A new goal-oriented formulation of the finite element method

Kenan Kergrene, Serge Prudhomme, Ludovic Chamoin, Laforest Marc

A new goal-oriented formulation of the finite element method

Kenan Kergrene[a], Serge Prudhomme[a,*], Ludovic Chamoin[b], Marc Laforest[a]

[a] *Département de mathématiques et de génie industriel, École Polytechnique de Montréal, Montréal, Québec, Canada, H3T 1J4* [b] *LMT, ENS Cachan, CNRS, Université Paris-Saclay, 61 Avenue du Président Wilson, 94230 Cachan, France*

Abstract

In this paper, we introduce, analyze, and numerically illustrate a method for taking into account quantities of interest during the finite element treatment of a boundary-value problem. The objective is to derive a method whose computational cost is of the same order as that of the classical approach for goal-oriented adaptivity, which involves the solution of the primal problem and of an adjoint problem used to weigh the residual and provide indicators for mesh refinement. In the current approach, we first solve the adjoint problem, then use the adjoint information as a minimization constraint for the primal problem. As a result, the constrained finite element solution is enhanced with respect to the quantities of interest, while maintaining near-optimality in energy norm. We describe the formulation in the case of a problem defined by a symmetric continuous coercive bilinear form and demonstrate the efficiency of the new approach on several numerical examples.

Keywords: Finite element method; Goal-oriented formulation; Mixed formulation; Error estimation; Adaptive mesh refinement; Multi-objective goal functionals

1. Introduction

Advances in Computational Science and Engineering have reached such a level of maturity that increasingly complex multiphysics and multiscale problems can now be simulated for decision-making and optimal design. The focus of such simulations has thus shifted towards efficiently and accurately predicting specific features of the solution rather than the whole solution itself. With that objective in mind, goal-oriented error estimation and adaptive methods [1,2], whose predominant instance is the dual-weighted residual method [3], have been developed since the late nineties in order to estimate and control errors with respect to quantities of interest. The principle of these methods essentially relies on the solution of adjoint problems associated with quantities of interest in order to identify and refine the sources of discretization or modeling errors that influence the most these quantities [4]. These approaches have been very successful so far in accelerating the convergence of the approximations towards the exact quantities of interest and thus at a lesser computational cost than classical a posteriori error estimation methods. However, dual-weighted residual methods are reminiscent of two-step predictor–corrector methods, in the sense that one first

* Corresponding author.
E-mail address: serge.prudhomme@polymtl.ca (S. Prudhomme).

computes an approximate solution of a boundary-value problem and then corrects the discrete solution space in order to better approximate the quantities of interest.

The objective of the paper is to propose an alternative paradigm: we aim at developing a novel finite element formulation of boundary-value problems whose approximate solutions are tailored towards the calculation of quantities of interest. The main idea is based on the reformulation of the problem as a minimization problem subjected to the additional constraint that the error in the quantities of interest be within some prescribed tolerance. Chaudhry et al. [5] have proposed a similar approach in which constraints are enforced via a penalization method. One main issue with that approach is concerned with the selection of suitable penalization parameters. We propose here to circumvent this issue by imposing the equality or inequality constraints through the use of Lagrange multipliers. The framework will be presented in the case of several quantities of interest in order to describe the method in a general setting. However, the treatment of several quantities of interest is not the primary goal of the paper and the reader interested in multi-objective error estimation is referred to the following literature [6–9].

The present paper is organized as follows: In Section 2, we present the model problem considered in this study and introduce some classical notations. In Section 3, we formulate the novel formulation of taking into account quantities of interest using a constrained minimization. We demonstrate the well-posedness of the formulation and the near-optimality of the corresponding solution. In Section 4, we investigate the case of inequality constraints using the KKT conditions (Karush–Kuhn–Tucker). Section 5 addresses the topic of error estimation and adaptivity. Numerical examples are presented in Section 6 and illustrate the performance of the method. In particular, we compare our approach to the classical goal-oriented adaptivity. We finally provide some concluding remarks in Section 7.

2. Preliminaries and model problem

Consider an abstract problem written in weak form as follows:

$$\text{Find } u \in V \text{ such that } a(u, v) = f(v), \quad \forall v \in V, \tag{2.1}$$

where $(V, \|\cdot\|)$ is a Hilbert space, and bilinear form a and linear form f satisfy the usual regularity assumptions: a is continuous and coercive and f is continuous over V. This problem will be referred to as the *primal problem* and its well-posedness is ensured by the Lax–Milgram theorem. For the sake of simplicity, we require in addition that a be symmetric so that the primal problem (2.1) is equivalent to minimizing the following energy functional

$$J(u) = \frac{1}{2}a(u, u) - f(u), \tag{2.2}$$

i.e.

$$\text{Find } u \in V \text{ such that } u = \underset{v \in V}{\operatorname{argmin}}\, J(v). \tag{2.3}$$

If a were not symmetric, the method presented in this work could be applied by considering a Least Squares approach [10], which in effect symmetrizes the problem.

We now turn to the finite element formulation of the primal problem (2.1). Here, and in the remainder of the paper, we consider a general conforming finite element space $V_h = \operatorname{span}\{\varphi_i\} \subset V$, where φ_i, $i = 1, \ldots, N$ are basis functions of V_h. We also assume that the corresponding mesh satisfies the usual regularity properties [11]. We denote by h the characteristic mesh size. The classical finite element problem associated to the primal problem (2.1) is given by

$$\text{Find } u_h \in V_h \text{ such that } a(u_h, v_h) = f(v_h), \quad \forall v_h \in V_h. \tag{2.4}$$

The objective of this work is to improve the accuracy in the approximation of scalar quantities of the solution u of the primal problem (2.1). Consider therefore the continuous quantities of interest $Q_i(u)$, $i = 1, \ldots, k$, with $k \in \mathbb{N}$ and assume these are linear, i.e. $Q_i \in V'$, the dual space of V. We will denote by Q the linear map from V to $\mathbb{R}^k$ whose ith component is Q_i. Further, we assume the linear forms to be linearly independent, i.e. the map Q is surjective. In other words, each functional Q_i provides independent information about u. These linear forms are associated with the k dual or adjoint problems

$$\text{For } i = 1, \ldots, k, \text{ find } p_i \in V \text{ such that } a(v, p_i) = Q_i(v), \quad \forall v \in V, \tag{2.5}$$

which allow one to derive the fundamental relations:

$$Q_i(u) = a(u, p_i) = f(p_i), \quad \forall i = 1, \ldots, k. \tag{2.6}$$

The finite element formulation of the adjoint problems (2.5) in space V_h are given by

$$\text{For } i = 1, \ldots, k, \text{ find } p_{i,h} \in V_h \text{ such that } a(v_h, p_{i,h}) = Q_i(v_h), \quad \forall v_h \in V_h. \tag{2.7}$$

The main idea is to derive a novel formulation of the problem based on the minimization of the energy functional J subjected to constraints in terms of the quantities of interest Q.

3. Goal-oriented formulation with equality constraints

3.1. Formulation and well-posedness

Suppose for a moment that we are interested in finding a solution $w \in V$ that satisfies the constraints $Q_i(w) = \alpha_i$, where $\alpha = (\alpha_1, \ldots, \alpha_k)^T \in \mathbb{R}^k$ is given. Instead of the minimization problem (2.3), we consider the constrained minimization problem

$$\text{Find } w \in V \text{ such that } w = \underset{\substack{v \in V \\ Q(v) = \alpha}}{\operatorname{argmin}} J(v). \tag{3.1}$$

The standard way to impose constraints is by the introduction of the Lagrangian functional $\mathcal{L} : V \times \mathbb{R}^k \to \mathbb{R}$ defined as

$$\mathcal{L}(w, \lambda) = J(w) + \sum_{i=1}^{k} \lambda_i (Q_i(w) - \alpha_i), \tag{3.2}$$

where $\lambda = (\lambda_1, \ldots, \lambda_k)^T \in \mathbb{R}^k$ is the vector collecting the so-called Lagrange multipliers. This functional can be written in compact form as

$$\mathcal{L}(w, \lambda) = J(w) + \lambda \cdot (Q(w) - \alpha). \tag{3.3}$$

The saddle-point formulation of the Lagrangian functional $\mathcal{L}$ over $V \times \mathbb{R}^k$ yields the mixed problem

$$\text{Find } (w, \lambda) \in V \times \mathbb{R}^k \text{ such that } \begin{cases} a(w, v) + \lambda \cdot Q(v) = f(v), & \forall v \in V, \\ \tau \cdot Q(w) = \tau \cdot \alpha, & \forall \tau \in \mathbb{R}^k. \end{cases} \tag{3.4}$$

Introducing the bilinear form $b(\tau, v) = \tau \cdot Q(v)$ defined on $\mathbb{R}^k \times V$, above problem can be recast in the classical form

$$\text{Find } (w, \lambda) \in V \times \mathbb{R}^k \text{ such that } \begin{cases} a(w, v) + b(\lambda, v) = f(v), & \forall v \in V, \\ b(\tau, w) = \tau \cdot \alpha, & \forall \tau \in \mathbb{R}^k. \end{cases} \tag{3.5}$$

Lemma 1 (*LBB Condition*)**.** *Let $\|\cdot\|_1$ be the 1-norm on $\mathbb{R}^k$, i.e. $\|\tau\|_1 = \sum_i |\tau_i|$. The bilinear form b satisfies the LBB condition*

$$\exists\, \beta > 0 \text{ such that } \forall \tau \in \mathbb{R}^k, \quad \sup_{v \in V} \frac{|b(\tau, v)|}{\|v\|} \geq \beta \|\tau\|_1. \tag{3.6}$$

Proof. We first consider the trivial case $k = 1$ and then the general case.

Case $k = 1$. Let $z \in V \setminus \operatorname{Ker} Q$ (there exists such a z since the linear form Q is assumed to be surjective, i.e. non-zero in this case) and define $\beta = \frac{|Q(z)|}{\|z\|}$. Then for any $\tau \in \mathbb{R}$, $\frac{|b(\tau, z)|}{\|z\|} = \beta\,|\tau|$, so that $\sup_{v \in V} \frac{|b(\tau, v)|}{\|v\|} \geq \beta\,|\tau|$.

General case. Similarly, using the surjectivity of Q, one can find functions in V such that all cases in terms of the signs of the components of $\tau \in \mathbb{R}^k$ will be accounted for. More specifically, let the "vector-valued sign function" defined over $\mathbb{R}^k$ as

$$\begin{aligned} \text{signs} : \ & \mathbb{R}^k \to \{-1, 0, 1\}^k \\ & \tau \mapsto (\operatorname{sign}(\tau_1), \ldots, \operatorname{sign}(\tau_k)) \end{aligned} \tag{3.7}$$

Function signs is surjective onto the set $\{-1, 1\}^k$. Indeed, for any $\sigma \in \{-1, 1\}^k$, it holds $\text{signs}(\sigma) = \sigma$. Since Q is also surjective onto $\mathbb{R}^k$, for any $\sigma \in \{-1, 1\}^k$ there exists $z_\sigma \in V$ such that $\text{signs}(Q(z_\sigma)) = \sigma$. The set $\{-1, 1\}^k$ is finite, as a result this process constructs a finite set $\mathcal{Z} \subset V$. Then we define

$$\beta = \min_{\substack{z_\sigma \in \mathcal{Z} \\ i=1,\dots,k}} \frac{|Q_i(z_\sigma)|}{\|z_\sigma\|} > 0. \tag{3.8}$$

Now for any $\tau \in \mathbb{R}^k$, let $\sigma = \text{signs}(\tau)$. To determine $z \in \mathcal{Z}$ associated with σ when σ contains components of value zero, we construct $\tilde{\sigma}$ where all zero components have been replaced by one and set $z_\sigma = z_{\tilde{\sigma}}$. As a result there always exists a well-defined $z_\sigma \in \mathcal{Z}$ and it holds

$$\frac{|b(\tau, z_\sigma)|}{\|z_\sigma\|} = \frac{\left|\sum_{i=1}^k \tau_i Q_i(z_\sigma)\right|}{\|z_\sigma\|} = \frac{\sum_{i=1}^k |\tau_i|\,|Q_i(z_\sigma)|}{\|z_\sigma\|} \geq \beta \|\tau\|_1. \tag{3.9}$$

Consequently, $\sup_{v \in V} \frac{|b(\tau,v)|}{\|v\|} \geq \beta \|\tau\|_1$. □

Theorem 1. *The constrained problem* (3.5) *has a unique solution.*

Proof. The proof directly follows from the LBB condition established in Lemma 1 and the Babuška–Lax–Milgram theorem [12–14]. □

In the specific case where $\alpha = Q(u)$, u being the solution of the original problem (2.1), the solution of the constrained problem (3.5) is given by $w = u$ and $\lambda = 0$.

We now establish a key relation between the solutions to the constrained and unconstrained problems. A similar relation will be established in Theorem 3 in the case of the solutions to the constrained and unconstrained finite element problems.

Theorem 2. *Let $(w, \lambda) \in V \times \mathbb{R}^k$ denote the solution of the constrained problem* (3.5)*, $p_i \in V$ denote the solutions of the dual problems* (2.5)*, and $u \in V$ denote the solution of the unconstrained problem* (2.1)*. Then*

$$u = w + \sum_{i=1}^k \lambda_i p_i. \tag{3.10}$$

Proof. Using the adjoint problems (2.5) and the bilinearity of a, it holds

$$\lambda \cdot Q(v) = \sum_{i=1}^k \lambda_i Q_i(v) = \sum_{i=1}^k \lambda_i a(v, p_i) = a\left(v, \sum_{i=1}^k \lambda_i p_i\right). \tag{3.11}$$

Substituting the new expression (3.11) for $\lambda \cdot Q(v)$ in the first equation of the constrained problem (3.5) yields

$$a(w, v) + a\left(v, \sum_{i=1}^k \lambda_i p_i\right) = f(v), \quad \forall v \in V. \tag{3.12}$$

Now, making use of the fact that a is bilinear and symmetric yields

$$a\left(w + \sum_{i=1}^k \lambda_i p_i, v\right) = f(v), \quad \forall v \in V. \tag{3.13}$$

Finally, the Lax–Milgram theorem applied to the unconstrained problem (2.1) ensures unicity of the solution so that

$$u = w + \sum_{i=1}^k \lambda_i p_i, \tag{3.14}$$

which completes the proof. □

We further note that Theorem 2 holds for any choice of $\alpha \in \mathbb{R}^k$.

The mixed finite element problem on $V_h \times \mathbb{R}^k$ corresponding to the Lagrangian approach (3.4) is given by

$$\text{Find } (w_h, \lambda_h) \in V_h \times \mathbb{R}^k \text{ such that } \begin{cases} a(w_h, v_h) + \lambda_h \cdot Q(v_h) = f(v_h), & \forall v_h \in V_h, \\ \tau_h \cdot Q(w_h) = \tau_h \cdot \alpha, & \forall \tau_h \in \mathbb{R}^k. \end{cases} \tag{3.15}$$

Note that the Lagrange multiplier is here denoted by λ_h, not because of the discretization of $\mathbb{R}^k$, but rather because it depends on $w_h \in V_h$ where V_h is a finite-dimensional subspace of V. Furthermore, we also use this notation in order to avoid confusion with the Lagrange multiplier λ appearing in the constrained problem (3.4).

Remark 1. If $Q : V_h \to \mathbb{R}^k$ is still surjective, existence and unicity are inherited from the infinite dimensional case; in particular, surjectivity implies here that: (1) $\dim \mathbb{R}^k \leq \dim V_h$, i.e. $k \leq N$: there are fewer constraints than degrees of freedom; (2) $\dim \operatorname{Im} Q = k$, i.e. the rows of the $k \times N$ constraint matrix are linearly independent: in other words it has full row-rank.

Similarly to Theorem 2, we can establish the following relation between the solutions to the constrained and unconstrained finite element problems. This result will be used when studying convergence in Section 3.2 and adaptivity in Section 5.

Theorem 3. *Let $(w_h, \lambda_h) \in V_h \times \mathbb{R}^k$ denote the solution of the constrained problem (3.15), $p_{i,h} \in V_h$ denote the solutions of the dual problems (2.7) and $u_h \in V_h$ denote the solution of the unconstrained problem (2.4). Then*

$$u_h = w_h + \sum_{i=1}^{k} \lambda_{h,i}\, p_{i,h}. \tag{3.16}$$

Proof. The proof is similar to that of Theorem 2. □

Remark 2. Note that the Galerkin orthogonality arising from the constrained problem (3.15) is slightly modified compared to the classical unconstrained approach. Indeed, subtracting the first equation of the constrained finite element problem (3.15) from the initial weak formulation (2.1) yields

$$a(u - w_h, v_h) - b(\lambda_h, v_h) = f(v_h) - f(v_h) = 0, \quad \forall v_h \in V_h, \tag{3.17}$$

that is $a(u - w_h, v_h) = b(\lambda_h, v_h) = \lambda_h \cdot Q(v_h),\ \forall v_h \in V_h$. In particular, $u - w_h$ is not orthogonal to the entire space V_h but at least to $V_h \cap \operatorname{Ker} Q$. This modified Galerkin orthogonality relation will be used when studying error estimation and adaptivity in Section 5.

Remark 3. In contrast with the Lagrangian approach, the penalization approach [5] seeks the minimizer of the modified energy functional

$$J_\beta(u) = J(u) + \sum_{i=1}^{k} \frac{\beta_i}{2} (Q_i(u) - \alpha_i)^2, \tag{3.18}$$

with a penalization parameter $\beta \in \mathbb{R}^k$ chosen to ensure convergence, efficiency, and accuracy. In that case, the relation between the penalized solution u_β and the unconstrained solution u is

$$u = u_\beta + \sum_{i=1}^{k} \beta_i \left(Q_i(u_\beta) - \alpha_i\right) p_i, \tag{3.19}$$

and similarly for their finite dimensional counterparts.

3.2. Selection of α and a priori convergence rate

In this work, the goal is to obtain an approximation w_h such that $Q(w_h) \approx Q(u)$, meaning that the target values α_i should be as close as possible to the quantities of interest $Q_i(u)$. In view of the fundamental relation (2.6), we propose to choose α by considering the k adjoint problems (2.5). However, for most problems of practical interest, the adjoint

problems cannot be solved exactly and have to be discretized, say using the finite element method. These approximate adjoint solutions $\tilde{p}_i$ are then used to derive the target values α_i, i.e. we set $\alpha_i = f(\tilde{p}_i)$, $i = 1, \ldots, k$ and then proceed to solve the constrained finite element problem (3.15).

Remark 4. We emphasize here that one needs to use a space larger than V_h to compute the adjoint finite element solutions $\tilde{p}_i$. Indeed let us assume that we were to solve each discrete adjoint problem in the same space V_h as the one used to solve the classical finite element problem (2.4), i.e. the adjoint problems (2.7), and then set $\alpha_i = f(p_{i,h})$. Choosing these target values α as constraints for the constrained primal problem (3.15) leads to $Q_i(w_h) = \alpha_i$.

Repeating the computation (2.6), now in the finite element space V_h, for $i = 1, \ldots, k$ we find

$$Q_i(u_h) = a(u_h, p_{i,h}) = f(p_{i,h}) = Q_i(w_h), \tag{3.20}$$

that is, the same approximation of the quantities of interest is found whether we proceed to a constrained minimization or not: the approach would thus be useless. Indeed, the unique solution (w_h, λ_h) of the constrained problem (3.15) would be given by $w_h = u_h$, the solution of the unconstrained problem (2.4), and $\lambda_h = 0$.

In the remainder of the paper, we shall use a larger finite element space for the adjoint problems, denoted by $\tilde{V}_h$, than the approximation space $V_h \subset \tilde{V}_h$ for the primal problem. In practice, $\tilde{V}_h$ consists of higher-order hierarchical elements on the same mesh. Consequently, in order to compute the target values α, we consider the following higher-order dual problems

$$\text{Find } \tilde{p}_i \in \tilde{V}_h \text{ such that } a(\tilde{v}, \tilde{p}_i) = Q_i(\tilde{v}), \quad \forall \tilde{v} \in \tilde{V}_h, \quad \forall i = 1, \ldots, k, \tag{3.21}$$

and set $\alpha_i = f(\tilde{p}_i)$ for all $i = 1, \ldots, k$. In later analysis, we will nonetheless also consider the dual problems in the same space V_h defined by (2.7).

We now turn our attention to the numerical method that will be used to solve the finite-dimensional mixed problem described above. The mixed formulation (3.15) yields the following system of equations

$$\begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} W \\ \lambda_h \end{bmatrix} = \begin{bmatrix} F \\ \alpha \end{bmatrix}, \tag{3.22}$$

with $A_{ij} = a(\varphi_j, \varphi_i)$, $B_{ij} = b(e_j, \varphi_i) = Q_j(\varphi_i)$, where $\left\{e_j\right\}_{j=1}^{k}$ denotes the canonical basis of $\mathbb{R}^k$, $F_i = f(\varphi_i)$, and the components w_i of W are the coefficients of the solution w_h with respect to the finite element basis functions φ_i, i.e. $w_h = \sum_i w_i \varphi_i$. This system could be solved directly as given since the augmented matrix is non-singular. However, its size is larger than that yielded by the classical unconstrained finite element method (2.4), which is simply $AU = F$. If the number of constraints is large, one may consider applying either the Uzawa or Augmented Lagrangian method [15].

Applying each of the k functionals $Q_1, \ldots, Q_k$ to the relation (3.16) and rearranging the terms, we obtain a linear system of size $k \times k$ with vector $\lambda_h \in \mathbb{R}^k$ as unknown

$$S_h \lambda_h = Q(u_h - w_h) = Q(u_h) - \alpha, \tag{3.23}$$

with $S_{h,ij} = Q_i(p_{j,h})$ and where we highlighted the dependency of this matrix on the mesh size h. As a result, instead of solving the augmented and possibly ill-conditioned linear system (3.22), one only needs to compute

1. the unconstrained solution u_h of (2.4),
2. the adjoint solutions $p_{i,h}$ of (2.7),
3. the Lagrange multipliers λ_h using (3.23), then
4. form the constrained solution w_h using (3.16).

In other words, one solves $k+1$ (1 primal and k dual) systems of the same size as the original finite element problem, as well as one $k \times k$ linear algebraic system. In fact, S_h is precisely the Schur complement arising from the augmented matrix featured in (3.22), usually defined as $B^T A^{-1} B$. Indeed, each vector counterpart to the adjoint solution $p_{j,h}$ is given by $A^{-1} B_j$, where B_j denotes the jth column of B. When concatenating these k column vectors, we form $A^{-1} B$. Applying now each of the k functionals Q_i, we get $B^T A^{-1} B$. Since B is injective and A^{-1} is symmetric positive-definite, S_h is symmetric positive-definite and thus invertible, so that the Schur complement equation (3.23) has a unique solution. We now provide the following result about the Schur complements S_h.

Lemma 2. *The Schur complements S_h converge towards a symmetric positive-definite matrix S as the mesh size h tends to zero.*

Proof. The entries of S_h are given by $S_{h,ij} = Q_i(p_{j,h})$. By continuity of Q_i, the matrices S_h converge to the matrix S defined by $S_{ij} = Q_i(p_j)$. Since the S_h are all symmetric and positive semi-definite, by closedness of $\mathcal{S}_k^+$ in $\mathcal{M}_k$, S is symmetric and positive semi-definite as well. To prove definiteness, we note that $S_{ij} = Q_i(p_j) = a(p_i, p_j)$. It follows that S is the Gram matrix of the family $\{p_i \in V, i = 1, \dots, k\}$ for the inner product $a(\cdot, \cdot)$. Since the linear forms $Q_i \in V'$ are assumed to be linearly independent, the adjoint solutions $p_i \in V$ are also linearly independent, as a result S is positive-definite. □

We now establish a theorem stating that the proposed approach yields the same rate of convergence as that of the classical approach.

Theorem 4. *Let $\|\cdot\|_{\mathcal{E}}$ denote the energy norm on V, i.e. the norm induced by the bilinear form a, and $\|\cdot\|_1$ denote the 1-norm on $\mathbb{R}^k$. Let $u \in V$ denote the solution of the primal problem (2.1), $w_h \in V_h$ the solution of the constrained problem (3.15), and $u_h \in V_h$ the solution of the unconstrained problem (2.4). Assume there exists $C > 0$, independent of the mesh size h, such that*

$$\|Q(u) - Q(w_h)\|_1 \leq C\|Q(u) - Q(u_h)\|_1. \tag{3.24}$$

Then there exists $D > 0$, independent of the mesh size h, such that

$$\|u - w_h\|_{\mathcal{E}} \leq D\|u - u_h\|_{\mathcal{E}}. \tag{3.25}$$

Proof. Using Theorem 3, it holds

$$\begin{aligned}
\|u - w_h\|_{\mathcal{E}} &\leq \|u - u_h\|_{\mathcal{E}} + \|u_h - w_h\|_{\mathcal{E}}, \\
&\leq \|u - u_h\|_{\mathcal{E}} + \left\| \sum_{i=1}^{k} \lambda_{h,i} p_{i,h} \right\|_{\mathcal{E}}, \\
&\leq \|u - u_h\|_{\mathcal{E}} + \sum_{i=1}^{k} \left|\lambda_{h,i}\right| \|p_{i,h}\|_{\mathcal{E}}, \\
&\leq \|u - u_h\|_{\mathcal{E}} + C_1\|\lambda_h\|_1,
\end{aligned} \tag{3.26}$$

where $C_1 = \max_{i=1,\dots,k}\|p_i\|_{\mathcal{E}} \geq \max_{i=1,\dots,k}\|p_{i,h}\|_{\mathcal{E}}$ is independent of the mesh size h. Now using (3.23) and the fact that the Schur complement S_h is non-singular, it follows

$$\lambda_h = S_h^{-1} Q(u_h - w_h), \tag{3.27}$$

from which we obtain

$$\|\lambda_h\|_1 = \|S_h^{-1} Q(u_h - w_h)\|_1 \leq C_{S_h^{-1}}\|Q(u_h - w_h)\|_1, \tag{3.28}$$

where $C_{S_h^{-1}}$ denotes the matrix norm of S_h^{-1} induced by $\|\cdot\|_1$, which is not independent of the mesh size h. To obtain a uniform bound, we use Lemma 2 and continuity of the matrix norm, so that the sequence $C_{S_h^{-1}}$ converges to $C_{S^{-1}}$, the matrix norm of S^{-1}. As a convergent sequence, it is bounded so there exists $\gamma \geq C_{S_h^{-1}}$, with γ independent of the mesh size h, such that

$$\|\lambda_h\|_1 \leq \gamma\|Q(u_h - w_h)\|_1. \tag{3.29}$$

Then, using assumption (3.24),

$$\begin{aligned}
\|Q(u_h - w_h)\|_1 &\leq \|Q(u - w_h)\|_1 + \|Q(u - u_h)\|_1, \\
&\leq (1 + C)\|Q(u - u_h)\|_1.
\end{aligned} \tag{3.30}$$

Using now the boundedness of Q, it holds

$$\|Q(u - u_h)\|_1 \leq C_Q\|u - u_h\|_{\mathcal{E}}, \tag{3.31}$$

where C_Q denotes the operator norm of Q induced by $\|\cdot\|_{\mathcal{E}}$ and $\|\cdot\|_1$. Finally, we obtain

$$\|u - w_h\|_{\mathcal{E}} \leq D\|u - u_h\|_{\mathcal{E}}, \tag{3.32}$$

where $D = 1 + C_1(1 + C)\gamma C_Q$ is independent of the mesh size h. $\square$

Essentially, Theorem 4 states that if the target values $\alpha \in \mathbb{R}^k$ are consistent with the problem, then the constrained solution w_h maintains near-optimality in the energy norm. In particular, we also demonstrated that the vector of the Lagrange multipliers $\lambda_h \in \mathbb{R}^k$ necessarily converges to zero as h tends to zero.

4. Inequality constraints

In this section, we focus on a slightly less restrictive approach where the equality constraints $Q(w_h) = \alpha$ are replaced by the following inequality constraints

$$|Q_i(w_h) - \alpha_i| \leq \varepsilon_i, \quad \forall i = 1, \ldots, k, \tag{4.1}$$

where each ε_i is a positive scalar, possibly quite small. The rationale for replacing equality constraints by inequality constraints is twofold. First, since we consider the computable approximates α using the discretized adjoint problems (3.21) instead of the exact quantities $Q(u)$, we introduce some error in the target values. As a result, there is no need to exactly impose those perturbed values. The quantities $\varepsilon \in \mathbb{R}^k$ could be user-specified or could represent tolerances on the errors in the quantities of interest. Second, recall that our objective is to derive a finite element formulation that adequately represents the solution globally as well as quantities of interest of that solution. Intuitively speaking, incorporating equality constraints comes down to sacrificing the energy in order to satisfy the constraints (minimization in an affine space strictly contained in the "surrounding" space). Replacing equality constraints by inequality constraints would allow one to reduce the impact of this sacrifice and better represent the solution globally while maintaining a controlled (through parameters ε) representation of the quantities of interest. In order to reduce the burden of the notation, we introduce

$$\alpha^- = \alpha - \varepsilon \text{ and } \alpha^+ = \alpha + \varepsilon. \tag{4.2}$$

The necessary conditions for the solution of an inequality constrained minimization problem are given by the KKT (Karush–Kuhn–Tucker) conditions [16,17], leading here to

$$\text{Find } (w_h, \lambda_h^+, \lambda_h^-) \in V_h \times \mathbb{R}^k \times \mathbb{R}^k \text{ such that}$$
$$\begin{cases} a(w_h, v_h) + \lambda_h^+ \cdot Q(v_h) + \lambda_h^- \cdot Q(v_h) = f(v_h), & \forall v_h \in V_h, \\ \lambda_h^+ \geq 0,\ \lambda_h^- \leq 0, \\ \alpha^+ \geq Q(w_h),\ \alpha^- \leq Q(w_h), \\ \lambda_{h,i}^+ \left(\alpha_i^+ - Q_i(w_h)\right) = 0,\ \lambda_{h,i}^- \left(Q_i(w_h) - \alpha_i^-\right) = 0, & \forall i = 1, \ldots, k, \end{cases} \tag{4.3}$$

where the notation $\tau \geq 0$ (resp. $\tau \leq 0$) for a vector $\tau \in \mathbb{R}^k$ is employed to mean that all components of the vector are positive (resp. negative). The last conditions of the inequality constrained system (4.3) are usually called the "complementary conditions" and essentially state that for each of the k constraints there are three possibilities:

1. $\lambda_{h,i}^+ = \lambda_{h,i}^- = 0$ and $\alpha_i^- \leq Q_i(w_h) \leq \alpha_i^+$,
2. $\lambda_{h,i}^+ = 0$, $\lambda_{h,i}^- < 0$ and $Q_i(w_h) = \alpha_i^-$,
3. $\lambda_{h,i}^- = 0$, $\lambda_{h,i}^+ > 0$ and $Q_i(w_h) = \alpha_i^+$.

In the first case, the ith constraint is usually referred to as "non-binding", in the sense that it is naturally satisfied by the unconstrained minimization and does not have to be enforced (observe in this case that the ith component of the Lagrange multipliers vanishes from the weak formulation — first equation of problem (4.3)); in the other cases, it is said "binding" and equality has to be enforced on the boundary of the admissible set: either $Q_i(w_h) = \alpha_i^+$ or $Q_i(w_h) = \alpha_i^-$. As a result, in an inequality constrained minimization, each constraint is either enforced with equality or discarded. The main difficulty in such problems rests on the determination of the set of active constraints. One could use a brute force approach and solve all 3^k problems, but there exist more efficient approaches. We mention for instance the existence of e.g. interior point or barrier methods as well as the IPOPT package [18,19], which can be used to solve the inequality constrained system (4.3) at the expense of an iterative scheme.

In the present study, the task of finding the set of active constraints is somewhat simplified compared to a general problem. We describe the rationale for solving the inequality constrained system (4.3) in the case where there is only one constraint, i.e. $k = 1$. We start by finding $u_h \in V_h$, the solution to the unconstrained problem (2.4). Equivalently this could be viewed as assuming that the set of active constraints is empty, i.e. all Lagrange multipliers are zero. Next we form the quantity of interest $Q(u_h) \in \mathbb{R}$ and determine whether we are in case (i) $\alpha^- \leq Q(u_h) \leq \alpha^+$; or (ii) $Q(u_h) < \alpha^-$; or (iii) $\alpha^+ < Q(u_h)$. If we are in the first case then the constraint is non-binding, i.e. the solution $(w_h, \lambda_h^+, \lambda_h^-)$ of the inequality constrained problem (4.3) is given by $(u_h, 0, 0)$, which is the unconstrained solution; if we are in the second case then the lower bound is binding, i.e. the solution is of the form $(w_h, 0, \lambda_h)$ with $\lambda_h < 0$; and if we are in the third case then the upper bound is binding, i.e. the solution is of the form $(w_h, \lambda_h, 0)$ with $\lambda_h > 0$. The reason follows from the fact that the Schur complement S_h is positive-definite, i.e. $S_h > 0$ since $k = 1$, and as a result equation (3.23) implies that the Lagrange multiplier and the quantity on the right-hand side have same sign. Unfortunately, the reasoning does not extend to more than one constraint ($k > 1$) since in that case being positive-definite does not yield enough information on the coefficients of the Schur complement S_h.

An alternative approach could be to exploit the Schur complement equation (3.23) and solve each of the 3^k problems of size less or equal to $k \times k$, each associated to a different set of active constraints. For each resulting vector of Lagrange multipliers, a first check is whether the KKT conditions relative to their signs are respected: a Lagrange multiplier associated to an upper (resp. lower) bound should be positive (resp. negative). For each of the remaining potential solutions, one should solve the primal unknowns and check whether the inequalities $\alpha^- \leq Q(w_h) \leq \alpha^+$ hold. Since the minimization problem is convex, the KKT conditions are necessary and sufficient so that in practice not all 3^k problems need to be solved.

We will not show numerical examples using inequality constraints as they do not bring new insight when compared to the results with equality constraints.

5. Error estimation and adaptivity

In this section, we derive error estimates and design an adaptive strategy for the numerical approach considered in this study. We will use what could be coined a "global implicit method". We note that implicit methods usually introduce auxiliary residual problems defined on patches of elements or single elements notably to spare computational effort [20]. However such a paradigm is not the primary focus of this paper, so that we consider a global method instead.

In the current approach, local contributions to the error are derived on elements, and the elements with the largest contributions are marked for refinement. For the sake of clarity, we first describe the method for error estimation in energy norm and with respect to quantities of interest in the case of the classical solution u_h of problem (2.4). We will then turn to error estimation for the solution w_h of the constrained problem (3.15).

Recall that the energy norm is defined on V by $\|v\|_{\mathcal{E}} = a(v, v)^{1/2}$. Let us introduce the residual functional R^h, defined with respect to u_h, the classical unconstrained finite element solution of (2.4)

$$R^h(u_h; v) = f(v) - a(u_h, v) = a(u - u_h, v), \quad \forall v \in V. \tag{5.1}$$

Thanks to the classical Galerkin orthogonality, we have that $R^h(u_h; v_h) = 0$ for any $v_h \in V_h$. As a result, $R^h(u_h; v) = R^h(u_h; v - v_h)$ for all $v \in V$ and $v_h \in V_h$. This residual is actually used for both the error estimation in the energy norm as well as in the quantities of interest. Indeed, the error in the energy norm is defined by

$$\mathcal{E}_h = \|u - u_h\|_{\mathcal{E}} = \sqrt{a(u - u_h, u - u_h)} = \sqrt{R^h(u_h; u - u_h)}. \tag{5.2}$$

Similarly, the error in each quantity of interest Q_i is defined by

$$\mathcal{E}_i = |Q_i(u) - Q_i(u_h)| = |f(p_i) - a(u_h, p_i)| = \left|R^h(u_h; p_i)\right| = \left|R^h(u_h; p_i - p_{i,h})\right|. \tag{5.3}$$

For any $v \in V$, the scalar quantity $R^h(u_h; v)$ can be decomposed into local contributions. In order to illustrate this process, consider the following example for bilinear form a and linear form f

$$a(u, v) = \int_\Omega a \nabla u \cdot \nabla v \, dx, \text{ and } f(v) = \int_\Omega f v \, dx + \int_{\Gamma_N} g v \, ds, \tag{5.4}$$

where $\Gamma_N \subset \partial\Omega$ denotes the Neumann part of the boundary of domain Ω. Then, we have

$$R^h(u_h; v) = f(v) - a(u_h, v) = \int_\Omega f v \, dx + \int_{\Gamma_N} g v \, ds - \int_\Omega a \nabla u_h \cdot \nabla v \, dx. \tag{5.5}$$

After splitting the integrals from Ω to each element $K \subset \Omega$, we can use Green's first identity and rearrange the integrals over the boundary of each element as integrals over the set of mesh edges Γ

$$R^h(u_h; v) = \sum_{K \subset \Omega} \int_K r_K v \, dx - \sum_{\gamma \subset \Gamma} \int_\gamma j_\gamma v \, ds, \tag{5.6}$$

where we have introduced the interior element residual term $r_K = (\nabla \cdot (a \nabla u_h) + f)\,|_K$ and the edge element residual term j_γ defined on Γ by

$$j_\gamma = \begin{cases} (a\nabla u_h)\,|_K \cdot n_K + (a \nabla u_h)\,|_{K'} \cdot n_{K'} & \text{if } \gamma = \partial K \cap \partial K', \\ (a\nabla u_h)\,|_K \cdot n_K - g & \text{if } \gamma = \partial K \cap \Gamma_N, \\ (a\nabla u_h)\,|_K \cdot n_K & \text{if } \gamma = \partial K \cap (\partial\Omega \setminus \Gamma_N). \end{cases} \tag{5.7}$$

For the adaptive procedure, we wish to obtain local contributions that can be computed and compared elementwise. We choose

$$R^h(u_h; v) = \sum_{K \subset \Omega} R^h_K(u_h; v), \tag{5.8}$$

where $R^h_K(u_h; v)$ is the elementary contribution to the error, defined by

$$R^h_K(u_h; v) = \int_K r_K v \, dx - \frac{1}{2} \sum_{\gamma \subset (\partial K \setminus \partial\Omega)} \int_\gamma j_\gamma v \, ds - \sum_{\gamma \subset (\partial K \cap \partial\Omega)} \int_\gamma j_\gamma v \, ds. \tag{5.9}$$

This process can be used to compute the error either in energy norm (5.2) by setting $v = u - u_h$ or in the quantities of interest (5.3) by setting $v = p_i - p_{i,h}$. Of course the exact solution u (resp. p_i) is unavailable in practice, so that it is replaced by an approximation, denoted $\tilde{u}$ (resp. $\tilde{p}_i$) computed in the same space $\tilde{V}_h$ as the one used to get the enhanced quantities of interest values.

As refinement criterion, we choose the so-called "maximum strategy" [1,2], i.e. we mark for refinement all elements that satisfy

$$\frac{\left|R^h_K(u_h; v)\right|}{\max_K \left|R^h_K(u_h; v)\right|} > \delta, \tag{5.10}$$

where $\delta \in (0, 1)$ is a chosen threshold. In the numerical experiments, we chose $\delta = 0.5$.

We now turn our attention to error estimates for the solution w_h of the constrained problem (3.15). Again, the approach is based on the use of the residual, which is evaluated this time with respect to the computed solution w_h

$$R^h(w_h; v) = f(v) - a(w_h, v) = a(u - w_h, v). \tag{5.11}$$

Note that the modified Galerkin orthogonality (3.17) yields

$$\begin{aligned} R^h(w_h; v) &= a(u - w_h, v - v_h) + \lambda_h \cdot Q(v_h), \\ &= R^h(w_h; v - v_h) + \sum_{j=1}^{k} \lambda_{h,j} a(v_h, p_{j,h}). \end{aligned} \tag{5.12}$$

The error in energy norm satisfies

$$\mathcal{E}_h = \|u - w_h\|_{\mathcal{E}} = \sqrt{a(u - w_h, u - w_h)} = \sqrt{R^h(w_h; u - w_h)}, \tag{5.13}$$

and the error in each quantity of interest is given by

$$\begin{aligned} \mathcal{E}_i = |Q_i(u) - Q_i(w_h)| &= |f(p_i) - a(w_h, p_i)|, \\ &= \left|R^h(w_h; p_i)\right|, \\ &= \left| R^h(w_h; p_i - p_{i,h}) + \sum_{j=1}^{k} \lambda_{h,j} a(p_{i,h}, p_{j,h}) \right|, \end{aligned} \tag{5.14}$$

where the modified Galerkin orthogonality (3.17) was used. In (5.13) (resp. (5.14)), we proceeded to a straightforward extension of the classical approach (5.2) (resp. (5.3)). Though this approach yields satisfying results, we investigated a different error representation approach aiming at separating the two sources of errors in the numerical solution w_h, namely the classical error due to the discretization of space V into the finite dimensional space V_h and the additional error term due to the constrained minimization. Using the classical Galerkin orthogonality between $u - u_h \in V$ and $u_h - w_h \in V_h$, it holds

$$\mathcal{E}_h^2 = \|u - w_h\|_{\mathcal{E}}^2 = \|u - u_h\|_{\mathcal{E}}^2 + \|u_h - w_h\|_{\mathcal{E}}^2, \tag{5.15}$$

where the first term is the discretization error in the classical solution u_h. Now using (5.2) and Theorem 3, it follows that

$$\mathcal{E}_h^2 = R^h(u_h; u - u_h) + \left\| \sum_{j=1}^{k} \lambda_{h,j} p_{j,h} \right\|_{\mathcal{E}}^2. \tag{5.16}$$

The second term can be interpreted as the error due to the introduction of the constraint. We further note that it can be computed exactly since the Lagrange multipliers $\lambda_h \in \mathbb{R}^k$ and the finite element adjoint solutions $p_{i,h} \in V_h$ are known at this stage. Moreover, local contributions can be derived by splitting the integral on Ω to integrals on each element $K \subset \Omega$.

As far as the error in the quantity of interest Q_i is concerned, it holds

$$\begin{aligned} \mathcal{E}_i = |Q_i(u) - Q_i(w_h)| &= |Q_i(u - u_h) + Q_i(u_h - w_h)|, \\ &= \left| R^h(u_h;\, p_i - p_{i,h}) + \sum_{j=1}^{k} \lambda_{h,j} Q_i(p_{j,h}) \right|, \\ &= \left| R^h(u_h;\, p_i - p_{i,h}) + \sum_{j=1}^{k} \lambda_{h,j} a(p_{j,h},\, p_{i,h}) \right|, \end{aligned} \tag{5.17}$$

where we used (5.3) and Theorem 3. The first term is the contribution due to the discretization error in the classical solution u_h. The second term can be interpreted as the error due to the introduction of the constraint. Again, the second term can be computed exactly and local contributions on each element can be derived.

Comparing (5.16) and (5.17), it appears that the additional error term scales with $\|\lambda_h\|_1^2$ for the error in energy norm while only with $\|\lambda_h\|_1$ for the error in each quantity of interest. The different contributions to the errors will be illustrated in the next section. However, by replacing the adjoint solution p_i by the computable approximation $\tilde{p}_i \in \tilde{V}_h$ either in (5.14) or in (5.17), one obtains an estimate of the error in the quantity of interest that is zero. Indeed, $Q_i(w_h) = \alpha_i = f(\tilde{p}_i) = Q_i(\tilde{u})$. Of course, one could use an even higher-order approximation for the purpose of error estimation, but the cost of the method would then be prohibitive compared to a traditional approach. Nevertheless, the local contributions can be used to mark the elements that contribute largely to the error.

In the case of adaptation for the error in the energy norm (5.16), the element contributions are defined as

$$\begin{aligned} \mathcal{E}_h^2 &= R^h(u_h; u - u_h) + \| \sum_{j=1}^{k} \lambda_{h,j} p_{j,h} \|_{\mathcal{E}}^2, \\ &= \sum_{K \subset \Omega} R_K^h(u_h; u - u_h) + \sum_{i,j=1}^{k} \lambda_{h,i} \lambda_{h,j} a(p_{i,h},\, p_{j,h}), \\ &= \sum_{K \subset \Omega} \left(R_K^h(u_h; u - u_h) + \sum_{i,j=1}^{k} \lambda_{h,i} \lambda_{h,j} a_K(p_{i,h},\, p_{j,h}) \right), \end{aligned} \tag{5.18}$$

where the first term $R_K^h(u_h; u - u_h)$ was defined in (5.9) and the bilinear form a_K relative to each element K is given by

$$a_K(u, v) = \int_K a \nabla u \cdot \nabla v \, dx, \tag{5.19}$$

following the example for the bilinear form a chosen in (5.4). To avoid eventual cancellation between the two sources during the marking process, we will consider the following refinement indicator

$$\left|R_K^h(u_h; u - u_h)\right| + \left|\sum_{i,j=1}^{k} \lambda_{h,i}\lambda_{h,j} a_K(p_{i,h}, p_{j,h})\right|. \tag{5.20}$$

The choice of considering absolute values when computing the local indicators may lead to pessimistic results. However, it is motivated by observing that the two contributions may cancel each other while still being large sources of errors that may need to be controlled. Again, in practice the exact solution $u \in V$ is replaced by the computable approximation $\tilde{u} \in \tilde{V}_h$.

In the case of adaptation for the error in the ith quantity of interest (5.17), the element contributions are defined as

$$\begin{aligned} \mathcal{E}_i &= \left|R^h(u_h; p_i - p_{i,h}) + \sum_{j=1}^{k} \lambda_{h,j} a(p_{j,h}, p_{i,h})\right|, \\ &= \left|\sum_{K\subset\Omega} \left(R_K^h(u_h; p_i - p_{i,h}) + \sum_{j=1}^{k} \lambda_{h,j} a_K(p_{j,h}, p_{i,h})\right)\right|. \end{aligned} \tag{5.21}$$

As previously, to avoid eventual cancellation between the two sources during the marking process, we will consider the following refinement indicator

$$\left|R_K^h(u_h; p_i - p_{i,h})\right| + \left|\sum_{j=1}^{k} \lambda_{h,j} a_K(p_{j,h}, p_{i,h})\right|. \tag{5.22}$$

Again, in practice the adjoint solution $p_i \in V$ is replaced by the computable approximation $\tilde{p}_i \in \tilde{V}_h$.

6. Numerical examples

In this section, we numerically illustrate the proposed approach on some academic boundary-value problems. For the finite element simulations, we use square elements and V_h is defined as the space spanned by the bilinear Lagrange functions. For the enhanced quantities of interest and error estimation, we use $\tilde{V}_h$ the space spanned by the hierarchical integrated Legendre polynomials up to quadratic order.

We will consider three examples: the first two consist of a Poisson equation. They are used to illustrate the efficiency of the method introduced in this work under uniform refinements. In the first example we consider a single quantity of interest that is conforming to the finite element mesh, while in the second example, we consider two quantities of interest that no longer conform to the mesh. The last example involves a diffusion equation with a piecewise constant coefficient, thus mimicking the so-called "L-shaped problem" in which the exact solution exhibits weak-singularities. It is used to illustrate the efficiency of the adaptive mesh refinement procedure introduced in this study.

Example 1.

The first model problem we consider consists of the Poisson equation with homogeneous Dirichlet conditions

$$\begin{cases} -\Delta u = 1, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{6.1}$$

where $\Omega = (0, 1)^2$. The exact solution of (6.1) can be found using Fourier series and is shown in Fig. 1b. We mention that $u \in H^3(\Omega)$ for this problem.

We also suppose that one is interested in the scalar quantity

$$Q(u) = \frac{1}{|\omega|}\int_\omega u\, dx, \tag{6.2}$$

where ω is a subdomain of Ω, illustrated in Fig. 1a, and defined as

$$\omega = \{(x, y) \in \Omega;\ 23/32 \le x \le 29/32,\ 3/32 \le y \le 11/32\}. \tag{6.3}$$

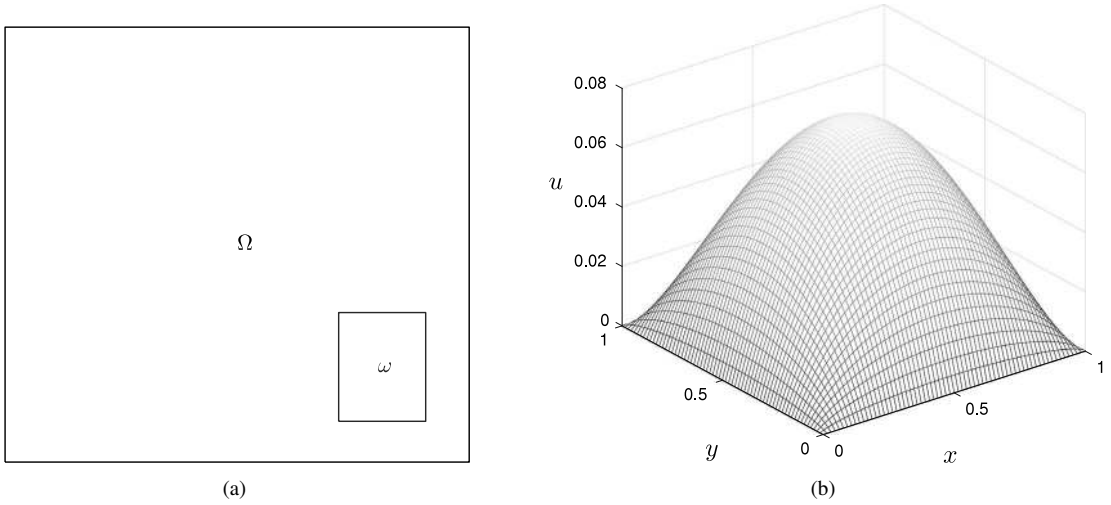

Fig. 1. Geometry and solution for Example 1.

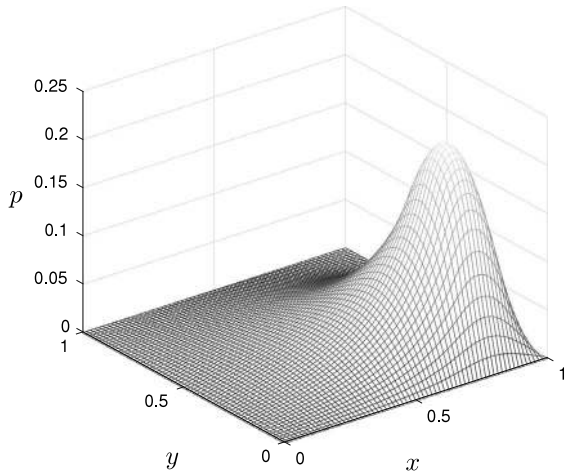

Fig. 2. Adjoint solution for Example 1.

In this first example, the region of interest ω coincides with the mesh (after a few uniform refinements).

The exact value of the quantity of interest (6.2) can be computed using the Fourier expansion of u. We mention that Q is continuous on $H^1(\Omega)$. The adjoint solution $p \in H^3(\Omega)$ is shown in Fig. 2.

In order to assess the efficiency of the constrained approach introduced in this paper as well as of the error estimation procedure, we perform a sequence of uniform refinements with inverse mesh sizes $h^{-1} = 4, 8, 16, \ldots, 256$, estimate the resulting errors, and measure the effectivity of the estimators. In Fig. 3a, we collect the normalized exact errors in energy norm and in the quantity of interest for both the classical unconstrained approach u_h and the constrained approach w_h proposed in this paper. In Fig. 3b, we also collect the effectivity indices i_{eff}, which is classically defined as the ratio of the estimated error over the exact error.

As can be seen from Fig. 3a, the errors in energy norm for the classical unconstrained approach u_h and for the constrained approach w_h have the same convergence rate $\mathcal{O}(h)$, as predicted by the results of *a priori* error estimation and Theorem 4. The effectivity indices for the errors in the energy norm are also very similar for both approaches and are in the [0.996, 1] interval.

Concerning the error in the quantity of interest, the results are striking on this simple example: the rate of convergence for the constrained approach is twice as large as that obtained by the classical approach: $\mathcal{O}(h^4)$ vs $\mathcal{O}(h^2)$,

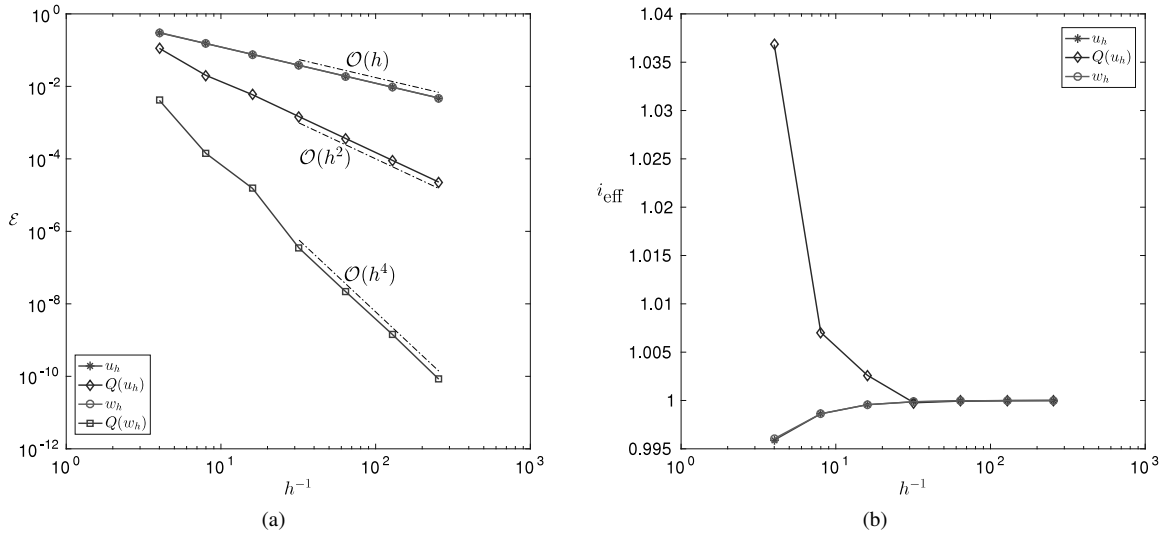

Fig. 3. Exact errors (a) and effectivity indices (b) as functions of the inverse mesh size h^{-1}.

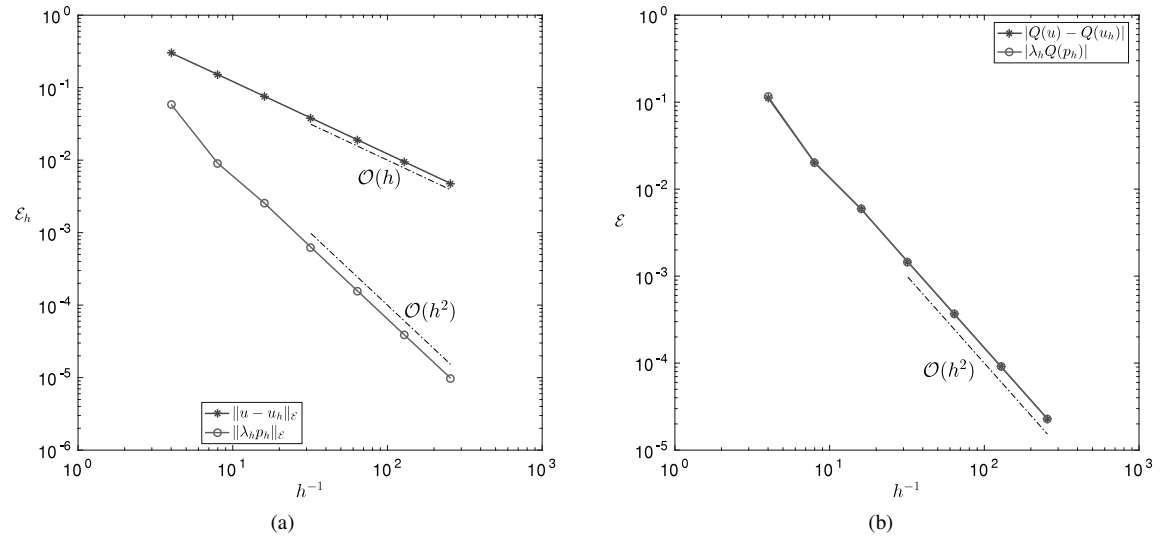

Fig. 4. Detail of the contributions as functions of the inverse mesh size h^{-1}: (a) error in the energy norm; (b) error in the quantity of interest.

again in agreement with the results of *a priori* error estimation [21], based on the fact here that $Q(w_h) = Q(\tilde{u})$ and that both u and p are sufficiently smooth. The error estimator for the quantities of interest is only available for the unconstrained approach (recall discussion about the error estimator for $Q_i(w_h)$ in Section 5) with an effectivity ranging in the [0.9998, 1.037] interval.

In Fig. 4, we compare the terms contributing to the error as defined in (5.16) and in (5.17).

In Fig. 4a, we observe that the two sources of error in the energy norm do not have the same convergence rate. The error term due to the constraint decreases much more rapidly than the classical discretization error. As a result, the total error is similar to the discretization error: indeed w_h is near-optimal in the energy norm. The situation is completely different for the two terms of the error in the quantity of interest: in Fig. 4b, we observe that the two terms

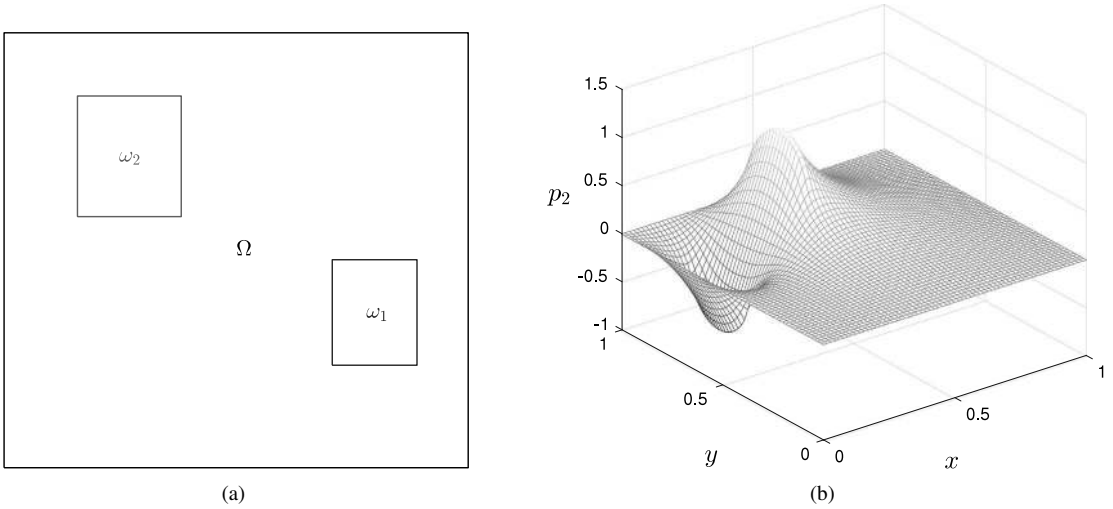

Fig. 5. Geometry and adjoint solution p_2 for Example 2.

are almost equal. In fact, they have opposite signs so that they mostly cancel when added. As a result the convergence rate of the error in the quantity of interest for w_h is increased.

Example 2.

The second model problem consists of the same boundary-value problem (6.1) introduced in Example 1. However, this time we suppose that we are interested in the two quantities

$$Q_1(u) = \frac{1}{|\omega_1|}\int_{\omega_1} u\,dx, \text{ and } Q_2(u) = \frac{1}{|\omega_2|}\int_{\omega_2} \boldsymbol{i}\cdot\nabla u\,dx, \tag{6.4}$$

where $\boldsymbol{i}$ denotes the horizontal unit vector, and ω_1, ω_2 are two subdomains of Ω, illustrated in Fig. 5a, and defined as

$$\begin{aligned}\omega_1 &= \left\{(x,y)\in\Omega;\ 1/\sqrt{2}\le x\le 1/\sqrt{2}+1/\sqrt{30},\ 1/\sqrt{18}\le y\le 1/\sqrt{18}+1/\sqrt{17}\right\},\\ \omega_2 &= \left\{(x,y)\in\Omega;\ 1/\sqrt{40}\le x\le 1/\sqrt{40}+1/\sqrt{20},\ 1/\sqrt{3}\le y\le 1/\sqrt{3}+1/\sqrt{13}\right\},\end{aligned} \tag{6.5}$$

where the irrational coordinates were chosen so that the regions of interest ω_1, ω_2 never coincide with the meshes. In Fig. 5b, we present the adjoint solution p_2. The adjoint solution p_1 is similar to that shown in Fig. 2 and is not shown here.

Again, the exact values of the quantities of interest (6.4) can be computed using the Fourier expansion of u. We mention that Q_1 and Q_2 are continuous on $H^1(\Omega)$. Furthermore, we have the following regularity for the adjoint solutions: $p_1 \in H^3(\Omega)$ while $p_2 \in H^2(\Omega)$, only.

Once more, we perform a sequence of uniform refinements, estimate the resulting errors, and measure the effectivity indices of the estimators. In Fig. 6a, we show the normalized exact errors in energy norm and in the two quantities of interest for both the classical unconstrained solution u_h and the constrained solution w_h. The effectivity indices i_{eff} are shown for this case in Fig. 6b.

We observe from Fig. 6a that the errors in energy norm in u_h and w_h are again almost equal: the error for the constrained solution is 2% larger than for the unconstrained solution on the coarsest mesh considered, and only 0.0002% larger for the finest mesh considered. This behavior is very similar to that observed in the first example. The effectivity indices for the errors in the energy norm are also very similar for both approaches and are in the [0.996, 1] interval.

Concerning the error in the quantity of interest Q_1, the rate of convergence for the constrained approach is again twice as large as that obtained by the classical approach: $\mathcal{O}(h^4)$ vs $\mathcal{O}(h^2)$. For the quantity of interest Q_2, the rate

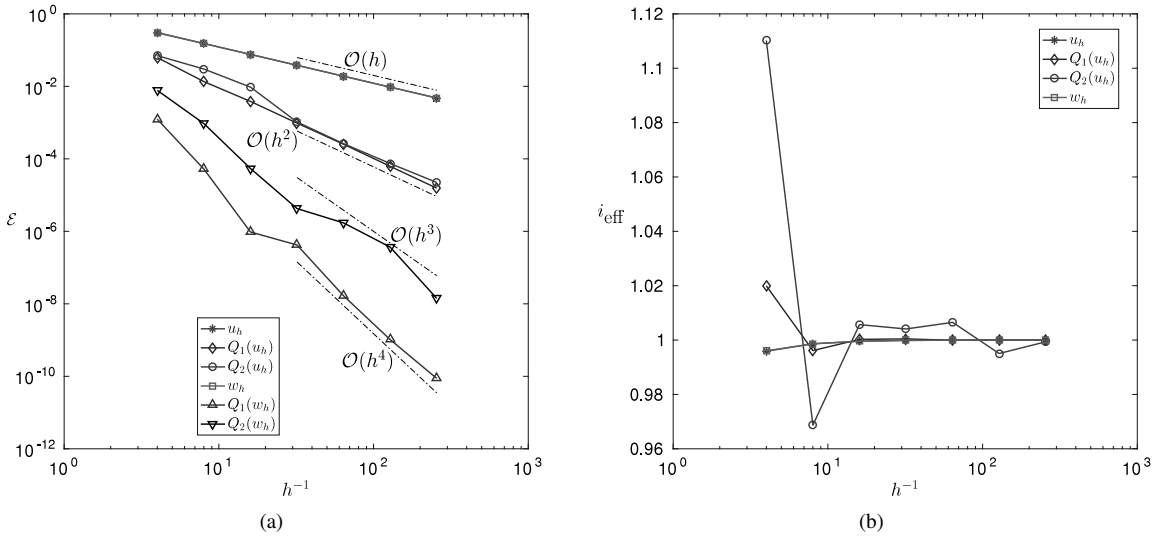

Fig. 6. Exact errors (a) and effectivity indices (b) as functions of the inverse mesh size h^{-1}.

of convergence only increases by one order: $\mathcal{O}(h^3)$ vs $\mathcal{O}(h^2)$. This difference is due to the limited regularity of the adjoint solution for the second quantity of interest: recall $p_1 \in H^3(\Omega)$ while $p_2 \in H^2(\Omega)$ only. Indeed, with such regularity but no more, we have $\|p_2 - \tilde{p}_2\|_{\mathcal{E}} = \mathcal{O}(h)$ as $h \to 0$ while $\|u - \tilde{u}\|_{\mathcal{E}} = \mathcal{O}(h^2)$. Hence the quantity $Q_2(u)$ is approximated with order $\mathcal{O}(h^3)$. The error estimator for the quantities of interest shows an effectivity ranging in the [0.969, 1.110] interval.

We emphasize here that the constrained approach requires solving the higher-order adjoint problems (3.21) beforehand, which is not the case for the unconstrained approach. As a result, when performing a uniform refinement (i.e. without any error estimation procedure) the two approaches are not on an equal footing when comparing their respective costs. In order to fairly compare the two methods, one has to consider an adaptive refinement procedure, which is the aim of the third example.

Example 3.

Again we consider $\Omega = (0,1)^2$, and choose a point $(l_x, l_y) \in \Omega$ so that Ω is split into two regions: $\Omega_1 = \{(x,y) \in \Omega;\, x > l_x \text{ and } y > l_y\}$, and the complementary region $\Omega_0 = \Omega \setminus \Omega_1$. We choose $l_x = l_y = 1/2$. We introduce on Ω the piecewise constant coefficient a such that $a_{|\Omega_i} = a_i$, $i = 0, 1$, with $a_0 = 1$ and $a_1 = 100$. The third problem consists of a diffusion equation with diffusivity coefficient a so that it features a weak-singularity (i.e. the gradient of the solution is singular), whose solution u is subjected to Robin boundary conditions on $\partial\Omega$

$$\begin{cases} -\nabla \cdot a\nabla u = f, & \text{in } \Omega, \\ \boldsymbol{n} \cdot a\nabla u + u = g, & \text{on } \partial\Omega. \end{cases} \tag{6.6}$$

The exact solution u is constructed using the so-called manufactured solution method and is chosen to be harmonic of the form

$$u = u(r,\theta) = \begin{cases} A_0 r^\mu \cos(\mu\theta) + B_0 r^\mu \sin(\mu\theta), & \text{in } \Omega_0, \\ A_1 r^\mu \cos(\mu\theta) + B_1 r^\mu \sin(\mu\theta), & \text{in } \Omega_1, \end{cases} \tag{6.7}$$

where (r,θ) are the polar coordinates centered at (l_x, l_y). The constants μ, A_0, B_0, A_1 and B_1 are chosen such that u is continuous in Ω and $\boldsymbol{n} \cdot a\nabla u$ is continuous across the interface between Ω_0 and Ω_1. The source term f and boundary datum g are derived by injecting (6.7) into (6.6). We mention that $f = 0$ because u is taken to be harmonic in Ω. Table 1 collects the values of the constant parameters μ, A_0 and B_0 while we have $A_1 = A_0$ and $B_1 = (a_0/a_1)B_0$.

Note that by construction we have $u \in H^{1+\mu-\epsilon}(\Omega)$, where $\epsilon > 0$ is arbitrarily small. The manufactured problem resembles the so-called "L-shaped problem" constructed here with a finite contrast a_1/a_0. As a result, the solution

Table 1
Values of the parameters μ, A_0 and B_0 used for the third example.

μ	A_0	B_0
0.6739	0.0171	0.9998

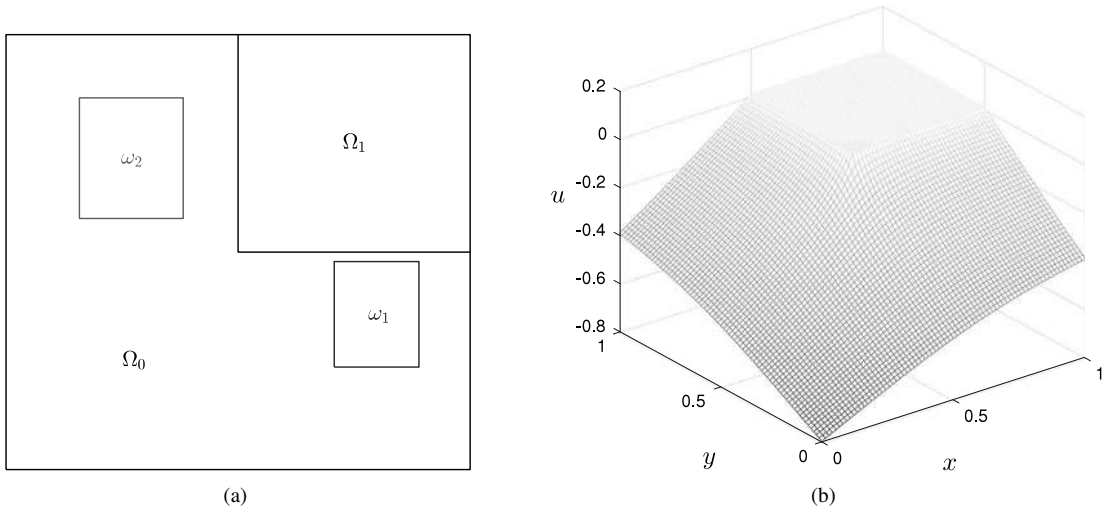

Fig. 7. Geometry and solution for Example 3.

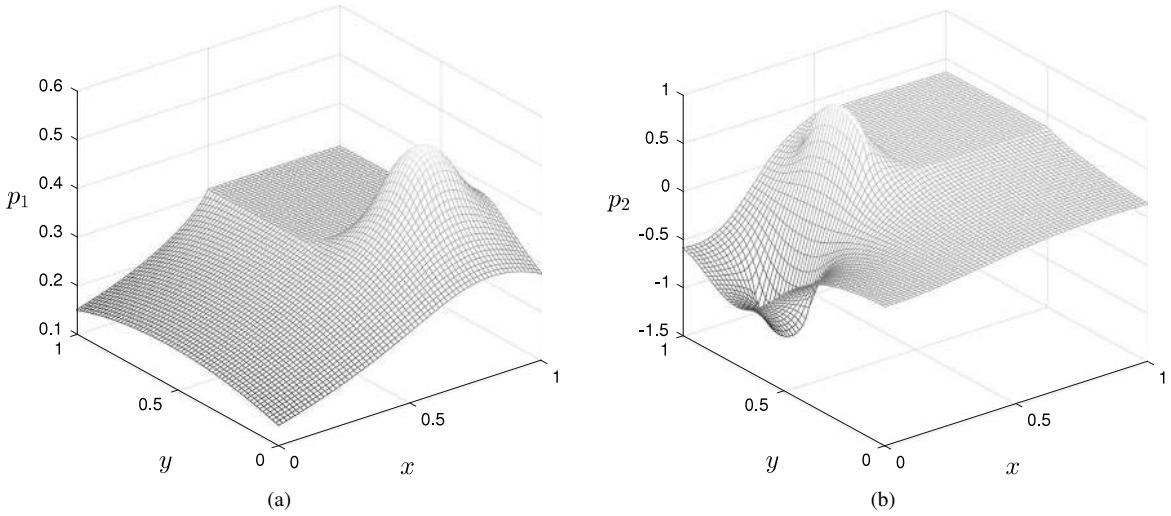

Fig. 8. Adjoint solutions for Example 3.

exhibits a weak-singularity at the corner (l_x, l_y) and its gradient is discontinuous along the interface $\partial\Omega_1 \setminus \partial\Omega$. In order to simplify the presentation, the initial mesh is chosen to be conforming to the interface by taking $h^{-1} = 2$.

We show in Fig. 7 the geometry and the manufactured solution for this third example. The quantities of interest are the same as in the second example, see (6.4)–(6.5). The adjoint solutions associated with these quantities of interest are illustrated in Fig. 8.

We now turn to the adaptive procedure for above problem. When an element is marked for refinement, it is divided into four squares of equal areas, which introduces hanging nodes [22]. We will compare four types of refinement based on the following criteria:

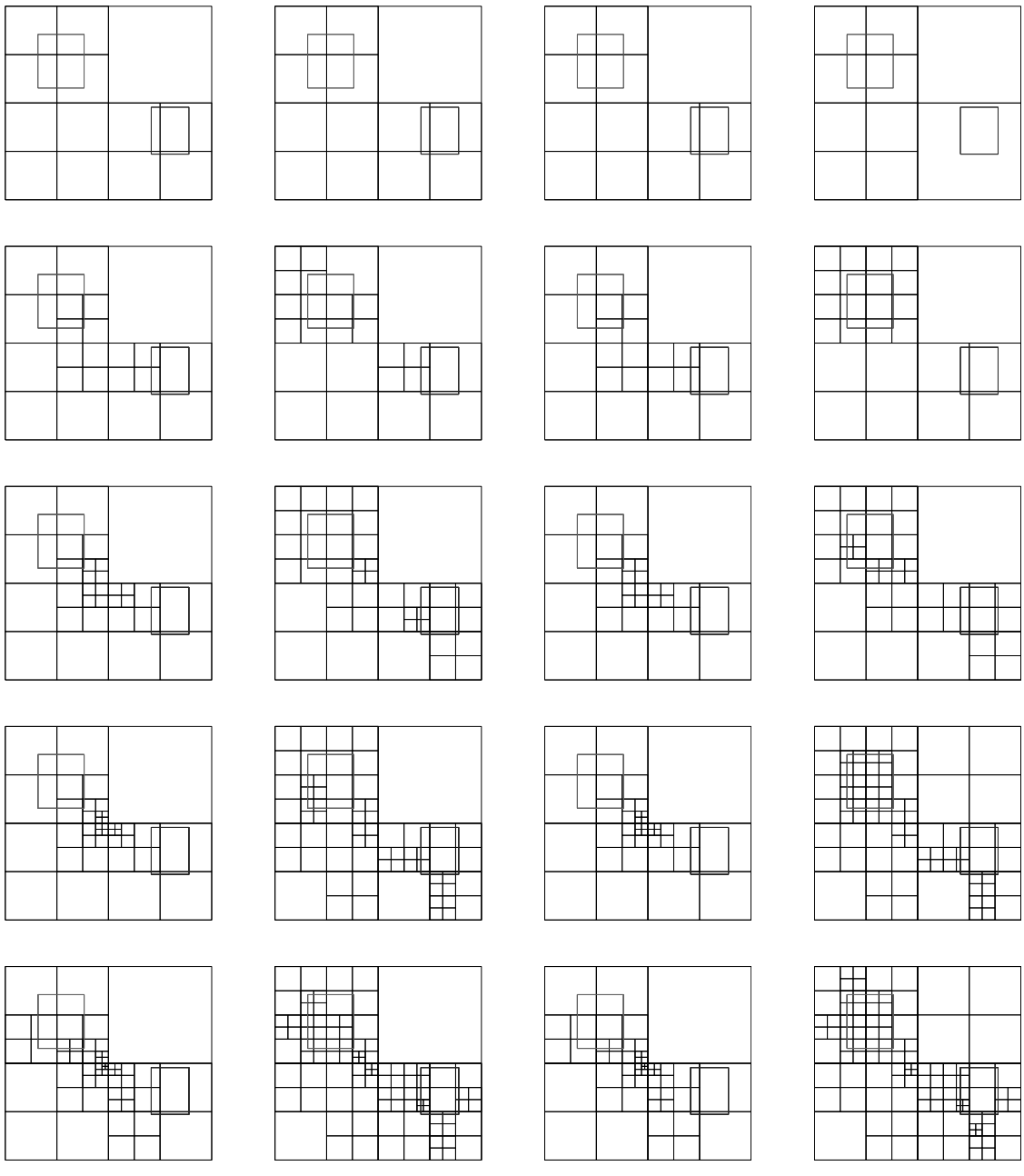

Fig. 9. Sequences of adapted meshes for Example 3. (a): refinement in energy norm for u_h; (b): refinement in the quantities of interest for u_h; (c): refinement in energy norm for w_h; (d): refinement in the quantities of interest for w_h.

(a) adaptation in norm for the unconstrained solution u_h,
(b) adaptation in the quantities of interest for the unconstrained solution u_h,
(c) adaptation in norm for the constrained solution w_h,
(d) adaptation in the quantities of interest for the constrained solution w_h.

We mention that for the adaptation based on the two quantities of interest, elements are marked for refinement if any of the two error indicators associated with Q_1 and Q_2 exceeds the prescribed threshold.

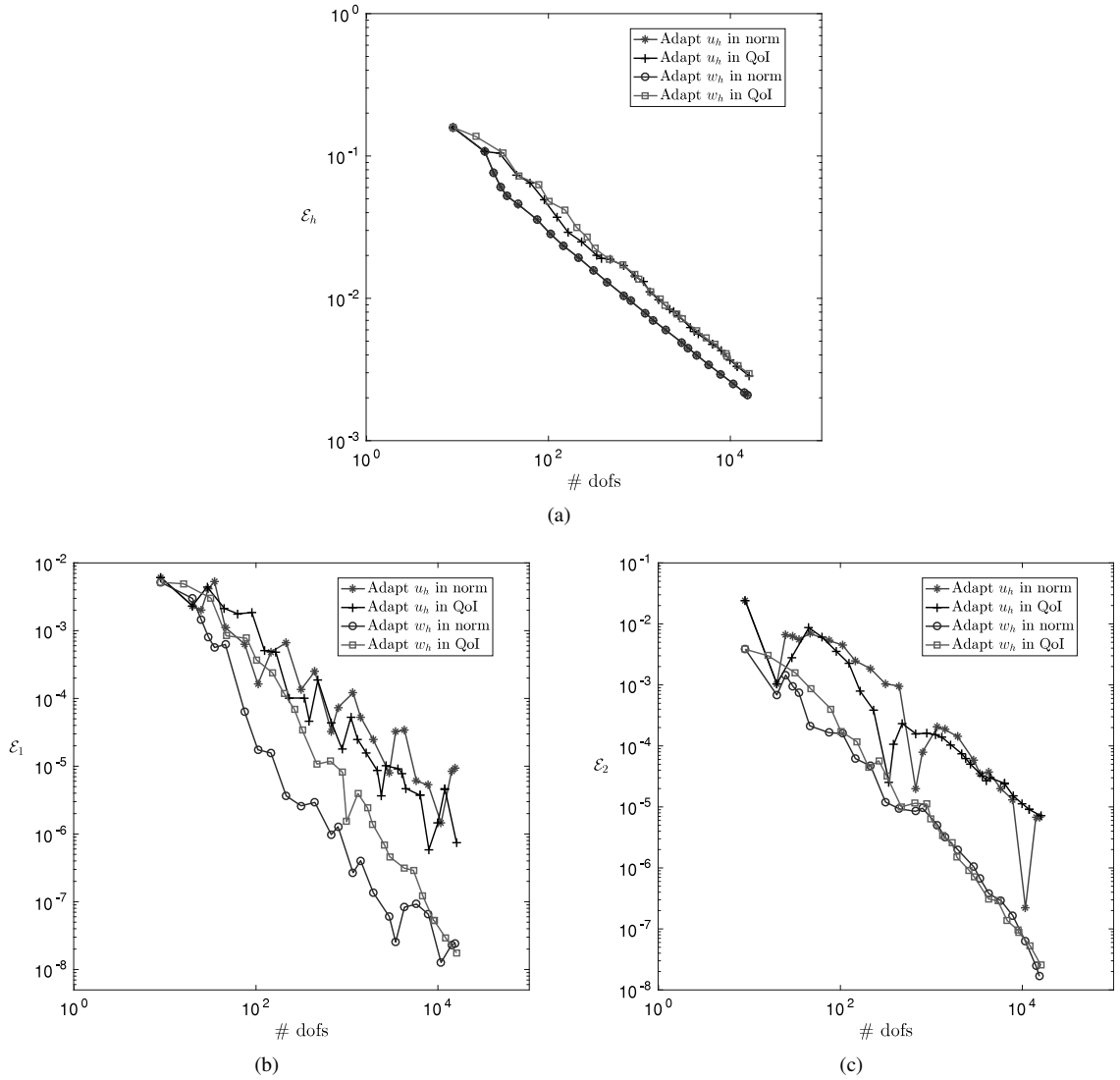

Fig. 10. Convergence results for the four considered methods. (a): convergence in energy norm; (b) and (c): convergence in the two quantities of interest.

We show in Fig. 9 the resulting sequences of adapted meshes. All four methods manage to capture the singularity at the corner of the interface. In addition, we note that the approaches based on the quantities of interest, (b) and (d), accentuate the refinement in the two regions of interest ω_1, ω_2. Furthermore, the refinements in energy norm (a) and (c) are very similar, which is due to the relatively small contribution of the term related to the constraint, recall Fig. 4a and (5.16). Conversely, the adapted meshes obtained for the refinement based on the quantities of interest (b) and (d) are less similar because the additional error term scales with $\|\lambda_h\|_1$, see (5.17).

The convergence plots for the four methods are shown in Fig. 10. We mention that 25 iterations (15,412 dofs) were considered for approaches (a) and (c), 26 iterations (16,020 dofs) for approach (d), and 30 iterations (16,041 dofs) for approach (b). The two approaches based on the constrained solution yield the best results in terms of convergence of the quantities of interest. In particular, all results converge asymptotically at optimal rates, namely, where N_{dof} denotes the number of degrees of freedom,

$$\|u - u_h\|_{\mathcal{E}} \leq C(N_{\text{dof}})^{-p/d} \text{ and } \|u - w_h\|_{\mathcal{E}} \leq C(N_{\text{dof}})^{-p/d}, \tag{6.8}$$

with $p = 1$ for both u_h and w_h, and

$$|Q_i(u) - Q_i(u_h)| \leq C(N_{\text{dof}})^{-2p/d} \text{ and } |Q_i(u) - Q_i(w_h)| \leq C(N_{\text{dof}})^{-2p/d}, \tag{6.9}$$

with $p = 1$ for u_h and $p = 2$ for w_h, for both $i = 1, 2$.

7. Conclusion

We have introduced a novel formulation for taking into account quantities of interest in finite element approximations. The approach is very different from classical procedures involving goal-oriented adaptivity. In the latter the adjoint problems are solved after computing the primal solution in order to assess and control the errors in the quantity of interest. In the proposed approach, the adjoint problems are solved beforehand in order to obtain enhanced values for the quantities of interest, which are then introduced in the formulation of the primal problem by means of a constraint. In this study, we have proved that the corresponding mixed formulation was well-posed, and that the constrained finite element solution retained near-optimality in the energy norm while being much more accurate in the quantities of interest. Error estimators were derived for the proposed approach, with an emphasis on explicitly identifying the two contributions to the error, namely the classical discretization error and the error due to the introduction of a constraint. The efficiency of the novel formulation and of the corresponding mesh refinement procedure was demonstrated on a series numerical examples.

Future work will focus on the extension of the present work to non-linear problems and non-linear quantities of interest. We also note that the methodology can be straightforwardly extended to a worst-case multi-objective formulation [9] by considering one dual problem using an approximate supporting functional of the objective set rather than solving a dual problem for each quantity of interest. In a forthcoming paper, we will extend the proposed method to reduced-order modeling methods, such as the Proper Generalized Decomposition [23,24] method, for which we have developed the framework needed to enforce constraints [25]. We anticipate that the construction of reduced models using such an approach could help provide accurate estimates of quantities of interest at very low computational cost. This would be particularly useful for the treatment of uncertainty quantification problems, in which case one has to estimate output quantities of interest for a very large number of parameter samples.

Acknowledgments

SP is grateful for the support by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada. He also acknowledges the support by KAUST under Award Number OCRF-2014-CRG3-2281. Moreover, the authors gratefully acknowledge Jonathan Vacher for fruitful discussions on the subject.

References

[1] J.T. Oden, S. Prudhomme, Goal-oriented error estimation and adaptivity for the finite element method, Comput. Math. Appl. 41 (5–6) (2001) 735–756.

[2] S. Prudhomme, J.T. Oden, On goal-oriented error estimation for elliptic problems: Application to the control of pointwise errors, Comput. Methods Appl. Mech. Engrg. 176 (1–4) (1999) 313–331.

[3] R. Becker, R. Rannacher, An optimal control approach to a posteriori error estimation in finite element methods, Acta Numer. 10 (2001) 1–102.

[4] J.T. Oden, S. Prudhomme, Estimation of modeling error in computational mechanics, J. Comput. Phys. 182 (2) (2002) 496–515.

[5] J.H. Chaudhry, E.C. Cyr, K. Liu, T.A. Manteuffel, L.N. Olson, L. Tang, Enhancing least-squares finite element methods through a quantity-of-interest, SIAM J. Numer. Anal. 52 (6) (2014) 3085–3105.

[6] D. Estep, M. Holst, M. Larson, Generalized Green's functions and the effective domain of influence, SIAM J. Sci. Comput. 26 (2005) 1314–1339.

[7] R. Hartmann, Multitarget error estimation and adaptivity in aerodynamic flow simulations, SIAM J. Sci. Comput. 31 (2008) 708–731.

[8] B. Endtmayer, T. Wick, A partition-of-unity dual-weighted residual approach for multi-objective goal functional error estimation applied to elliptic problems, Comput. Methods Appl. Math. (2017) Published online.

[9] E.H. van Brummelen, S. Zhuk, G.J. van Zwieten, Worst-case multi-objective error estimation and adaptivity, Comput. Methods Appl. Mech. Engrg. 313 (2017) 723–743.

[10] P.B. Bochev, M.D. Gunzburger, Least-Squares Finite Element Methods, Springer Science & Business Media, 2009.

[11] P.G. Ciarlet, The Finite Element Method for Elliptic Problems, North-Holland, Amsterdam, 1978.

[12] I. Babuška, Error-bounds for finite element method, Numer. Math. 16 (4) (1971) 322–333.

[13] F. Brezzi, On the existence, uniqueness and approximation of saddle-point problems arising from lagrangian multipliers, Rev. Fr. Autom. Inform. Rech. Oper. 8 (2) (1974) 129–151.

[14] J.T. Oden, L.F. Demkowicz, Applied Functional Analysis, CRC Press Incorporated, 1996.
[15] Y. Saad, Iterative Methods for Sparse Linear Systems, SIAM, 2003.
[16] W. Karush, Minima of functions of several variables with inequalities as side constraints, Master's thesis, Dept. of Mathematics, Univ. of Chicago, Chicago, Illinois, 1939.
[17] H.W. Kuhn, A.W. Tucker, Nonlinear programming, in: Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, Calif, 1951, pp. 481–492.
[18] Interior Point OPTimizer (IPOPT) software package, https://projects.coin-or.org/Ipopt/.
[19] A. Wächter, L.T. Biegler, On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming, Math. Program. 106 (1) (2006) 25–57.
[20] M. Ainsworth, J.T. Oden, A posteriori error estimation in finite element analysis, Comput. Methods Appl. Mech. Engrg. 142 (1–2) (1997) 1–88.
[21] I. Babuška, A. Miller, 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 Engrg. 20 (6) (1984) 1085–1109.
[22] G.F. Carey, J.T. Oden, Finite Elements: Computational Aspects, Prentice Hall, 1984.
[23] F. Chinesta, R. Keunings, A. Leygue, The Proper Generalized Decomposition for Advanced Numerical Simulations, Springer International Publishing, 2014.
[24] A. Nouy, A priori model reduction through proper generalized decomposition for solving time-dependent partial differential equations, Comput. Methods Appl. Mech. Engrg. 23–24 (199) (2010) 1603–1626.
[25] K. Kergrene, S. Prudhomme, L. Chamoin, M. Laforest, Approximation of constrained problems using the PGD method with application to pure Neumann problems, Comput. Methods Appl. Mech. Engrg. 317 (2016) 507–525.