(u - u^h) \\ &\qquad - B(u - u^h, v^h) - \big[B(u^h, v^h) - B_h(u^h, v^h)\big]. \end{aligned} \tag{8.6}$$

On the other hand, as $u - u^h \in U_0$, we deduce from (2.13) that

$$B(u - u^h, z) = m(u - u^h),$$

which we can use to eliminate $m(u - u^h)$ from (8.6) and deduce that

$$\begin{aligned} J_p - &\big\{J_p^h + \big[(\ell(v^h) - B(u^h, v^h)) - (\ell_h(v^h) - B_h(u^h, v^h))\big]\big\} \\ &= B(u - u^h, z - v^h), \end{aligned}$$

for all v^h from V_d^h; hence (8.4). The proof of the identity (3.5) is completely analogous. □

Our next result is a counterpart of the discrete Primal–Dual Equivalence Theorem, Theorem 3.3.

Theorem 8.2. Suppose that $(J_p^h, u^h) \in \mathbb{R} \times U_p^h$ and $(J_d^H, z^H) \in \mathbb{R} \times V_d^H$ denote the solutions to the primal problem $(\hat{\mathrm{P}}^h)$ and the dual problem $(\hat{\mathrm{D}}^H)$, respectively, and define $\hat{J}_p^h$ and $\hat{J}_d^H$ as in (8.2) and (8.3) above; then,

$$\hat{J}_p^h = \hat{J}_d^H + \rho^{hH},$$

where

$$\rho^{hH} = B(u - u^H, z - z^H) - B(u - u^h, z - z^h),$$

for any $u^H \in U_p^H$ and any $z^h \in V_d^h$.

Proof. The result is a direct consequence of the previous theorem, on subtracting (8.4) from (8.5), and recalling from the Primal–Dual Equivalence Theorem that $J_p = J_d$. □

Next, we shall consider an application of these abstract results to a class of stabilized finite element methods that includes the streamline diffusion finite element method (SDFEM), and the least-squares stabilized finite element method for a scalar linear hyperbolic problem. Such stabilized methods arise by perturbing the classical Galerkin finite element method in a consistent manner through the inclusion of a least-squares stabilization term, so as to enhance numerical dissipation in the direction of the characteristic curves of the hyperbolic operator.

8.2. Hyperbolic model problem: the effects of stabilization

Let us consider the transport problem

$$\mathcal{L}u \equiv \mathbf{b}\cdot\nabla u + cu = f, \quad x \in \Omega, \qquad u = g, \quad x \in \Gamma_-, \tag{8.7}$$

where $\Omega = (0,1)^n$, Γ is the union of open faces of Ω, and Γ_- denotes the inflow part of Γ, namely the set of all points $x \in \Gamma$ where the vector $\mathbf{b}(x)$ points into Ω; Γ_+, the outflow part of Γ, is defined analogously.

As before, we assume that the entries $b_1, \ldots, b_n$ of the n-component vector function $\mathbf{b}$ are continuously differentiable and positive on $\bar{\Omega}$; this hypothesis ensures that Γ is noncharacteristic for the operator $\mathcal{L}$ at each point $x \in \Gamma$. Also, we shall suppose that $c \in C(\bar{\Omega})$, $f \in L^2(\Omega)$ and $g \in L^2(\Gamma_-)$. In addition, it will be assumed that there exists $\gamma > 0$ such that

$$c_0^2(x) \equiv c(x) - \frac{1}{2}\nabla\cdot\mathbf{b}(x) \geq \gamma^2 \qquad \text{for all } x \in \bar{\Omega}. \tag{8.8}$$

In order to introduce the variational formulation of the boundary value problem (8.7), we associate with $\mathcal{L}$ the graph space

$$H(\mathcal{L},\Omega) = \{v \in L^2(\Omega) \,:\, \mathcal{L}v \in L^2(\Omega)\}.$$

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

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

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

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

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 ds (with analogous definition of $(\cdot\,,\cdot)_{\Gamma_+}$). In terms of this notation, the boundary value problem (8.7) can be restated in the following variational form: find $u \in H(\mathcal{L},\Omega)$ such that

$$B(u,v) = \ell(v) \qquad \forall v \in H(\mathcal{L},\Omega). \tag{8.9}$$

Suppose that $\mathcal{T}_h$ is a finite element partition of the computational domain Ω into open simplicial element domains κ. It will be assumed that the family $\{\mathcal{T}_h\}_h$ is shape-regular. We then consider on $\mathcal{T}_h$ the finite element trial and test spaces $U^h = V^h \subset H^1(\Omega) \subset H(\mathcal{L}, \Omega)$ consisting of continuous piecewise polynomial functions of maximum degree k, $k \geq 1$. The finite element space U^h will be assumed to possess the following approximation property.

(H) Given that $v \in H^{s+1}(\Omega)$ and $v|_{\Gamma_-} \in H^{s+1}(\Gamma_-)$ for some s, $0 \leq s \leq k$, there exists $\pi^h v$ in U^h and a positive constant c_{int}, independent of v and the mesh function h, such that

$$\|v - \pi^h v\|_{L^2(\kappa)} + h_\kappa |v - \pi^h v|_{H^1(\kappa)} \leq c_{\text{int}} h_\kappa^{s+1} |v|_{H^{s+1}(\hat{\kappa})} \quad \forall \kappa \in \mathcal{T}_h,$$
$$\|v - \pi^h v\|_{L^2(\partial\kappa \cap \Gamma_-)} \leq c_{\text{int}} h_\kappa^{s+1} |v|_{H^{s+1}(\partial\hat{\kappa} \cap \Gamma_-)} \ \forall \kappa \in \mathcal{T}_h \,:\, \partial\kappa \cap \Gamma_- \neq \emptyset.$$

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 taking $\pi^h v$ to be the quasi-interpolant of v based on local averaging that involves the neighbours of κ (see Brenner and Scott (1994), for example). A further possibility is to define $\pi^h v \in U^h$ at the degrees of freedom interior to $\Omega \cup \Gamma_+$ as indicated in the previous sentence, while on Γ_- one can define $\pi^h v|_{\Gamma_-}$ as the orthogonal projection of $v|_{\Gamma_-}$ onto $U^h|_{\Gamma_-}$ with respect to the inner product $\langle \cdot\,, \cdot \rangle_- = ((\mathbf{b} \cdot \boldsymbol{\nu}) \cdot\,, \cdot)_{\Gamma_-}$; the inner product $\langle \cdot\,, \cdot \rangle_+$ on $L^2(\Gamma_+)$ is defined analogously.

Next we introduce the stabilized finite element approximation of our model problem. Let δ be a positive function contained in $L^\infty(\Omega)$; δ will be referred to as the *stabilization parameter*. A typical 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 and $x \mapsto h(x)$ is the local mesh size; for instance, $h|_\kappa = h_\kappa$, the diameter of element $\kappa \in \mathcal{T}_h$. The stabilized finite element approximation of (8.7) is then defined as follows:

Find $u^h \in U^h$ such that

$$B_\delta(u^h, v^h) = \ell_\delta(v^h) \quad \forall v^h \in U^h, \tag{8.10}$$

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 given by

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

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 the coefficient $\hat{c}$, we obtain different stabilization techniques; some typical choices are listed below:

$$\hat{c} = \begin{cases} 0 & \text{SDFEM,} \\ c & \text{least-squares FEM,} \\ \nabla \cdot \mathbf{b} - c & \text{Douglas–Wang stabilization.} \end{cases}$$

Condition (8.8) implies that $B_\delta(v,v) > 0$ for all $v \in U^h \setminus \{0\}$; if the Douglas–Wang stabilization is used, it has to be assumed additionally that $0 < \delta \leq \frac{1}{2}\gamma^2[c^2+(\nabla\cdot\mathbf{b})^2]^{-1}$ on $\bar\Omega$ to ensure positivity of $B_\delta(v,v)$ for nontrivial v from U^h. Since (8.10) is a linear problem over a finite-dimensional space U^h, the existence of a unique solution u^h to (8.10) follows from the positivity of $B_\delta(v,v)$, $v \in U^h \setminus \{0\}$.

Let us suppose that we wish to control the discretization error in some linear functional $J(\cdot)$ defined on $H(\mathcal{L},\Omega)+U^h$. To be more precise, suppose that a certain tolerance $\texttt{TOL} > 0$ is given and that 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 U^h associated with $\mathcal{T}_h$ such that

$$|J(u) - J(u^h)| < \texttt{TOL}.$$

In order to solve this measurement problem, we consider the *a posteriori* error analysis of the stabilized finite element method (8.10) 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. The derivation of the *a posteriori* error bound will be based on a duality argument. The dual problem is defined as follows:

Find $z \in H(\mathcal{L},\Omega)$ such that

$$B(w,z) = J(w) \quad \forall w \in H(\mathcal{L},\Omega). \tag{8.11}$$

Error representation formula and error correction. Our starting point is the following theorem.

Theorem 8.3. The dual problem (8.11) gives rise to the following error representation formula:

$$J(u) - J(u^h) = -\langle r^{h,-}, z - z^h\rangle_- + (r^h, z - z^h) - (r^h, \delta\hat{\mathcal{L}}z^h), \tag{8.12}$$

for all $z^h \in U^h$. Hence,

$$J(u) - \hat{J}_p^h(u^h; z^h) = -\langle r^{h,-}, z - z^h\rangle_- + (r^h, z - z^h) \tag{8.13}$$

for all $z^h \in U^h$, where

$$\hat{J}_p^h(u^h; z^h) = J(u^h) - (r^h, \delta\hat{\mathcal{L}}z^h),$$

$r^h = f - \mathcal{L}u^h$ is the *internal residual*, $r^{h,-} = g - u^h$.

Proof. By virtue of the linearity of J and the definition of the dual problem (8.11), we have that

$$\begin{aligned} J(u) - J(u^h) &= J(u - u^h) \\ &= B(u - u^h, z) \\ &= B(u,z) - B(u^h, z) \end{aligned}$$

$$\begin{aligned}
&= \ell(z) - B(u^h, z) \\
&= \ell(z - z^h) - B(u^h, z - z^h) + \ell(z^h) - B(u^h, z^h) \\
&= \ell(z - z^h) - B(u^h, z - z^h) \\
&\qquad + \ell(z^h) - \ell_\delta(z^h) - B(u^h, z^h) + B_\delta(u^h, z^h) \\
&= \ell(z - z^h) - B(u^h, z - z^h) \\
&\qquad + \big[(\ell - \ell_\delta)(z^h) - (B - B_\delta)(u^h, z^h)\big],
\end{aligned}$$

where z_h is any element in U^h. Hence, in agreement with Theorem 8.1, we let

$$\hat{J}^h_p(u^h; z^h) = J(u^h) + \big[(\ell - \ell_\delta)(z^h) - (B - B_\delta)(u^h, z^h)\big]$$

and note that

$$\ell(z - z^h) - B(u^h, z - z^h) = -\langle r^{h,-}, z - z^h\rangle_- + (r^h, z - z^h)$$

and

$$\hat{J}^h_p(u^h; z^h) = J(u^h) - (r^h, \delta\hat{\mathcal{L}} z^h)$$

to complete the proof. □

If we label the three terms on the right-hand side of (8.12) by I_1, II_1 and III_1, then, on general unstructured shape-regular meshes, and continuous piecewise polynomial finite elements of degree $k \geq 1$, hypothesis (H) implies that

$$\mathrm{I}_1 = \mathcal{O}(h^{2k+2}), \qquad \mathrm{II}_1 = \mathcal{O}(h^{2k+1}), \qquad \mathrm{III}_1 = \mathcal{O}(h^{k+1}),$$

and therefore $J(u) - J(u^h) = \mathcal{O}(h^{k+1})$. In fact, we shall see in the next numerical example that, on structured uniform triangular meshes, these rates of convergence may be exceeded, leading to

$$\mathrm{I}_1 = \mathcal{O}(h^{2k+2}), \qquad \mathrm{II}_1 = \mathcal{O}(h^{2k+2}), \qquad \mathrm{III}_1 = \mathcal{O}(h^{k+2}),$$

and hence $J(u) - J(u^h) = \mathcal{O}(h^{k+2})$. One way or the other, the rate of convergence of $J(u) - J(u^h)$ is dominated by III_1 whose convergence order is always inferior to those of I_1 and II_1. This motivates us to move III_1 from the right-hand side of (8.12) to the left-hand side and, as a correction term, combine it with $J(u^h)$. This then leads to the error representation formula (8.13) whose right-hand side is of size $\mathcal{O}(h^{2k+1})$. Hence $\hat{J}^h_p(u^h, z^h)$ will be a better approximation to $J(u)$ than $J(u^h)$ is. Indeed, since the term III_1 is structurally different from terms I_1 and II_1 in that it does not involve the analytical dual solution z, $\hat{J}^h_p(u^h; z^h)$ is a *computable* approximation to $J(u)$. We shall now illustrate these points through a simple numerical example using the streamline diffusion finite element method (SDFEM) with continuous piecewise polynomials of degree $k = 1$ (see Houston *et al.* (2000*a*) for details).

Example 1. Let us take $\Omega = (0,1)^2$, $\mathbf{b} = (1+x, 1+y)$, $c = 0$ and $f = 0$ with boundary condition

$$u(x,y) = \begin{cases} 1 - y^5 & \text{for } x = 0,\ 0 \le y \le 1, \\ \mathrm{e}^{-50x^4} & \text{for } 0 \le x \le 1,\ y = 0. \end{cases}$$

We select $\delta = C_\delta h$ with $C_\delta = 1/4$ and define

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

We wish to compute the weighted normal flux

$$J(u) = N_\psi(u) = \int_{\Gamma_+} (\mathbf{b} \cdot \boldsymbol{\nu}) u \, \psi \, \mathrm{d}s$$

of the analytical solution u over the outflow boundary Γ_+. For purposes of comparison, the analytical solution u and the dual solution z 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 $N_\psi(u) = 2.4676$.

In Table 8.1 we have displayed the orders of convergence, ρ, 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 U^h consisting of continuous piecewise polynomials of degree 1 ($k = 1$). 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 is of second order.

In Table 8.2 we show the convergence of each of the terms in the error representation formula (8.12). We see, in particular, that the second term in the error representation formula (8.12), *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 $\mathcal{O}(h^3)$ convergence and entirely dominates the error in the weighted outward normal flux. Thus, when term III_1 is interpreted as a computable *correction term* to the functional and is combined with $J(u^h)$, the remaining two terms, I_1 and II_1 exhibit $\mathcal{O}(h^4)$ convergence: hence, by the error representation formula (8.13) for the corrected functional $\hat{J}_p^h(u^h; z^h)$ we see that this approximates $J(u)$ with error $\mathcal{O}(h^4)$. Similar behaviour is observed on unstructured triangular meshes: there, $\mathrm{I}_1 = \mathcal{O}(h^4)$, $\mathrm{II}_1 = \mathcal{O}(h^3)$, so then $\hat{J}_p^h(u^h; z^h)$ approximates $J(u)$ with error $\mathcal{O}(h^3)$.

Table 8.1. Example 1: Convergence of $\|u - u^h\|_{L^2(\Omega)}$ with $\delta = h/4$, and the rate of convergence ρ

Mesh	$\|u - u^h\|_{L^2(\Omega)}$	ρ	$\|N_\psi(u) - N_\psi(u^h)\|$	ρ
17×17	2.927×10^{-3}	–	2.957×10^{-4}	–
33×33	5.195×10^{-4}	2.49	3.860×10^{-5}	2.94
65×65	1.079×10^{-4}	2.27	4.944×10^{-6}	2.96
129×129	2.544×10^{-5}	2.08	6.257×10^{-7}	2.98
257×257	6.260×10^{-6}	2.02	7.874×10^{-8}	2.99

Table 8.2. Example 1: Convergence of the terms in the error representation formula (8.13) with $\delta = h/4$, and the rate of convergence ρ

Mesh	I_1	ρ	II_1	ρ	III_1	ρ
17×17	3.31×10^{-6}	–	3.35×10^{-6}	–	2.96×10^{-4}	–
33×33	1.91×10^{-7}	4.12	2.30×10^{-7}	3.87	3.86×10^{-5}	2.94
65×65	1.17×10^{-8}	4.03	1.52×10^{-8}	3.92	4.95×10^{-6}	2.97
129×129	7.24×10^{-10}	4.01	9.74×10^{-10}	3.96	6.26×10^{-7}	2.98
257×257	4.51×10^{-11}	4.00	6.18×10^{-11}	3.99	7.87×10^{-8}	2.99

Table 8.3. Example 1: Convergence of the terms I_2, II_2 and III_2

Mesh	I_2	ρ	II_2	ρ	III_2	ρ
17×17	1.01×10^{-5}	–	7.43×10^{-5}	–	3.20×10^{-3}	–
33×33	4.94×10^{-7}	4.35	9.12×10^{-6}	3.03	8.16×10^{-4}	1.97
65×65	2.85×10^{-8}	4.12	1.14×10^{-6}	3.00	2.04×10^{-4}	2.00
129×129	1.73×10^{-9}	4.04	1.42×10^{-7}	3.00	5.11×10^{-5}	2.00
257×257	1.08×10^{-10}	4.01	1.78×10^{-8}	3.00	1.28×10^{-5}	2.00

Now let us consider the localized counterparts of the terms I_1, II_1 and III_1, defined by

$$\mathrm{I}_2 = \sum_{\kappa\in\mathcal{T}^h} |\langle r^{h,-}, z - z^h\rangle_{\partial\kappa\cap\Gamma_-}|,$$

$$\mathrm{II}_2 = \sum_{\kappa\in\mathcal{T}^h} |(r^h, z - z^h)_\kappa|, \qquad \mathrm{III}_2 = \sum_{\kappa\in\mathcal{T}^h} |(r^h, \delta\hat{\mathcal{L}}z^h)_\kappa|,$$

respectively.

Table 8.3 demonstrates that localization does not adversely affect the term I_1 but it does slightly affect II_1 whose localized counterpart, II_2, is now only $\mathcal{O}(h^3)$. On unstructured triangular meshes, I_2 and II_2 exhibit the same rates of convergence as I_1 and II_1, namely, $\mathcal{O}(h^4)$ and $\mathcal{O}(h^3)$, respectively. We therefore conclude that on unstructured triangular meshes the convergence rates of I_1 and II_1 are preserved under localization.

Table 8.3 also shows that global superconvergence of the term III_1 is lost under localization: term III_2 is only $\mathcal{O}(h^2)$. However, this is irrelevant, since the error representation formula (8.13) for the corrected functional $\hat{J}_p^h(u^h; z^h)$ only involves the terms I_1 and II_1, while term III_1 has become part of the corrected functional, so its localization is not required.

The insensitivity of the terms I_1 and II_1 in the error representation formula (8.13) to localization on unstructured triangular meshes implies that a Type I error bound on $|J(u) - \hat{J}_p^h(u^h; z^h)|$ will exhibit the same asymptotic rate of convergence as the error itself. Moreover, as any standard mesh refinement criterion will require the localizations I_2 and II_2 to define the local refinement indicators η_κ, the fact that I_2 and II_2 exhibit the same rates of convergence as I_1 and II_1 will be essential for ensuring the optimality of the resulting adaptive meshes.

A mesh-dependent dual problem. Still assuming that the quantity of interest is a certain linear output functional, we now explore an alternative approach to deriving an *a posteriori* error bound where, following Houston *et al.* (2000*a*), instead of $B(\cdot\,,\cdot)$, we use the stabilization-dependent bilinear form $B_\delta(\cdot\,,\cdot)$ to define the dual solution; namely, we now define the dual solution, z_δ, as the solution to the following problem:

$$B_\delta(w, z_\delta) = J(w) \qquad \forall w \in H(\mathcal{L}, \Omega). \tag{8.14}$$

Noting the Galerkin orthogonality property with respect to the bilinear functional

$$B_\delta(u - u^h, v^h) = 0 \qquad \forall v^h \in U^h,$$

we deduce the following error representation formula.

Theorem 8.4. The dual problem (8.14) gives rise to the following error representation formula:

$$J(u) - J(u^h) = -\langle r^{h,-}, z_\delta - z_\delta^h \rangle_- + (r^h, z_\delta - z_\delta^h) + (r^h, \delta \hat{\mathcal{L}}(z_\delta - z_\delta^h)) \quad (8.15)$$

for all $z_\delta^h \in U^h$.

On comparing (8.15) with the error representation formula (8.12) which stems from using the bilinear form $B(\cdot\,,\cdot)$ in the definition of the dual problem, we see that while the first two terms in the two formulae are analogous, the third term in (8.15) has now become more similar to the other terms in the representation formula in that it, too, contains the difference $z_\delta - z_\delta^h$. This is due to the fact that Galerkin orthogonality is with respect to $B_\delta(\cdot\,,\cdot)$ rather than $B(\cdot\,,\cdot)$; indeed, Galerkin orthogonality did not even enter into the derivation of (8.14). This, in turn, has some important consequences.

If we label the three terms on the right-hand side of (8.15) as $\mathrm{I}_{1,\delta}$, $\mathrm{II}_{1,\delta}$ and $\mathrm{III}_{1,\delta}$, and denote their localizations by $\mathrm{I}_{2,\delta}$, $\mathrm{II}_{2,\delta}$ and $\mathrm{III}_{2,\delta}$, repeating the numerical experiment from the previous example, we now find (see Houston *et al.* (2000*a*)) that, both on uniform and on unstructured triangular meshes and continuous piecewise linear basis functions (*i.e.*, $k = 1$),

$$\mathrm{I}_{1,\delta} = \mathcal{O}(h^4), \quad \mathrm{II}_{1,\delta} = \mathcal{O}(h^3), \quad \mathrm{III}_{1,\delta} = \mathcal{O}(h^3).$$

Furthermore,

$$\mathrm{I}_{2,\delta} = \mathcal{O}(h^4), \quad \mathrm{II}_{2,\delta} = \mathcal{O}(h^3), \quad \mathrm{III}_{2,\delta} = \mathcal{O}(h^3).$$

Thus, none of the terms in the error representation formula (8.15) is now sensitive to localization.

In the next example, we show a numerical experiment based on the stabilization-dependent dual problem (8.14). Adaptive mesh refinement is performed based on a Type I *a posteriori* error bound which stems from the error representation formula (8.15), together with the optimized mesh criterion presented in Section 7.3.

More precisely, we define the local refinement indicator

$$\begin{aligned}\eta_\kappa(u^h; z_\delta - z_\delta^h) = &-((\mathbf{b}\cdot\boldsymbol{\nu})\, r^{h,-}, z_\delta - z_\delta^h)_{\Gamma_- \cap \partial\kappa} + (r^h, z_\delta - z_\delta^h)_\kappa \\ &+ (r^h, \delta\hat{\mathcal{L}}(z_\delta - z_\delta^h))_\kappa,\end{aligned}$$

and our Type I *a posteriori* error bound is then

$$\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}} = \sum_{\kappa \in \mathcal{T}_h} |\tilde{\eta}_\kappa| \le \mathtt{TOL},$$

with

$$\tilde{\eta}_\kappa = \eta_\kappa(u^h; z_\delta^H - z_\delta^h),$$

where z_δ^H denotes the numerical solution of the stabilization-dependent dual problem (8.14) and TOL is the prescribed tolerance.

Let us suppose that the finite element space U^h consists of continuous piecewise polynomials of degree $k = 1$. Guided by the asymptotic behaviour or the terms $\mathrm{I}_{1,\delta}$, $\mathrm{II}_{1,\delta}$ and $\mathrm{III}_{1,\delta}$ and their localizations under mesh refinement, we define

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

where h_Ω and h_{Γ_-} are the mesh functions on $\Omega \cup \Gamma_+$ and Γ_-, respectively. Assuming, for example, that $\Omega \subset \mathbb{R}^2$, *i.e.*, $n = 2$, after an elementary calculation based on the use of Lagrange multipliers, as described in Section 7.3, 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. We may then explicitly solve for λ in terms of $\mathtt{TOL}$ and the integral of $A^{2/5}$, and substitute the resulting expression into the formula for h_Ω^{opt} to obtain

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

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

Example 2. Let us again consider the transport equation $\mathbf{b}\cdot\nabla u + cu = f$ in $\Omega = (0,1)^2$, but this time with $\mathbf{b} = (10y^2 - 12x + 1, 1 + y)$, $c = 0$ and $f = 0$. In this problem the characteristics enter Ω through the bottom edge and through the two vertical sides, and exit through the top edge. Thus it is admissible to impose the following inflow boundary condition:

$$u(x,y) = \begin{cases} 0 & \text{for } x = 0,\ \ 0.5 < y \le 1, \\ 1 & \text{for } x = 0,\ \ 0 < y \le 0.5, \\ 1 & \text{for } 0 \le x \le 0.5,\ \ y = 0, \\ 0 & \text{for } 0.5 < x \le 1,\ \ y = 0, \\ \sin^2(\pi y) & \text{for } x = 1,\ \ 0 \le y \le 1. \end{cases}$$

Let us suppose that the objective of the computation is to calculate the weighted normal flux $N_\psi(u)$ of the analytical solution u through the outflow

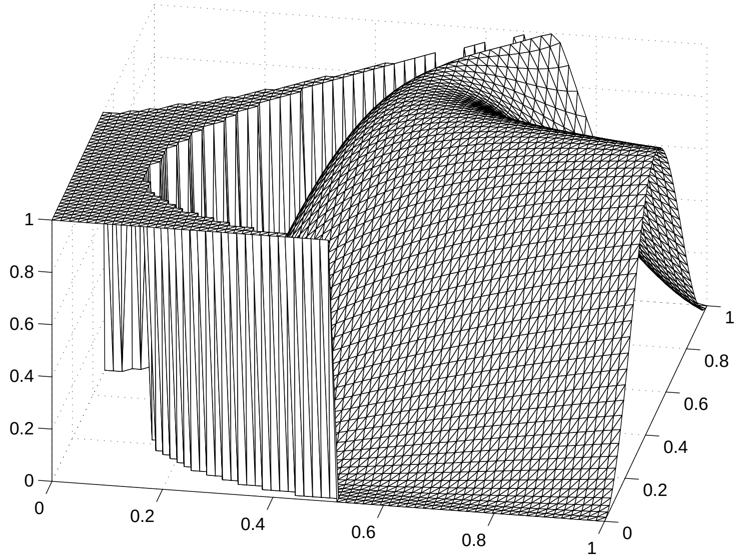


Figure 8.1. Example 2: The analytical solution to the primal problem (from Houston *et al.* (2000*a*))

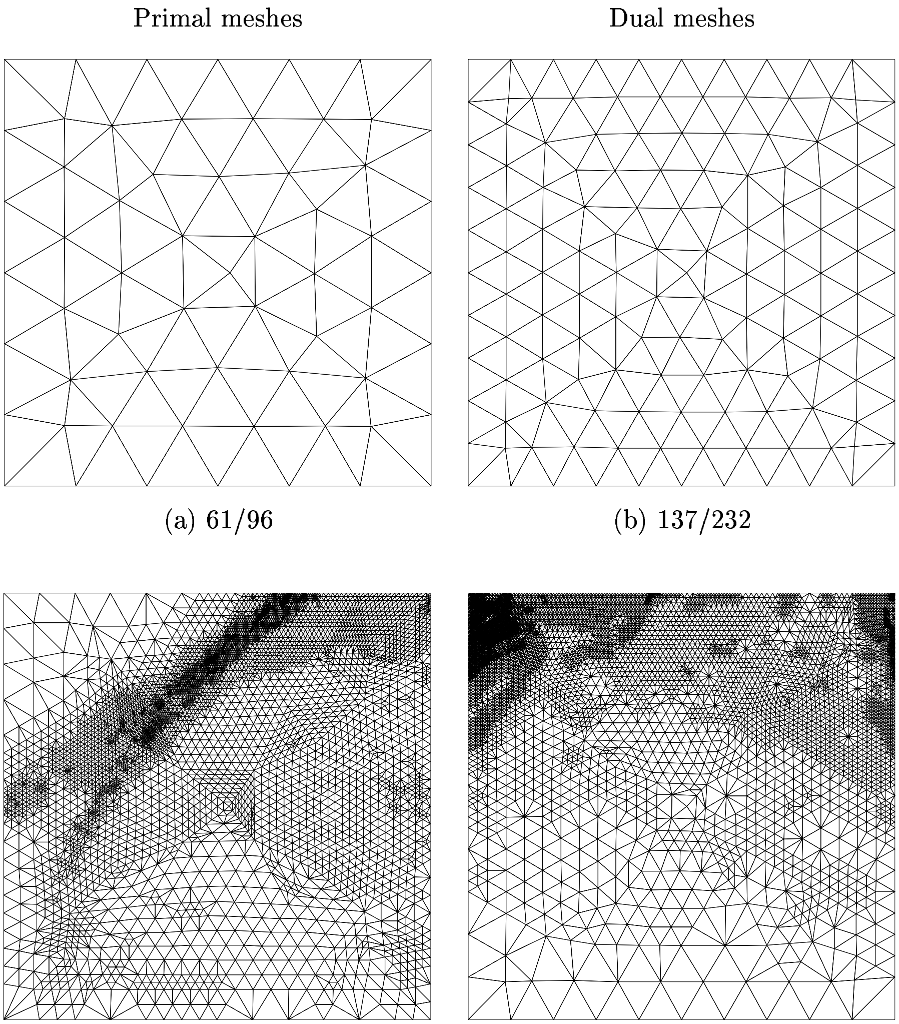
edge of the square where the weight function ψ is defined by

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

Using the method of characteristics one may compute a highly accurate approximation to u and thereby deduce that $N_\psi(u) = 0.24650$.

The analytical solution to this hyperbolic boundary value problem is shown in Figure 8.1. As the boundary datum is discontinuous along the vertical face $x = 0$ and along the horizontal face $y = 0$, the analytical solution exhibits discontinuities in Ω along the two characteristic curves that stem from the points of discontinuity on the inflow boundary. Nevertheless, numerical experiments analogous to those in Example 1 indicate that the error in the weighted outward normal flux $N_\psi(u)$ is $\mathcal{O}(h^4)$ on uniform triangular meshes and $\mathcal{O}(h^3)$ on unstructured triangular meshes.

We shall aim to compute $N_\psi(u)$ to within a prescribed tolerance $\mathtt{TOL}$. Our adaptive mesh refinement is driven by the Type I *a posteriori* error bound which stems from the error representation formula (8.15) for the stabilization-dependent dual problem. The mesh design for the primal problem is based on the *optimal mesh criterion* with $\mathtt{TOL} = 5.0 \times 10^{-5}$, and $\delta = h/4$. The background meshes for the primal and dual problems and the adaptively refined meshes which result from them are shown in Figure 8.2. We can see from Figure 8.2(c) that most of the nodes in the adaptively refined mesh for the primal problem are concentrated near the outflow



Primal meshes

Dual meshes

(a) 61/96

(b) 137/232

(c) 5648/11132

(d) 8607/17038

Figure 8.2. 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) mesh for the primal problem based on the use of the optimal mesh criterion ($\mathtt{TOL} = 5.0 \times 10^{-5}$) with 5648 nodes and 11132 elements ($|N_\psi(u) - N_\psi(u^h)| = 6.764 \times 10^{-6}$); (d) adaptively refined mesh for the stabilization-dependent dual problem (*cf.* equation (8.14)) with 7594 nodes and 14199 elements which has been constructed using the fixed fraction criterion (from Houston *et al.* (2000*a*))

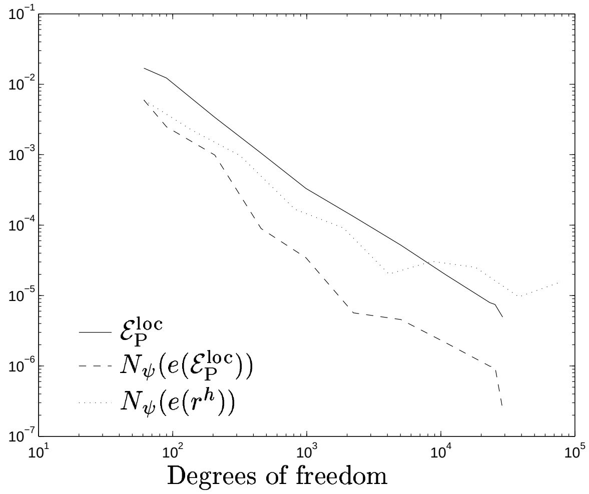


Figure 8.3. Example 2: Performance of the adaptive algorithm for TOL $= 5.0 \times 10^{-6}$ and $\delta = h/4$. $N_\psi(e(\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}}))$ denotes the error in the functional on a sequence of adaptively refined meshes using the stabilization-dependent dual problem; $\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}}$ is the corresponding Type I *a posteriori* error bound; $N_\psi(e(r^h))$ denotes the error in the functional when the adaptive meshes are generated by using $\|r^h\|_{L^2(\kappa)}$ alone as refinement indicator (from Houston *et al.* (2000*a*))

boundary where the quantity of interest, $N_\psi(u)$, is concentrated. We note, in particular, the lack of mesh refinement in the vicinity of the discontinuities as they enter the domain from $y = 0$ and $x = 0$. Had the mesh adaptation been based on disregarding the size of the dual solution and refining according to the local size of the residual alone, the resulting mesh for the primal problem would have contained heavy (and, clearly, unnecessary) refinement along the discontinuities in the primal solution.

This last point is illustrated further by the computations whose results are depicted in Figure 8.3. Here we show the performance of our adaptive algorithm with TOL $= 5.0 \times 10^{-6}$. The initial meshes are as in Figure 8.2. We see that, even though the stabilization-dependent dual problem has been solved numerically, our Type I *a posteriori* error bound $\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}}$, based on the use of the numerically computed dual solution z^H and $\pi^h z^H$ in place of z and $z^h = \pi^h z$, respectively, remains an upper bound on the true error in the approximation of the output functional $N_\psi(u)$. In Figure 8.3, $N_\psi(e(\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}}))$ 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. Figure 8.3 also shows the true error in the outward normal flux on a sequence of adaptively refined meshes which have been constructed using an empirical refinement indicator based on the local L^2-norm of the residual, $\|r^h\|_{L^2(\kappa)}$ on each element κ in $\mathcal{T}^h$ in conjunction with a fixed fraction strategy. It is clear from Figure 8.3 that the latter approach is inferior: stagnation of the error in the output functional is observed in the course of the adaptive mesh refinement.

9. *hp*-adaptivity by duality

In this section we shall briefly discuss the derivation of Type I *a posteriori* error bounds, based on a duality argument, for *hp*-version finite element methods, and the application of such bounds in *hp*-adaptive finite element algorithms. In addition to local variation of the mesh size h, adaptive algorithms of this kind admit local variation of the degree p of the approximating polynomial in the finite element space, and thereby offer greater flexibility than traditional h-version finite element methods. For the sake of brevity, here we shall focus on one particular algorithm, based on the *hp*-version of the discontinuous Galerkin finite element method (*hp*-DGFEM). The results presented in this section are based on the papers by Süli *et al.* (1999), Houston and Süli (2001*a*), Süli, Houston and Senior (2001). For the *a priori* error analysis of the *hp*-DGFEM and *hp*-SDFEM, we refer to Bey and Oden (1996), Houston, Schwab and Süli (2000*b*), Houston and Süli (2001*b*). The *hp*-version finite element methods were traditionally developed in the context of elliptic boundary value problems. Their use for the numerical solution of first-order hyperbolic systems is more recent and is motivated by the fact that, even though solutions to hyperbolic problems may exhibit local singularities and discontinuities, they are typically piecewise analytic functions. Thus, away from singularities, one may use high-degree piecewise polynomial approximations on course meshes. The relevance of *hp*-version finite element methods for the numerical solution of hyperbolic conservation laws is discussed in Bey and Oden (1996) and Adjerid, Aiffa and Flaherty (1998); see also Flaherty, Loy, Shephard and Teresco (2000) concerning implementational aspects of DGFEM. For a review of recent developments concerning the theory and application of *hp*-version finite element methods, see Ainsworth and Oden (2000), Szabó and Babuška (1991), Schwab (1998).

9.1. The model problem and its hp-DGFEM approximation

Suppose that Ω is a bounded open polyhedral domain in $\mathbb{R}^n$, $n \geq 2$, and let Γ denote the union of open faces of Ω. Let us further assume that $\mathbf{B} = (\mathbf{B}_1, \ldots, \mathbf{B}_n)$ is an n-component matrix function defined on $\bar{\Omega}$ with $\mathbf{B}_i \in [W^1_\infty(\Omega)]^{m\times m}_{\text{symm}}$, $i = 1, \ldots, n$. We shall let $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)$ denote the unit outward normal vector to Γ, and we consider the matrix

$$\mathbf{B}(\boldsymbol{\nu}) \equiv \boldsymbol{\nu} \cdot \mathbf{B} = \nu_1 \mathbf{B}_1 + \cdots + \nu_n \mathbf{B}_n.$$

Since $\mathbf{B}(\boldsymbol{\nu})$ is a symmetric matrix, it can be diagonalized; that is, we can write

$$\mathbf{B}(\boldsymbol{\nu}) = X(\boldsymbol{\nu})^{-1} \Lambda(\boldsymbol{\nu}) X(\boldsymbol{\nu}),$$

where $\Lambda(\boldsymbol{\nu})$ is a diagonal matrix, with the (real) eigenvalues of $\mathbf{B}(\boldsymbol{\nu})$ appearing along its diagonal. We shall suppose that Γ is nowhere characteristic, in the sense that none of the diagonal entries of $\Lambda(\boldsymbol{\nu})$ is zero for any choice of the unit outward normal vector $\boldsymbol{\nu}$ on Γ. Let us additively decompose the matrix $\Lambda(\boldsymbol{\nu})$ as

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

where $\Lambda_-(\boldsymbol{\nu})$ is diagonal and negative semidefinite, and $\Lambda_+(\boldsymbol{\nu})$ is diagonal and positive semidefinite. With this notation, we now define the $m \times m$ matrices

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

We then have the following induced decomposition of $\mathbf{B}(\boldsymbol{\nu})$ for each choice of the unit outward normal vector $\boldsymbol{\nu}$ on Γ:

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

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

$$\mathcal{L}\mathbf{u} \equiv \nabla \cdot (\mathbf{B}\mathbf{u}) + \mathbf{C}\mathbf{u} = \mathbf{f} \qquad \text{in } \Omega, \qquad \mathbf{B}_-(\boldsymbol{\nu})\mathbf{u} = \mathbf{B}_-(\boldsymbol{\nu})\mathbf{g} \quad \text{on } \Gamma, \quad (9.1)$$

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 $L^2(\Omega)$.

Now, let us formulate the *hp*-DGFEM for (9.1). We begin by considering a regular or 1-irregular subdivision $\mathcal{T}_h$ of Ω into disjoint open element domains κ such that $\bar{\Omega} = \cup_{\kappa \in \mathcal{T}_h} \bar{\kappa}$. By *regular or* 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 hanging node is chosen as the barycentre of the face, although this is not essential for what follows. We shall further suppose that the family of subdivisions $\mathcal{T}_h$ is shape-regular and that each $\kappa \in \mathcal{T}_h$ is a bijective affine 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 $(\hat{x}_1, \ldots, \hat{x}_n) \in \hat{\kappa}$ and $(\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0^n$, we define spaces of polynomials of degree $p \geq 1$ as follows:

$$\mathcal{Q}_p = \operatorname{span}\{\hat{x}_1^{\alpha_1} \cdots \hat{x}_n^{\alpha_n} : 0 \leq \alpha_i \leq p, \quad 1 \leq i \leq n\},$$

$$*\mathcal{P}_p = \operatorname{span}\{\hat{x}_1^{\alpha_1} \cdots \hat{x}_n^{\alpha_n} : 0 \leq \alpha_1 + \cdots + \alpha_n \leq p\}.$$

Now, to each $\kappa \in \mathcal{T}_h$ we assign an integer $p_\kappa \geq 1$; collecting the local polynomial degrees p_κ and mappings 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

$$\begin{aligned} S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F}) = \{\mathbf{v} \in [L^2(\Omega)]^m : \mathbf{v}|_\kappa \circ F_\kappa \in [\mathcal{Q}_{p_\kappa}]^m &\text{ if } F_\kappa^{-1}(\kappa) \text{ is the open} \\ \text{unit hypercube and } \mathbf{v}|_\kappa \circ F_\kappa \in [\mathcal{P}_{p_\kappa}]^m &\text{ if } F_\kappa^{-1}(\kappa) \text{ is the} \\ \text{open unit simplex; } \kappa \in \mathcal{T}_h\}. \end{aligned}$$

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

$$[H^{\mathbf{s}}(\Omega, \mathcal{T}_h)]^m = \{\mathbf{v} \in [L^2(\Omega)]^m : \mathbf{v}|_\kappa \in [H^{s_\kappa}(\kappa)]^m \quad \forall \kappa \in \mathcal{T}_h\}.$$

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

Let us suppose that κ is an element in the subdivision $\mathcal{T}_h$ of the computational domain Ω. We shall let $\partial\kappa$ denote the union of $(n-1)$-dimensional open faces of κ. Let $x \in \partial\kappa$ and suppose that $\mathbf{n}_\kappa(x)$ denotes the unit outward normal vector to $\partial\kappa$ at x. We then define $\mathbf{B}(\mathbf{n}_\kappa)$, $\mathbf{B}_-(\mathbf{n}_\kappa)$ and $\mathbf{B}_+(\mathbf{n}_\kappa)$ analogously to $\mathbf{B}(\boldsymbol{\nu})$, $\mathbf{B}_-(\boldsymbol{\nu})$ and $\mathbf{B}_+(\boldsymbol{\nu})$ above, respectively.

For each $\kappa \in \mathcal{T}_h$ and any $\mathbf{v} \in [H^1(\kappa)]^m$ we let $\mathbf{v}_\kappa^+$ denote the interior trace of $\mathbf{v}$ on $\partial\kappa$ (the trace taken from within κ). Now consider an element κ such that the set $\partial\kappa\backslash\Gamma$ is nonempty; then, for each $x \in \partial\kappa\backslash\Gamma$ (with the exception of a set of $(n-1)$-dimensional measure zero), there exists a unique element κ', depending on the choice of x, such that $x \in \partial\kappa'$. Suppose that $\mathbf{v} \in [H^1(\Omega, \mathcal{T}_h)]^m$. 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. The context should always make it clear to which element κ in the subdivision $\mathcal{T}_h$ the quantities $\mathbf{n}_\kappa$, $\mathbf{v}_\kappa^+$ and $\mathbf{v}_\kappa^-$ correspond. Thus, for the sake of simplicity of notation, we shall suppress the letter κ in the subscript and write, respectively, $\mathbf{n}$, $\mathbf{v}^+$, and $\mathbf{v}^-$ instead.

For $\mathbf{v}, \mathbf{w} \in [H^1(\Omega, \mathcal{T}_h)]^m$, we define the bilinear form of the hp-DGFEM by

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

where $\mathcal{L}^*$ is the formal adjoint of $\mathcal{L}$ defined by $\mathcal{L}^*\mathbf{v} \equiv -(\mathbf{B} \cdot \nabla)\mathbf{v} + \mathbf{C}^T\mathbf{v}$; $\mathcal{H}(\cdot\,,\cdot\,,\cdot)$ is a numerical flux function, assumed to be Lipschitz-continuous, and such that:

(i) $\mathcal{H}$ is consistent, *i.e.*, $\mathcal{H}(\mathbf{u},\mathbf{u},\mathbf{n})|_{\partial\kappa\setminus\Gamma}=\mathbf{B}(\mathbf{n})\mathbf{u}|_{\partial\kappa\setminus\Gamma}$ for all κ in $\mathcal{T}_h$;

(ii) $\mathcal{H}(\cdot,\cdot,\cdot)$ is conservative, *i.e.*, $\mathcal{H}(\mathbf{u}^+,\mathbf{u}^-,\mathbf{n})|_{\partial\kappa\setminus\Gamma}=-\mathcal{H}(\mathbf{u}^-,\mathbf{u}^+,-\mathbf{n})|_{\partial\kappa\setminus\Gamma}$.

For example, we may take

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

where $|\mathbf{B}(\mathbf{n})| = \mathbf{B}_+(\mathbf{n}) - \mathbf{B}_-(\mathbf{n})$. For $\mathbf{v}\in[H^1(\Omega,\mathcal{T}_h)]^m$, we introduce the linear functional

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

With this notation, the *hp*-DGFEM for (9.1) is defined as follows: find $\mathbf{u}_{\mathrm{DG}}\in S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$ such that

$$B_{\mathrm{DG}}(\mathbf{u}_{\mathrm{DG}},\mathbf{v}) = \ell_{\mathrm{DG}}(\mathbf{v}) \quad \forall \mathbf{v}\in S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F}). \tag{9.2}$$

9.2. A posteriori *error analysis by duality*

Let us suppose that the aim of the computation is to control the error in some linear output functional $J(\cdot)$ defined on a linear space which contains $H(\mathcal{L},\Omega)+S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$. We shall do so by deriving a Type I *a posteriori* bound on the error between $J(\mathbf{u})$ and $J(\mathbf{u}_{\mathrm{DG}})$. For this purpose we introduce the following dual problem: find $\mathbf{z}$ in $H(\mathcal{L}^*,\Omega)$ such that

$$B_{\mathrm{DG}}(\mathbf{w},\mathbf{z}) = J(\mathbf{w}) \quad \forall \mathbf{w}\in H(\mathcal{L},\Omega), \tag{9.3}$$

where $H(\mathcal{L}^*,\Omega)$ denotes the graph space of the adjoint operator $\mathcal{L}^*$ in $L^2(\Omega)$. We shall tacitly assume that (9.3) has a unique solution.

Let us define the *internal residual* $\mathbf{r}_{h,\mathbf{p}}$ on $\kappa\in\mathcal{T}_h$ by

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

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 $\mathcal{T}_h$; and, for each element κ with $\partial\kappa\cap\Gamma$ nonempty, we define the *boundary residual* $\boldsymbol{\rho}_{h,\mathbf{p}}$ by

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

Analogously, on $\partial\kappa\setminus\Gamma$, we define the *interelement flux residual* $\boldsymbol{\sigma}_{h,\mathbf{p}}$ by

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

Assuming that $\mathbf{u}$ is sufficiently smooth, it is then a simple matter to verify the following Galerkin orthogonality property:

$$
\begin{aligned}
B_{\mathrm{DG}}(\mathbf{u}-\mathbf{u}_{\mathrm{DG}},\mathbf{v}) = \sum_{\kappa\in\mathcal{T}_h}(\mathbf{r}_{h,\mathbf{p}},\mathbf{v})_\kappa &+ \sum_{\kappa\in\mathcal{T}_h}(\boldsymbol{\sigma}_{h,\mathbf{p}},\mathbf{v}^+)_{\partial\kappa\setminus\Gamma}\\
&+ \sum_{\kappa\in\mathcal{T}_h}(\boldsymbol{\rho}_{h,\mathbf{p}},\mathbf{v}^+)_{\partial\kappa\cap\Gamma} = 0 \qquad (9.4)
\end{aligned}
$$

for all $\mathbf{v}$ in $S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$. On selecting $\mathbf{w}=\mathbf{u}-\mathbf{u}_{\mathrm{DG}}$ in (9.3), the linearity of $J(\cdot)$ and (9.4) yield the following error representation formula:

$$
\begin{aligned}
J(\mathbf{u})-J(\mathbf{u}_{\mathrm{DG}}) &= J(\mathbf{u}-\mathbf{u}_{\mathrm{DG}})\\
&= B_{\mathrm{DG}}(\mathbf{u}-\mathbf{u}_{\mathrm{DG}},\mathbf{z})\\
&= B_{\mathrm{DG}}(\mathbf{u}-\mathbf{u}_{\mathrm{DG}},\mathbf{z}-\mathbf{z}_{h,\mathbf{p}})\\
&\equiv \mathcal{E}_\Omega(\mathbf{u}_{\mathrm{DG}},h,\mathbf{p},\mathbf{z}-\mathbf{z}_{h,\mathbf{p}})\\
&= \sum_{\kappa\in\mathcal{T}_h}\eta_\kappa, \qquad (9.5)
\end{aligned}
$$

where

$$
\begin{aligned}
\eta_\kappa = (\mathbf{r}_{h,\mathbf{p}},\mathbf{z}-\mathbf{z}_{h,\mathbf{p}})_\kappa &+ (\boldsymbol{\sigma}_{h,\mathbf{p}},(\mathbf{z}-\mathbf{z}_{h,\mathbf{p}})^+)_{\partial\kappa\setminus\Gamma}\\
&+ (\boldsymbol{\rho}_{h,\mathbf{p}},(\mathbf{z}-\mathbf{z}_{h,\mathbf{p}})^+)_{\partial\kappa\cap\Gamma}. \qquad (9.6)
\end{aligned}
$$

For a user-defined tolerance TOL, we now consider the problem of designing an hp-finite element space $S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$ such that

$$
|J(\mathbf{u})-J(\mathbf{u}_{\mathrm{DG}})| \le \mathtt{TOL}. \qquad (9.7)
$$

To do so, we shall use the following Type I *a posteriori* error bound, which stems from (9.5):

$$
\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}} \equiv \sum_{\kappa\in\mathcal{T}_h}|\tilde{\eta}_\kappa| \le \mathtt{TOL}, \qquad (9.8)
$$

where $\tilde{\eta}_\kappa$ is defined analogously to η_κ (*cf.* (9.6)), with $\mathbf{z}$ replaced by its numerical approximation $\tilde{\mathbf{z}}_{\mathrm{DG}}$ computed by means of the hp-DGFEM.

If the stopping criterion (9.8) is violated, then certain elements $\kappa\in\mathcal{T}_h$ will be marked for refinement; in addition to h- and p-refinement, the adaptive algorithm discussed here will also admit h- and p-derefinement. Here we shall employ the fixed fraction mesh refinement criterion, based on $\tilde{\eta}_\kappa$, with refinement and derefinement fractions set to 20% and 10%, respectively, to identify elements which will be refined/derefined. In this way, $\mathcal{T}_h$ is partitioned into three disjoint subsets:

$\mathcal{T}_h^{\mathrm{ref}}$, consisting of those elements κ in $\mathcal{T}_h$ that are marked for *refinement*;

$\mathcal{T}_h^{\mathrm{deref}}$, containing those elements $\kappa\in\mathcal{T}_h$ that are marked for *derefinement*;

$\mathcal{T}_h^{\mathrm{idle}}=\mathcal{T}_h\setminus(\mathcal{T}_h^{\mathrm{ref}}\cup\mathcal{T}_h^{\mathrm{deref}})$, the set of *idle* elements where no action is required.

If $\kappa\in\mathcal{T}_h^{\mathrm{ref}}\cup\mathcal{T}_h^{\mathrm{deref}}$, then a decision must be made as to whether the local mesh size h_κ or the local degree p_κ of the approximating polynomial should be altered. The choice between h-refinement/derefinement and p-refinement/derefinement is made by assessing the local smoothness of the primal and dual solutions $\mathbf{u}$ and $\mathbf{z}$, respectively. The various possibilities are discussed below in more detail.

`Refinement:` suppose that $\kappa \in \mathcal{T}_h^{\mathrm{ref}}$. If $\mathbf{u}$ or $\mathbf{z}$ are smooth on κ, then p-refinement will be more effective than h-refinement, since the error is then expected to decay quickly within κ as p_κ is increased. If, on the other hand, $\mathbf{u}$ and $\mathbf{z}$ have low regularity within κ, then h-refinement will be performed. In this way, regions in the computational domain where the primal or dual solutions are locally non-smooth are isolated from regions of smoothness; this then reduces the influence of singularities/discontinuities and makes p-refinement more effective. In order to ensure that the desired level of accuracy is achieved efficiently, Houston and Süli (2001*a*) developed an automatic procedure for deciding when to h- or p-refine, based on the Sobolev smoothness estimation strategy proposed in Ainsworth and Senior (1998) in the context of hp-adaptive norm control for second-order elliptic problems; for a review of recent developments on Sobolev regularity estimation, we refer to Houston and Süli (2002*a*).

`Derefinement:` suppose that $\kappa \in \mathcal{T}_h^{\mathrm{deref}}$. The derefinement strategy implemented here is to coarsen the mesh around κ 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.* Adjerid *et al.* (1998) and Houston and Süli (2001*a*)).

In Houston and Süli (2001*a*) a fully hp-adaptive algorithm has been developed; hp-adaptivity for the primal problem is controlled by a Type I *a posteriori* error bound, while the hp-adaptive algorithm for the dual is driven by a (cruder) Type II *a posteriori* error bound. Here, for the sake of 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 $\tilde{\mathbf{p}} = \mathbf{p} + \mathbf{1}$; this possibility for the numerical approximation of the dual problem was mentioned in Section 3.2. The reader is referred to Houston and Süli (2001*a*) for details concerning the implementation of a more general algorithm where $\tilde{\mathcal{T}}_h \neq \mathcal{T}_h$ and $\tilde{\mathbf{p}}$ is not required to be related to $\mathbf{p}$.

9.3. Numerical experiments

We present a numerical experiment to demonstrate the performance of the hp-adaptive algorithm. The extension to nonlinear problems will be considered in the next section.

Linear advection. In this example, we consider the scalar hyperbolic equation $\nabla \cdot (\mathbf{b}u) + cu = f$ on $\Omega = (0,1)^2$, where $\mathbf{b} = (10y^2 - 12x + 1, 1 + y)$, $c = -\nabla \cdot \mathbf{b}$ and $f = 0$. The characteristics enter the square Ω across three of its sides, *i.e.*, the two vertical faces and the bottom; they exit Ω through the top edge. We prescribe the boundary condition

$$u(x,y) = \begin{cases} 0 & \text{for } x = 0, \quad 0.5 < y \leq 1, \\ 1 & \text{for } x = 0, \quad 0 \leq y \leq 0.5, \\ 1 & \text{for } 0 \leq x \leq 0.75, \quad y = 0, \\ 0 & \text{for } 0.75 < x \leq 1, \quad y = 0, \\ \sin^2(\pi y) & \text{for } x = 1, \quad 0 \leq y \leq 1, \end{cases}$$

on the union Γ_- of the three inflow sides. The objective is to compute the weighted normal flux through the outflow side Γ_+ defined by

$$J(u) = \int_{\Gamma_+} \psi(x) u(x,1) \, \mathrm{d}x,$$

where the weight function ψ is defined by $\psi(x) = \sin(\pi x/2)$ for $0 \leq x \leq 1$; thereby, the true value of the functional is $J(u) = 0.246500283257585$ (*cf.* Houston *et al.* (2000*a*)).

In Table 9.1, we show the performance of the adaptive algorithm: we give the number of nodes (Nds), elements (Els) and degrees of freedom (DOF)

Table 9.1. Adaptive algorithm for the linear advection problem

Nds	Els	DOF	$J(u - u_{\mathrm{DG}})$	$\sum_\kappa \tilde{\eta}_\kappa$	θ_1	$\sum_\kappa \lvert\tilde{\eta}_\kappa\rvert$	θ_2
25	16	64	0.1207E-02	0.1023E-02	0.85	0.1938E-01	16.06
30	19	86	-0.8405E-02	-0.8203E-02	0.98	0.1006E-01	1.20
48	31	202	-0.6729E-02	-0.6002E-02	0.89	0.7279E-02	1.08
48	31	244	-0.1611E-02	-0.1623E-02	1.01	0.1927E-02	1.20
57	37	330	-0.9690E-03	-0.9756E-03	1.01	0.1043E-02	1.08
87	61	595	-0.8424E-03	-0.8581E-03	1.02	0.8654E-03	1.03
129	91	1078	-0.1075E-04	-0.4017E-04	3.74	0.4731E-04	4.40
139	100	1439	0.2691E-04	0.2906E-04	1.08	0.3580E-04	1.33
201	148	2490	-0.1456E-05	-0.1290E-05	0.89	0.2808E-05	1.93
263	199	3723	-0.4938E-06	-0.6040E-06	1.22	0.6721E-06	1.36
308	232	4876	-0.1196E-07	-0.1123E-07	0.94	0.4792E-07	4.01
383	292	6793	-0.5294E-08	-0.5296E-08	1.00	0.6621E-08	1.25
429	328	8548	-0.3450E-08	-0.3457E-08	1.00	0.4322E-08	1.25
542	418	12325	-0.1650E-09	-0.1676E-09	1.02	0.2047E-09	1.24

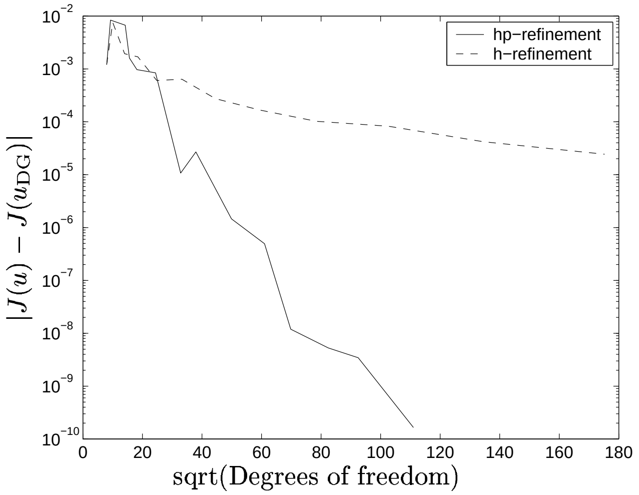


Figure 9.1. Comparison between h- and hp-adaptive mesh refinement for the linear problem

in $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$, the true error in the functional $J(u - u_{\mathrm{DG}})$, the computed error representation formula $\sum_{\kappa\in\mathcal{T}_h} \tilde{\eta}_\kappa$, the Type I *a posteriori* error bound $\mathcal{E}_{\mathrm{P}}^{\mathrm{loc}} = \sum_{\kappa\in\mathcal{T}_h} |\tilde{\eta}_\kappa|$, and their respective effectivity indices θ_1 and θ_2. We see that initially, on very coarse meshes, the quality of the computed error representation formula is quite poor, in the sense that θ_1, the ratio of the error representation formula and the error $J(u - u^h)$, is not close to one; however, as the mesh is refined the effectivity index θ_1 approaches unity. Furthermore, we observe that the Type I *a posteriori* error bound is indeed sharp, in the sense that the second effectivity index $\theta_2 = \mathcal{E}_{\mathrm{P}}^{\mathrm{loc}}/J(u - u_{\mathrm{DG}})$ overestimates the true error in the computed functional by a consistent factor as the finite element space $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$ is enriched.

Figure 9.1 shows $|J(u) - J(u_{\mathrm{DG}})|$, 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 an initial increase, the error in the approximation to the output functional using hp-refinement becomes (on average) a straight line, thereby indicating exponential convergence of $J(u_{\mathrm{DG}})$ to $J(u)$, despite the fact that u is only piecewise continuous; this occurs since z is a real analytic function on $\bar{\Omega}$. Figure 9.1 also highlights the superiority of the adaptive hp-refinement strategy over a traditional adaptive h-refinement algorithm. On the final mesh, the true error between $J(u)$ and $J(u_{\mathrm{DG}})$ using hp-refinement is almost 6 orders of magnitude smaller than the corresponding quantity when h-refinement is employed alone.

9.4. Adaptivity for nonlinear problems

Multi-dimensional compressible fluid flows are modelled by nonlinear conservation laws whose solutions exhibit a wide range of localized structures, such as shock waves, contact discontinuities and rarefaction waves. The accurate numerical resolution of these features necessitates the use of locally refined, adaptive computational meshes. Here we describe the development of Type I *a posteriori* error bounds for *hp*-version discontinuous Galerkin finite element approximations of nonlinear systems of conservation laws, following Süli *et al.* (2001), which is the extension of the *h*-version *a posteriori* error analysis in Larson and Barth (2000), Hartmann (2001), Hartmann and Houston (2001) and Süli *et al.* (2001).

Given a bounded open polyhedral domain Ω in $\mathbb{R}^n$, $n \geq 1$, let Γ denote the union of open faces contained in $\partial\Omega$. We consider the following problem: find $\mathbf{u} : \Omega \to \mathbb{R}^m$, $m \geq 1$, such that

$$\mathrm{div}\mathcal{F}(\mathbf{u}) = 0 \qquad \text{in } \Omega, \tag{9.9}$$

where $\mathcal{F} : \mathbb{R}^m \to \mathbb{R}^{m\times n}$ is continuously differentiable. We assume that the system of conservation laws (9.9) may be supplemented by appropriate initial/boundary conditions. In other words, we assume that

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

has m real eigenvalues and a complete set of linearly independent eigenvectors for all $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$; then at inflow/outflow boundaries, we require that $B_-(\mathbf{u}, \boldsymbol{\nu})(\mathbf{u} - \mathbf{g}) = \mathbf{0}$, where $\boldsymbol{\nu}$ denotes the unit outward normal vector to $\partial\Omega$, $B_-(\mathbf{u}, \boldsymbol{\nu})$ is the negative part of $B(\mathbf{u}, \boldsymbol{\nu})$ and $\mathbf{g}$ is a (given) real-valued function.

The *hp*-DGFEM for (9.9) is defined as follows: find $\mathbf{u}_{\mathrm{DG}} \in S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$ such that

$$\sum_{\kappa\in\mathcal{T}_h} \left\{ -\int_\kappa \mathcal{F}(\mathbf{u}_{\mathrm{DG}}) \cdot \nabla \mathbf{v}_{h,\mathbf{p}} \,\mathrm{d}x + \int_{\partial\kappa} \mathcal{H}(\mathbf{u}_{\mathrm{DG}}^+, \mathbf{u}_{\mathrm{DG}}^-, \boldsymbol{\nu}_\kappa)\, \mathbf{v}_{h,\mathbf{p}}^+ \,\mathrm{d}s \right.$$
$$\left. + \int_\kappa \varepsilon \nabla \mathbf{u}_{\mathrm{DG}} \cdot \nabla \mathbf{v}_{h,\mathbf{p}} \,\mathrm{d}x \right\} = 0 \tag{9.10}$$

for all $\mathbf{v}_{h,\mathbf{p}} \in S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$ (*cf.* Jaffre, Johnson and Szepessy (1995), Hartmann and Houston (2001), Houston, Hartmann and Süli (2001), Süli *et al.* (2001), for example); here $\boldsymbol{\nu}_\kappa$ denotes the unit outward normal vector to κ, and $\mathcal{H}(\cdot\,,\cdot\,,\cdot)$ is a *numerical flux* function, assumed to be Lipschitz-continuous, consistent and conservative. We emphasize that the choice of the numerical flux function is completely independent of the finite element space employed; in the numerical experiments we use the (local) Lax–Friedrichs flux. Further, the parameter ε denotes the coefficient of artificial viscosity defined, on $\kappa \in \mathcal{T}_h$, by

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

where C_ε is a positive constant and $0 < \beta < 1/2$; see Jaffre *et al.* (1995). For elements $\kappa \in \mathcal{T}_h$ whose boundary intersects $\partial\Omega$, $\mathbf{u}_h^-$ is replaced by appropriate boundary/initial conditions on $\partial\kappa \cap \partial\Omega$.

Let us suppose that we are concerned with computing $J(\mathbf{u})$, where $J(\cdot)$ is a given linear output functional. Letting $\mathcal{N}(\mathbf{u}_{\mathrm{DG}}, \mathbf{v}_{h,\mathbf{p}})$ denote the left-hand side of (9.10), we write $\mathcal{M}(\mathbf{u}, \mathbf{u}_{\mathrm{DG}}; \cdot\,, \cdot)$ to denote the mean-value linearization of $\mathcal{N}(\cdot\,, \cdot)$ given by

$$\begin{aligned}\mathcal{M}(\mathbf{u}, \mathbf{u}_{\mathrm{DG}}; \mathbf{u} - \mathbf{u}_{\mathrm{DG}}, \mathbf{v}) &= \mathcal{N}(\mathbf{u}, \mathbf{v}) - \mathcal{N}(\mathbf{u}_{\mathrm{DG}}, \mathbf{v}) \\ &= \int_0^1 \mathcal{N}'_{\mathbf{u}}[\theta\mathbf{u} + (1-\theta)\mathbf{u}_{\mathrm{DG}}](\mathbf{u} - \mathbf{u}_{\mathrm{DG}}, \mathbf{v})\,\mathrm{d}\theta \qquad (9.11)\end{aligned}$$

for all $\mathbf{v}$ in V, where V is a suitable function space such that $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F}) \subset V$. 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 introduced in (9.11) 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.* Hartmann and Houston (2001)). Further, we shall suppose that the linearization (9.11) is well defined. Under these hypotheses, we introduce the following *dual* problem: find $\mathbf{z} \in V$ such that

$$\mathcal{M}(\mathbf{u}, \mathbf{u}_{\mathrm{DG}}; \mathbf{w}, \mathbf{z}) = J(\mathbf{w}) \quad \forall \mathbf{w} \in V. \qquad (9.12)$$

As in the linear case considered earlier, we shall tacitly assume that (9.12) possesses a unique solution. We then have the following error representation formula.

Theorem 9.1. Let $\mathbf{u}$ and $\mathbf{u}_{\mathrm{DG}}$ denote the solutions of (9.9) and (9.10), respectively, and suppose that the dual problem (9.12) is well posed. Then,

$$J(\mathbf{u}) - J(\mathbf{u}_{\mathrm{DG}}) = \mathcal{E}_\Omega(\mathbf{u}_{\mathrm{DG}}, h, \mathbf{p}, \mathbf{z} - \mathbf{z}_{h,\mathbf{p}}) \equiv \sum_{\kappa \in \mathcal{T}_h} \eta_\kappa, \qquad (9.13)$$

where

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

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

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

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

Proof. The proof is elementary. We choose $\mathbf{w} = \mathbf{u} - \mathbf{u}_{\mathrm{DG}}$ in (9.12), recall the linearity of $J(\cdot)$, and exploit the Galerkin orthogonality property $\mathcal{N}(\mathbf{u}, \mathbf{v}_{h,\mathbf{p}}) - \mathcal{N}(\mathbf{u}_{\mathrm{DG}}, \mathbf{v}_{h,\mathbf{p}}) = 0$ for all $\mathbf{v}_{h,\mathbf{p}}$ in $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$, to deduce that

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

for all $\mathbf{z}_{h,\mathbf{p}}$ in $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$. Equation (9.13) now follows by employing the divergence theorem. □

The error representation formula implies the following Type I *a posteriori* error bound.

Corollary 9.2. Under the assumptions of Theorem 9.1, we have

$$|J(\mathbf{u}) - J(\mathbf{u}_{\mathrm{DG}})| \lessapprox \mathcal{E}_{\mathrm{P}}^{\mathrm{loc}} \equiv \sum_{\kappa \in \mathcal{T}_h} |\tilde{\eta}_\kappa|, \tag{9.14}$$

where $\tilde{\eta}_\kappa$ is defined in the same way as η_κ, except that a numerical approximation to the dual solution is used in $\tilde{\eta}_\kappa$, in place of the analytical dual solution $\mathbf{z}$ appearing in η_κ.

We see that, in contrast with the linear problems considered earlier on, for nonlinear hyperbolic conservation laws the error representation formula (9.13) depends on the unknown analytical solutions to the primal and dual problems. Thus, to render the Type I *a posteriori* error bound (9.14) computable, now both $\mathbf{u}$ and $\mathbf{z}$ must be replaced by suitable approximations. In particular, the linearization leading to $\mathcal{M}(\mathbf{u}, \mathbf{u}_{\mathrm{DG}}; \cdot, \cdot)$ is performed about $\mathbf{u}_{\mathrm{DG}}$ and the dual solution $\mathbf{z}$ is replaced by a discontinuous Galerkin approximation computed on the same mesh $\mathcal{T}_h$ used for $\mathbf{u}_{\mathrm{DG}}$, but using piecewise polynomials whose local degree is by 1 higher than the local degree of $\mathbf{u}_{\mathrm{DG}}$. Our final example concerns the steady compressible Euler equations of gas dynamics.

Example: Ringleb's flow. We consider the steady compressible Euler equations

$$\sum_{j=1}^{n} \frac{\partial}{\partial x_j} \mathbf{F}_j(\mathbf{U}) = \mathbf{0} \tag{9.15}$$

where

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

is the vector of *conserved variables*,

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

are the fluxes. For an ideal gas, the density ρ and pressure p are related through the *equation of state*

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

involving the total energy E and the velocity vector $\mathbf{u} = (u_1, \ldots, u_n)^{\mathrm{T}}$ in Cartesian coordinates. Here, κ is the ratio of specific heats; for dry air, $\kappa = 1.405$.

We consider Ringleb's flow, in two space dimensions, for which an analytical solution may be obtained using the hodograph method. This problem represents a transonic flow which turns around an obstacle; the flow is mostly subsonic, with a small supersonic pocket around the nose of the obstacle (see Barth (1998), Süli *et al.* (2001)).

We take the functional of interest to be the value of the density at the point $(-0.4, 2)$, that is, $J(\mathbf{u}) = \rho(-0.4, 2)$; consequently the true value of the functional is given by $J(\mathbf{u}) = 0.8616065996968034$. Table 9.2 shows the performance of our *hp*-adaptive algorithm; here we 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 (9.8) overestimates the true error in the computed functional by

Table 9.2. Adaptive algorithm for Ringleb's flow

Nds	Els	DOF	$J(u-u^h)$	$\sum_\kappa \tilde{\eta}_\kappa$	θ_1	$\sum_\kappa \lvert\tilde{\eta}_\kappa\rvert$	θ_2
91	144	1728	-0.2995E-03	-0.3024E-03	1.01	0.2914E-02	9.73
129	219	3228	0.5143E-04	0.4975E-04	0.97	0.9527E-03	18.52
179	315	5312	-0.1884E-04	-0.1885E-04	1.00	0.2484E-03	13.18
245	445	8560	0.4813E-05	0.4303E-05	0.89	0.1164E-03	24.19
302	554	13480	-0.2541E-05	-0.2662E-05	1.05	0.4230E-04	16.65
352	650	17944	-0.1489E-05	-0.1520E-05	1.02	0.1824E-04	12.25
426	792	26260	-0.5522E-06	-0.5662E-06	1.03	0.5515E-05	9.99
487	912	35280	-0.2602E-07	-0.2615E-07	1.01	0.6618E-06	25.44
622	1171	51544	0.6738E-09	0.6335E-09	0.94	0.1119E-06	166.02

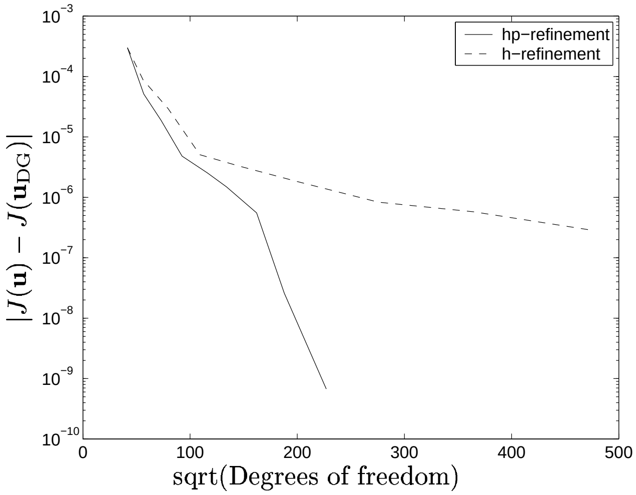


Figure 9.2. Comparison between *h*- and *hp*-adaptive mesh refinement for Ringleb's flow

about an order of magnitude, though there is a sharp increase on the last refinement. Figure 9.2 indicates exponential convergence for the error in the computed functional and again highlights the computational advantages of employing *hp*-mesh refinement when compared with the standard *h*-method, particularly when the output functional is required with high accuracy.

10. Conclusions and outlook

In this paper we have been concerned with the application of duality arguments to the derivation of *a priori* and *a posteriori* error bounds on the error in output functionals. We also discussed the role of adjoint equations in the process of error correction.

Looking to the future, there are many challenges to be addressed; below we discuss some of these.

Reconstruction on unstructured grids

Section 6 presented some preliminary ideas for reconstruction on unstructured grids. The analysis showed that if the error in the original solution is $\mathcal{O}(h^2)$ in the $L^2(\Omega)$-norm, but $\mathcal{O}(h)$ in $H^1(\Omega)$, then the reconstruction will have an improved accuracy of at least $\mathcal{O}(h^{3/2})$ in $H^1(\Omega)$. However, the ideal would be to achieve an accuracy of $\mathcal{O}(h^2)$ in $H^1(\Omega)$. There was also no discussion of how the analytic reconstruction equation might be approximated numerically.

Clearly an appropriate finite element discretization needs to be formulated using $H^2(\Omega)$-conforming (*e.g.*, C^1) finite elements. Numerical experiments must then be performed for a variety of test cases to establish the accuracy of the reconstruction in practice. If the error of the reconstructed solution is found to be $\mathcal{O}(h^2)$ in the $H^1(\Omega)$-norm, then further research is in order to try to improve the analysis, perhaps by including further assumptions concerning the formulation of the reconstruction algorithm.

Grid adaptation for multiple functionals

The error analysis and grid adaptation in this paper has been driven by concern for one particular output functional. However, in practice one might be interested in the simultaneous approximation of several functionals, such as both lift and drag in a CFD calculation.

One way to treat this situation would be to perform separate error analyses for each functional of interest, and then define a composite grid adaptation criterion. The obvious drawback of this, however, is the increased computational cost.

An alternative approach to grid adaptation might be to use a refinement criterion like (7.25), but instead of the weight $\omega_{\kappa,\tau}$ being based on the dual solution for a particular functional, it could instead be constructed to be representative of the dual solutions for a class of functionals. For example, when performing airfoil or aircraft calculations, the functionals of interest are almost always surface integrals. Analysis of the homogeneous adjoint flow equations will reveal the asymptotic behaviour of the dual solution in the far-field, and thus one might construct a weight $\omega_{\kappa,\tau}$ which would, at least qualitatively, have approximately the correct magnitude for a range of smoothly weighted boundary integral functionals.

Singularities and discontinuities

A priori error analysis usually leads to results proving that the error in the approximate solution (or the value of an output functional if that is of more interest) is $\mathcal{O}(h^p)$ for some p, provided the analytic solution is sufficiently smooth. Here h is the maximum element size (defined perhaps as the diameter of the smallest enclosing circumsphere) which is proportional to $N^{-1/n}$ for a quasi-uniform n-dimensional grid with N elements.

In practice, the analytic solution often does not satisfy the smoothness conditions required for the maximum value for p. This is frequently due to singularities because of corners in the domain boundary, or due to discontinuities in the boundary data. Under such circumstances, the best that can be hoped for is that the accuracy remains $\mathcal{O}(N^{-p/n})$, with $h \ll N^{-1/n}$ in the neighbourhood of the singularity.

For certain problems, there are indeed *a priori* proofs that this can be achieved with the appropriate local grid resolution. For such cases, to be considered *quasi-optimal*, an adaptive grid strategy should automatically generate such local grid resolution and hence the optimal order of accuracy. However, there has been little work so far on *a priori* proofs of the optimality of grid refinement indicators (see, for example, the papers of Gui and Babuška (1986), Section 3.3.7 of the monograph of Schwab (1998) and the work of Larson (1996)).

Anisotropic adaptation

The adaptive grid strategies discussed in this paper all use grid refinement, adding additional nodes/cells through an isotropic refinement process that improves the grid resolution in each direction. This is appropriate in many applications, but far from ideal in others.

One example is the inviscid flow around a wing. Here the grid resolution normal to the leading edge needs to be much finer than the spanwise resolution. In this case, anisotropic refinement is probably the best solution. This means adding nodes in such a way that the resolution normal to the leading edge is greater than in the spanwise resolution. Another, more extreme, example of the need for anisotropic resolution is a contact discontinuity in the solution to a hyperbolic partial differential equation. In this case, the best solution may well be grid redistribution, moving existing grid nodes to provide the resolution where it is needed.

The questions are how to decide which direction requires additional resolution, and how to move the nodes in grid redistribution? There are existing methods for doing this (Habashi, Fortin, Dompierre, Vallet and Bourgault 1998) but they are somewhat *ad hoc* in nature, although they often work well in practice. The challenge for those developing *a posteriori* adjoint-based refinement indicators is to formulate extensions to address this issue and provide a reliably good adaptive strategy. For recent work in the area of error estimation on anisotropic meshes we refer to Dobrowolski, Gräf and Pflaum (1999), Skalický and Roos (1999), Schötzau, Schwab and Stenberg (1998), Schötzau, Schwab and Stenberg (1999), Dolejši (2001), Apel, Nicaise and Schöberl (2001), Formaggia, Perotto and Zunino (2002).

Shocks

One last challenge we wish to highlight is the problem of shocks. With the quasi-1D Euler equations, it can be proved that, with an appropriate conservative formulation, and a numerical discretization that is second-order accurate when the solution is smooth, the accuracy of output functionals such as the integrated pressure is also second-order (Giles 1996). However, numerical evidence suggests this is not the case in multiple dimensions, and instead there is an error in quantities such as the lift on a transonic airfoil that is proportional to the local grid spacing at the shock. Thus, to get even second-order accuracy in the lift and in the solution on either side of the shock would require anisotropic grid adaptation so that the grid spacing at the shock is $\mathcal{O}(h^2)$, with h here being the average grid spacing in the rest of the grid.

There is another much more fundamental problem in the use of adjoint solutions for error analysis and correction. The approximate primal solution will have an $\mathcal{O}(1)$ error at the shock. This violates the whole basis for the adjoint error analysis since it relies on a linearization of the nonlinear equations that is valid only for small perturbations. The solution to this problem may be to use a regularization in which one numerically approximates a viscous shock with the level of viscosity being $\mathcal{O}(h^2)$. Grid adaptation would be based on the error in approximating the viscous equations, which would automatically lead to termination of the grid refinement in the neighbourhood of the shock once it is sufficiently well resolved. To apply the adjoint error correction to an improved order of accuracy for functionals, one would have to correct for the numerical error in approximating the viscous shock, plus the analytic error in using the viscous shock problem to approximate the inviscid shock problem. Through the use of matched asymptotic expansions, it can be proved that, to leading order, there is a linear dependence of integral functionals on the level of viscosity. Thus the analytic error can be compensated for by using the viscous dual solution to give the sensitivity of the lift to a change in the level of the viscosity.

Acknowledgements

The authors would like to express their sincere gratitude to Rémi Abgrall, Mark Ainsworth, Ivo Babuška, Tim Barth, Bernardo Cockburn, Leszek Demkowicz, Joe Flaherty, Kathryn Harriman, Paul Houston, Claes Johnson, Mats Larson, John Mackenzie, Peter Monk, Niles Pierce, Rolf Rannacher, Christoph Schwab, Thomas Sonar, and Gerald Warnecke for helpful discussions on the subject of this paper. We are particularly indebted to Kathryn Harriman, Paul Houston and Niles Pierce for their kind permission to include several of their computational results.

REFERENCES

R. A. Adams (1975), *Sobolev Spaces*, Academic Press, New York.

S. Adjerid, M. Aiffa and J. E. Flaherty (1998), Computational methods for singularly perturbed systems, in *Singular Perturbation Concepts of Differential Equations* (J. Cronin and R. E. O'Malley, eds), AMS, Providence, RI.

M. Ainsworth and J. T. Oden (2000), *A posteriori Error Estimation in Finite Element Analysis*, Wiley.

M. Ainsworth and B. Senior (1998), 'An adaptive refinement strategy for *hp*-finite element computations', *Appl. Numer. Math.* **26**, 165–178.

C. Airiau (2001), 'Non-parallel acoustic receptivity of a Blasius boundary using an adjoint approach', *Flow, Turbulence and Combustion* **65**, 347–367.

J. A. Alden and R. G. Compton (1997), 'A general method for electrochemical simulations, Part 1: Formulation of the strategy for two-dimensional simulations', *J. Phys. Chem. B* **101**, 8941–8954.

T. Apel, S. Nicaise and J. Schöberl (2001), 'Crouzeix–Raviart type finite elements on anisotropic meshes', *Numer. Math.* **89**, 193–223.

J.-P. Aubin (1967), 'Behavior of the error of the approximate solutions of boundary value problems for linear elliptic operators by Galerkin and finite difference methods', *Ann. Scuola. Norm. Sup. Pisa* **3**, 599–637.

I. Babuška and A. Miller (1984*a*), 'The post processing approach in the finite element method, Part 1: Calculation of displacements, stresses and other higher derivatives of the displacements', *Internat. J. Numer. Methods Engr.* **34**, 1085–1109.

I. Babuška and A. Miller (1984*b*), 'The post processing approach in the finite element method, Part 2: The calculation of stress intensity factors', *Internat. J. Numer. Methods Engr.* **34**, 1111–1129.

I. Babuška and A. Miller (1984*c*), 'The post processing approach in the finite element method, Part 3: *A posteriori* estimates and adaptive mesh selection', *Internat. J. Numer. Methods Engr.* **34**, 1131–1151.

C. Bardos (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.

J. W. Barrett, G. Moore and K. W. Morton (1988), 'Optimal recovery in the finite element method, Part 2: Defect correction for ordinary differential equations', *IMA J. Numer. Anal.* **8**, 527–540.

J. W. Barrett and C. M. Elliott (1987), 'Total flux estimates for a finite element approximation of elliptic equations', *IMA J. Numer. Anal.* **7**, 129–148.

T. J. Barth (1998), Numerical methods for gas dynamics systems on unstructured meshes, in *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws* (D. Kröner, M. Ohlberger and C. Rohde, eds), Vol. 5 of *Lecture Notes in Computational Science and Engineering*, Springer, Berlin/Heidelberg, pp. 195–285.

R. Becker and R. Rannacher (1996), A feed-back approach to error control in finite element methods: Basic analysis and examples, *East–West J. Numer. Math.* **4**, 237–264.

R. Becker and R. Rannacher (2001), An optimal control approach to *a posteriori* error estimation in finite element methods, in *Acta Numerica*, Vol. 10 (A. Iserles, ed.), Cambridge University Press, Cambridge, pp. 1–102.

T. R. Bewley (2001), 'Flow control: New challenges for a new Renaissance', *Progress in Aerospace Sciences* **37**, 21–58.

K. S. Bey and J. T. Oden (1996), '*hp*-version discontinuous Galerkin methods for hyperbolic conservation laws', *Comput. Methods Appl. Mech. Engr.* **133**, 259–286.

A. Bottaro, J. Mauss and D. S. Henningson, eds (2001), *Flow, Turbulence and Combustion*, Special Issue.

D. Braess (1997), *Finite Elements. Theory, Fast Solvers Applications in Solid Mechanics*, Cambridge University Press, Cambridge.

S. C. Brenner and L. R. Scott (1994), *The Mathematical Theory of Finite Element Methods*, Springer, Berlin/Heidelberg.

C. Carstensen and S. A. Funken (2000), 'Constants in Clément-interpolation error and residual based *a posteriori* error estimates in finite element methods', *East–West J. Numer. Math.* **8**, 153–175.

P. G. Ciarlet (1978), *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam.

B. Cockburn, G. E. Karniadakis and C.-W. Shu (2000), The development of discontinuous Galerkin methods, in *Discontinuous Galerkin Finite Element Methods* (B. Cockburn, G. E. Karniadakis and C.-W. Shu, eds), Vol. 11 of *Lecture Notes in Computational Science and Engineering*, Springer, Berlin/Heidelberg, pp. 3–50.

D. Colton and R. Kress (1991), *Inverse Acoustic and Electromagnetic Scattering Theory*, Vol. 93 of *Applied Mathematical Sciences*, Springer.

R. Dautray and J.-L. Lions (1993), *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 6: *Evolution Problems II*, Springer, Berlin/Heidelberg.

M. Dobrowolski, S. Gräf and C. Pflaum (1999), 'On *a posteriori* error estimators in the finite element method on anisotropic meshes', *Electron. Trans. Numer. Anal.* **8**, 36–45.

V. Dolejši (2001), 'Anisotropic mesh adaptation technique for viscous flow simulation', *East–West J. Numer. Math.* **1**, 1–24.

K. Eriksson, D. Estep, P. Hansbo and C. Johnson (1995), Introduction to adaptive methods for differential equations, in *Acta Numerica*, Vol. 4 (A. Iserles, ed.), Cambridge University Press, Cambridge, pp. 105–158.

K. Eriksson, D. Estep, P. Hansbo and C. Johnson (1996), *Computational Differential Equations*, Cambridge University Press.

J. E. Flaherty, R. M. Loy, M. S. Shephard and J. D. Teresco (2000), Software for the parallel adaptive solution of conservation laws by discontinuous Galerkin methods, in *Discontinuous Galerkin Finite Element Methods* (B. Cockburn, G. E. Karniadakis and C.-W. Shu, eds), Vol. 11 of *Lecture Notes in Computational Science and Engineering*, Springer, Berlin/Heidelberg, pp. 113–124.

L. Formaggia, S. Perotto and P. Zunino (2002), An anisotropic *a-posteriori* error estimate for a convection diffusion equation. EPFL-DMA Analyse et Analyse Numérique Report no. 04. To appear in *Computing and Visualization in Science.*

D. Gilbarg and N. S. Trudinger (1983), *Elliptic Partial Differential Equations of Second Order*, 2nd edn, Springer, Berlin/Heidelberg.

M. B. Giles (1996), 'Analysis of the accuracy of shock-capturing in the steady quasi-1D Euler equations', *Internat. J. Comput. Fluid Dynamics* **5**, 247–258.

M. B. Giles (1998), On adjoint equations for error analysis and optimal grid adaptation in CFD, in *Frontiers of Computational Fluid Dynamics 1998* (D. Caughey and M. Hafez, eds), World Scientific, pp. 155–170.

M. B. Giles (2000), An introduction to the adjoint design approach and analysis, Numerical Analysis Group Research Report NA-00/04, University of Oxford.

M. B. Giles (2001), Defect and adjoint error correction, in *Computational Fluid Dynamics 2000* (N. Satofuka, ed.), Springer.

M. B. Giles and N. A. Pierce (1997), 'Adjoint equations in CFD: Duality, boundary conditions and solution behaviour', AIAA Paper 97–1850.

M. B. Giles and N. A. Pierce (1999), 'Improved lift and drag estimates using adjoint Euler equations', AIAA Paper 99–3293.

M. B. Giles and N. A. Pierce (2001), 'Analysis of adjoint error correction for superconvergent functional estimates'. Submitted to *SIAM J. Numer. Anal.*

M. B. Giles and N. A. Pierce (2002), 'Adjoint error correction for integral outputs', *NASA Ames/VKI Lecture Series on Error Estimation and Solution Adaptive Discretization in CFD* (T. J. Barth and H. Deconinck, eds), *Lecture Notes in Computational Science and Engineering*, Springer. To appear.

M. B. Giles, M. Larsson, M. Levenstam and E. Süli (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.

V. Girault and P.-A. Raviart (1986), *Finite Element Methods for Navier–Stokes Equations*, Springer, Berlin/Heidelberg.

W. Gui and I. Babuška (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.

W. G. Habashi, M. Fortin, J. Dompierre, M.-G. Vallet and Y. Bourgault (1998), Anisotropic mesh adaptation: A step towards mesh-independent and user-independent CFD, in *Barriers and Challenges in Fluid Dynamics* (Hampton, VA, 1996), Vol. 6 of *ICASE/LaRC Interdiscip. Ser. Sci. Engr.*, pp. 99–117.

P. Hansbo and C. Johnson (1991), 'Adaptive streamline diffusion finite element methods for compressible flow using conservative variables.', *Comput. Methods Appl. Mech. Engr.* **87**, 267–280.

R. Hartmann (2001), Adaptive FE-methods for conservation equations, in *Eighth International Conference on Hyperbolic Problems: Theory, Numerics, Applications* (HYP2000) (G. Warnecke and H. Freistühler, eds), Birkhäuser, Basel.

R. Hartmann and P. Houston (2001), 'Adaptive discontinuous Galerkin finite element methods for nonlinear hyperbolic conservation laws'. Submitted for publication.

P. Houston and E. Süli (2001*a*), *hp*-adaptive discontinuous Galerkin finite element methods for hyperbolic problems, Numerical Analysis Group Research Report NA-01/05, University of Oxford. *SIAM J. Sci. Comput.* **23**, 1225–1251.

P. Houston and E. Süli (2001*b*), 'Stabilized *hp*-finite element approximation of partial differential equations with non-negative characteristic form', *Computing* **66**, 99–119.

P. Houston and E. Süli (2002*a*), 'Sobolev regularity estimation for *hp*-adaptive finite element methods', in *Proceedings of the ENUMATH 2001 Conference* (F. Brezzi, ed.), Springer. To appear.

P. Houston and E. Süli (2002*b*), 'Adaptive finite element approximation of hyperbolic problems', *NASA Ames/VK1 Lecture Series on Error Estimation and Solution Adaptive Discretization in CFD* (T. J. Barth and H. Deconinck, eds), *Lecture Notes in Computational Science and Engineering*, Springer. To appear.

P. Houston, R. Rannacher and E. Süli (2000*a*), '*A posteriori* error analysis for stabilized finite element approximations of transport problems', *Comput. Methods Appl. Mech. Engr.* **190**, 1483–1508.

P. Houston, C. Schwab and E. Süli (2000*b*), Discontinuous *hp*-finite element methods for advection–diffusion problems, Numerical Analysis Group Research Report NA-00/15, University of Oxford. To appear in *SIAM J. Numer. Anal.*

P. Houston, R. Hartmann and E. Süli (2001), Adaptive discontinuous Galerkin finite element methods for nonlinear hyperbolic problems, Numerical Analysis Group Research Report NA-01/06, University of Oxford. In *Numerical Methods for Fluid Dynamics VII* (M. Baines, ed.), ICFD, Oxford, pp. 347–353.

J. Jaffre, C. Johnson and A. Szepessy (1995), 'Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws', *Math. Mod. Methods Appl. Sci.* **5**, 367–386.

A. Jameson (1988), 'Aerodynamic design via control theory', *J. Sci. Comput.* **3**, 233–260.

A. Jameson (1995), Optimum aerodynamic design using control theory, in *Computational Fluid Dynamics Review 1995* (M. Hafez and K. Oshima, eds), Wiley, pp. 495–528.

A. Jameson (1999), 'Re-engineering the design process through computation', *J. Aircraft* **36**, 36–50.

A. Jameson, N. A. Pierce and L. Martinelli (1998), 'Optimum aerodynamic design using the Navier–Stokes equations', *J. Theoret. Comput. Fluid Mech.* **10**, 213–237.

B. Koren (1988), 'Defect correction and multigrid for an efficient and accurate computation of airfoil flows', *J. Comput. Phys.* **77**, 183–206.

M. G. Larson (1996), A new error analysis for finite element approximations of indefinite linear elliptic problems, Technical report, Department of Mathematics, Chalmers University of Technology, Sweden.

M. G. Larson and T. J. Barth (2000), *A posteriori* error estimation for adaptive discontinuous Galerkin approximations of hyperbolic systems, in *Discontinuous Galerkin Finite Element Methods* (B. Cockburn, G. E. Karniadakis and C.-W. Shu, eds), Vol. 11 of *Lecture Notes in Computational Science and Engineering*, Springer, Berlin/Heidelberg, pp. 363–368.

J.-L. Lions (1971), *Optimal Control of Systems Governed by Partial Differential Equations*, Springer. Translated by S. K. Mitter.

M. Melenk and C. Schwab (1999), 'An *hp* finite element method for convection-diffusion problems', *IMA J. Numer. Anal.* **19**, 425–453.

P. Monk and E. Süli (1998), 'The adaptive computation of far field patterns by *a posteriori* error estimates of linear functionals', *SIAM J. Numer. Anal.* **36**, 251–274.

J. C. Newman, A. C. Taylor, R. W. Barnwell, P. A. Newman and G. J.-W. Hou (1999), 'Overview of sensitivity analysis and shape optimization for complex aerodynamic configurations', *J. Aircraft* **36**, 87–96.

J. Nitsche (1968), 'Ein Kriterium für die Quasi-Optimalität des ritzschen Verfahrens', *Numer. Math.* **11**, 346–348.

J. T. Oden and S. Prudhomme (1999), 'On goal-oriented error estimation for elliptic problems: Application to control of pointwise errors', *Comput. Methods Appl. Mech. Engr.* **176**, 313–331.

J. T. Oden and J. N. Reddy (1983), *Variational Methods in Theoretical Mechanics*, 2nd edn, Springer, Berlin/Heidelberg.

L. A. Oganesjan and L. A. Ruhovec (1969), 'Investigation of the rate of convergence of variation-difference schemes for second order elliptic equations in two-dimensional region with smooth boundary', *Ž. Vyčisl. Mat. i Mat. Fiz.* **9**, 1102–1120.

M. Paraschivoiu, J. Peraire and A. T. Patera (1997), '*A posteriori* finite element bounds for linear functional outputs of elliptic partial differential equations', *Comput. Methods Appl. Mech. Engr.* **150**, 289–312.

J. Peraire and A. T. Patera (1997), Bounds for linear functional outputs of coercive partial differential equations: Local indicators and adaptive refinement, in *New Advances in Adaptive Computational Methods in Mechanics* (P. Ladeveze and J. T. Oden, eds), Elsevier.

N. A. Pierce and M. B. Giles (1998), Adjoint recovery of superconvergent functionals from approximate solutions of partial differential equations, Numerical Analysis Group Research Report NA-98/18, University of Oxford.

N. A. Pierce and M. B. Giles (2000), 'Adjoint recovery of superconvergent functionals from PDE approximations', *SIAM Review* **42**, 247–264.

O. Pironneau (1974), 'On optimum design in fluid mechanics', *J. Fluid Mech.* **64**, 97–110.

R. Rannacher (1998), Adaptive finite element methods, in *Proc. NATO Summer School on Error Control and Adaptivity in Scientific Computing*, Kluwer Academic, pp. 247–278.

D. Schötzau, C. Schwab and R. Stenberg (1998), 'Mixed *hp*-FEM on anisotropic meshes', *Math. Mod. Methods Appl. Sci.* **8**, 787–820.

D. Schötzau, C. Schwab and R. Stenberg (1999), 'Mixed *hp*-FEM on anisotropic meshes II: Hanging nodes and tensor products of boundary layer meshes', *Math. Mod. Methods Appl. Sci.* **4**, 667–697.

C. Schwab (1998), *p- and hp-Finite Element Methods: Theory and Applications to Solid and Fluid Mechanics*, Oxford University Press, Oxford.

T. Skalický and H. G. Roos (1999), 'Anisotropic mesh refinement for problems with internal and boundary layers', *Internat. J. Numer. Methods Engr.* **11**, 1933–1953.

R. D. Skeel (1981), 'A theoretical framework for proving accuracy results for deferred corrections', *SIAM J. Numer. Anal.* **19**, 171–196.

H. J. Stetter (1978), 'The defect correction principle and discretization methods', *Numer. Math.* **29**, 425–443.

G. Strang and G. J. Fix (1973), *An Analysis of the Finite Element Method*, Prentice-Hall.

E. Süli (1998), *A posteriori* error analysis and adaptivity for finite element approximations of hyperbolic problems, in *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws* (D. Kröner, M. Ohlberger and C. Rohde, eds), Vol. 5 of *Lecture Notes in Computational Science and Engineering*, Springer, Berlin/Heidelberg, pp. 123–194.

E. Süli, P. Houston and C. Schwab (1999), *hp*-finite element methods for hyperbolic problems, in *The Mathematics of Finite Elements and Applications X: MAFELAP 1999* (J. R. Whiteman, ed.), Elsevier, pp. 143–162.

E. Süli, P. Houston and B. Senior (2001), *hp*-Discontinuous Galerkin finite element methods for nonlinear hyperbolic problems, Numerical Analysis Group Research Report NA-01/07, University of Oxford. In *Numerical Methods for Fluid Dynamics VII* (M. Baines, ed.), ICFD, Oxford, pp. 73–86.

B. Szabó and I. Babuška (1991), *Finite Element Analysis*, Wiley, New York.

O. Talagrand and P. Courtier (1997), 'Variational assimilation of meteorological observations with the adjoint vorticity equation, Part 1: Theory', *Quart. J. Royal Met. Soc.* **113**, 1311–1328.

R. Verfürth (1996), *A Review of a posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Teubner, Stuttgart.

J. A. Wheeler (1973), 'Simulation of heat transfer from a warm pipeline buried in permafrost', in *Proceedings of the 74th National Meeting of the American Institute of Chemical Engineering.*