Jyväskylän yliopiston julkaisuarkisto
Jyväskylä University Digital Archive

UNIVERSITY OF JYVÄSKYLÄ

This is an electronic reprint of the original article.
This reprint *may differ* from the original in pagination and typographic detail.

Author(s): Mali, Olli

Title: Error Estimates for a Class of Elliptic Optimal Control Problems

Year: 2017

Version:

Please cite the original version:

Mali, O. (2017). Error Estimates for a Class of Elliptic Optimal Control Problems. Numerical Functional Analysis and Optimization, 38(1), 58-79. https://doi.org/10.1080/01630563.2016.1217881

All material supplied via JYX is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

Error Estimates for a Class of Elliptic Optimal Control Problems

O. Mali

Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland

Abstract

In this paper, functional type a posteriori error estimates are presented for a certain class of optimal control problems with elliptic PDE-constraints. It is assumed that in the cost functional the state is measured in terms of the energy norm generated by the state equation. The functional a posteriori error estimates developed by Repin in late 90's are applied to estimate the cost function value from both sides without requiring the exact solution of the state equation. Moreover, a lower bound for the minimal cost functional value is derived. A meaningful error quantity coinciding with the gap between the cost functional values of an arbitrary admissible control and the optimal control is introduced. This error quantity can be estimated from both sides using the estimates for the cost functional value. The theoretical results are confirmed by numerical tests.

Accepted Manuscript

Received 23 October 2015; Revised 22 July 2016; Accepted 24 July 2016

O. Mali, Department of Mathematical Information Technology, University of Jyväskylä, Mattilanniemi 2, 40100 Jyväskylä, Finland; E-mail: olli.mali@jyu.fi

1. INTRODUCTION

The optimal control problem with elliptic PDE constraint has been fully discussed in the book of Lions (1, Chap. 2) from the viewpoint of the existence and uniqueness of the solution, necessary conditions, different type of controls and observation, etc. Since 70's, the main development has occurred on numerical methods to generate approximate solution to the respective problem (see, e.g., (2; 3; 4) and references therein).

This paper presents guaranteed two-sided estimates for the value of the cost functional (assuming that the state equation can not be solved exactly) and shows how they can be used to generate estimates for a certain error quantity (cf. (3.13) and Theorem 3.4). In the case of unconstrained control, some estimates and numerical tests have been in presented in (5) and in (6), the case of "box constraints" is treated. Here, these results are extended considerably for constraints of more general type, a new error quantity is introduced, and the results are confirmed by numerical tests. The functional type approach has been studied for a different type const functional in (7), where a different type cost functional is studied and in (8), where similar techniques have been used for time periodic parabolic optimal control problem.

In contrast to the typical residual based error indicators associated with adaptive methods, these estimates are always valid regardless of the accuracy of the approximation itself.

Accepted Manuscript

In particular, the estimates are valid also outside the asymptotic regime and do not contain any unknown constants.

In section 2, definitions and standard results related to optimal control problems with elliptic state equation are recalled. The main assumption is that cost functionals are assumed to be of a certain type, where the state is measured in terms of the energy norm generated by the state equation. This is a special case of the general theory which can be found, e.g., from monographs (1; 3; 4). This assumption is not very restrictive, since in the case of less regular objective state, the cost functional can be shifted by a projection (cf. Remark 2.1). The goal of this paper is to show how the principal relations recalled in Sect. 2 generate the natural error measure and the respective functional type a posteriori error estimates.

In section 3, the functional a posteriori error estimates (see monographs (9; 6; 10) and references therein) for the state equation are applied to generate two-sided bounds for the value of the cost functional. The strong connections between the functional type a posteriori error estimates and the principal relations generating the optimal control problem are underlined. This is a joint property of all functional type estimates. Since the estimates are derived on a functional level (not associated to any particular numerical method), their building blocks are the fundamental relations generating the respective problem. Theorem 3.4 (generalization of (6, Ch. 9, Th. 9.14) for the case of constrained control) is the analog of the Mikhlin identity (cf. Theorem 3.4) for the optimal control

Accepted Manuscript

problem. The proposed error quantity differs from the one presented in (7, Eq. (9)), where some terms are omitted and one obtains an inequality instead of identity.

Some examples of optimal control problem of the type described in Sect. 2 are discussed in Sect. 4.1. Numerical tests in Sect. 4.3 depict how the estimates can be combined with an arbitrary (conforming) numerical method. The objective of these tests is not to generate an efficient solver (Algorithm 2possesses only a poor linear convergence), but to demonstrate that the derived error estimates satisfy the presented theoretical properties, i.e., they are indeed computable (see Algorithm 1),guaranteed, two-sided, and always valid (in contrast of asymptotic properties of error indicators).

2. ELLIPTIC OPTIMAL CONTROL PROBLEM

2.1. Definitions

Let W, H, and U be Hilbert spaces. Their inner products and norms are denoted by subscripts, e.g., $(\cdot,\cdot)_{\mathsf{W}}$ and $\|\cdot\|_{\mathsf{W}}$. Moreover, $\mathsf{V} \subset \mathsf{W}$ is a Hilbert space generated by the inner product $(q, z)_{\mathsf{V}} := (q, z)_{\mathsf{W}} + (\Lambda q, \Lambda z)_{\mathsf{H}}$, where $\Lambda : \mathsf{V} \to \mathsf{H}$ is a linear, bounded operator. The injection from V to W is continuous and V is dense in W. Operator Λ is assumed to satisfy a Friedrichs type inequality

$$\|q\|_{\mathsf{W}} \leq c\|\Lambda q\|_{\mathsf{H}}, \quad \forall\, q \in \mathsf{V}_0, \tag{2.1}$$

Accepted Manuscript

where a subspace $\mathsf{V}_0 \subset \mathsf{V}$ is closed. Assume $\mathsf{V}_0 \subset \mathsf{V} \subset \mathsf{W} \subset \mathsf{V}_0^*$, where V_0^* is the dual space of V_0.

Define linear bounded operators $\mathscr{B} : \mathsf{U} \to \mathsf{V}_0^*$, $\mathscr{A} : \mathsf{H} \to \mathsf{H}$, $\mathscr{N} : \mathsf{U} \to \mathsf{U}$, where $\mathscr{A}$ and $\mathscr{N}$ are symmetric and positive definite,

$$\underline{c}\|q\|_{\mathsf{H}}^2 \le (\mathscr{A}q, q)_{\mathsf{H}} \le \overline{c}\|q\|_{\mathsf{H}}^2, \quad \forall\, q \in \mathsf{H}$$

and

$$\underline{\kappa}\|v\|_{\mathsf{U}}^2 \le (\mathscr{N}v, v)_{\mathsf{U}} \le \overline{\kappa}\|v\|_{\mathsf{U}}^2, \quad \forall\, v \in \mathsf{U},$$

where $\underline{c}$ and $\overline{c}$ ($\underline{\kappa}$ and $\overline{\kappa}$) are positive constants. Thus, they generate inner products

$$(q, z)_{\mathscr{A}} := (\mathscr{A}q, z)_{\mathsf{H}}, \quad (q, z)_{\mathscr{A}^{-1}} := (\mathscr{A}^{-1}q, z)_{\mathsf{H}}, \quad (v, w)_{\mathscr{N}} := (\mathscr{N}v, w)_{\mathsf{U}},$$

and the respective norms

$$\|q\|_{\mathscr{A}} := \sqrt{(\mathscr{A}q, q)_{\mathsf{H}}}, \quad \|q\|_{\mathscr{A}^{-1}} := \sqrt{(\mathscr{A}^{-1}q, q)_{\mathsf{H}}}, \quad \|v\|_{\mathscr{N}} := \sqrt{(\mathscr{N}v, v)_{\mathsf{U}}}.$$

The adjoint operators $\Lambda^* : \mathsf{H} \to \mathsf{V}^*$ and $\mathscr{B}^* : \mathsf{V}_0 \to \mathsf{U}^*$ are defined by the relations

$$\langle \Lambda^* z, q\rangle_{\mathsf{V}} = (z, \Lambda q)_{\mathsf{H}}, \quad \forall\, z \in \mathsf{H},\ q \in \mathsf{V} \tag{2.2}$$

and

$$\langle \mathscr{B}v, q\rangle_{\mathsf{V}_0} = \langle v, \mathscr{B}^* q\rangle_{\mathsf{U}}, \quad \forall\, v \in \mathsf{U},\ q \in \mathsf{V}_0, \tag{2.3}$$

Accepted Manuscript

where $\langle \cdot, \cdot \rangle_{\mathsf{V}}$ denotes the pairing of V and its dual space V^*. By the Riesz representation theorem, there exists an isomorphism (denoted, e.g., by $\mathcal{J}_{\mathsf{U}} : \mathsf{U} \to \mathsf{U}^*$) from any Hilbert space onto the corresponding dual space. It is relevant to define a subspace, where $\Lambda^* z$ has additional regularity,

$$\mathsf{Q} := \{z \in \mathsf{H} \mid \Lambda^* z \in \mathsf{W}\} \subset \mathsf{H}.$$

The norm to V_0^* is

$$\| \ell \| := \sup_{\substack{q \in \mathsf{V}_0 \\ q \neq 0}} \frac{|\langle \ell, q \rangle_{\mathsf{V}_0}|}{\|\Lambda q\|_{\mathcal{A}}}. \tag{2.4}$$

Consider a bilinear form $a : \mathsf{V}_0 \times \mathsf{V}_0 \to \mathbb{R}$,

$$a(q, z) := (\mathcal{A}\Lambda q, \Lambda z)_{\mathsf{H}}. \tag{2.5}$$

It is V_0 -elliptic and continuous. Moreover, in V_0, it generates an energy norm

$$|||q||| := \sqrt{a(q, q)}. \tag{2.6}$$

Example 2.1. In following examples, we demonstrate the general framework on the paradigm of the Dirichlet optimal control problem with distributed control. Here, we present the spaces and operators which generate the Poisson problem, which will be our respective state equation.

Accepted Manuscript

The selection of spaces

$$L_2(\Omega) := \mathsf{W}, \quad L_2(\Omega, \mathbb{R}^d) := \mathsf{H}, \quad H_0^1(\Omega) := \{y \in L_2(\Omega) \mid \|\nabla y\|_{L_2(\Omega,\mathbb{R}^d)} < \infty\} := \mathsf{V},$$

and operators

$$\nabla := \Lambda, \quad -\mathrm{div} := \Lambda^*, \quad \mathrm{Id} := \mathscr{A},$$

generates the following bilinear form

$$a(q, z) := \int_\Omega \nabla q \cdot \nabla z \, \mathrm{d}x.$$

Together with linear right-hand side ,

$$\langle \ell, q \rangle := \int_\Omega fq \, \mathrm{d}x,$$

(in general $\ell \in H^{-1}(\Omega)$, but we assume that $f \in L^2(\Omega)$) it defines the weak form of the Poisson problem with homogeneous Dirichlet boundary conditions: Find $y \in H_0^1(\Omega)$, s.t.,

$$\int_\Omega \nabla y \cdot \nabla q \, \mathrm{d}x = \int_\Omega fq \, \mathrm{d}x, \quad \forall q \in H_0^1(\Omega).$$

The corresponding classical (strong) form of the problem is

$$-\Delta y = f, \quad \text{in } \Omega \subset \mathbb{R}^d,$$

$$y = 0, \quad \text{on } \partial\Omega.$$

Accepted Manuscript

The flux of the solution, i.e., ∇y belongs to the space

$$H(\mathrm{div}, \Omega) := \{z \in L_2(\Omega, \mathbb{R}^d) \mid \|\mathrm{div} z\|_{L_2(\Omega)} < \infty\} := \mathsf{Q}.$$

2.2. Optimal control problem

The state equation is

$$a(y(v), q) = \langle \ell + \mathscr{B}v, q\rangle_{\mathsf{V}_0}, \quad \forall q \in \mathsf{V}_0, \tag{2.7}$$

where $\ell \in \mathsf{V}_0^*$, $v \in \mathsf{U}_{\mathrm{ad}} \subset \mathsf{U}$ is the control, and $y(v) \in \mathsf{V}_0$ is the corresponding state. Let $\mathsf{U}_{\mathrm{ad}} \subset \mathsf{U}$ be a non-empty, convex, and closed set. The cost functional $J : \mathsf{U} \to \mathbb{R}$ is

$$J(v) := |||y(v) - y^d|||^2 + \|v - u^d\|_{\mathcal{N}}^2, \tag{2.8}$$

where $u^d \in \mathsf{U}$ and $y^d \in \mathsf{V}_0$. The optimal control problem is to find $u \in \mathsf{U}_{\mathrm{ad}}$, such that

$$J(u) \leq J(v), \quad \forall v \in \mathsf{U}_{\mathrm{ad}}. \tag{2.9}$$

Under earlier assumptions, J is U-elliptic, coercive, and lower semi-continuous. Thus, the solution of the optimal control problem exists and is unique (see, e.g., (1, Chap. II, Th. 1.2)).

Example 2.2. The Dirichlet problem with distributed control is

$$\min_{v \in \mathsf{U}_{\mathrm{ad}} \subset L^2(\Omega)} J(v) := \|\nabla\left(y(v) - y^d\right)\|_{L_2(\Omega, \mathbb{R}^d)}^2 + \alpha\|v - v^d\|_{L_2(\Omega)}^2,$$

Accepted Manuscript

where

$$-\Delta y(v) = f + v, \quad \text{in } \Omega \subset \mathbb{R}^d,$$

$$y(v) = 0, \quad \text{on } \partial\Omega.$$

The Dirichlet problem is obtained, if besides of the selection of spaces and operators as defined in Example 2.1, we select

$$\mathrm{Id} := \mathscr{B}, \quad \text{and} \quad \alpha := \mathscr{N}.$$

Selection of positive scalar multiplication for $\mathscr{N}$ is typical. It leads to the Tikhonov type regularization.

Remark 2.1. In elliptic optimal control problems, the cost functional is typically of the form

$$\widehat{\mathcal{J}}(v) := \|y(v) - \hat{y}^d\|_{\mathsf{W}}^2 + \|v - u^d\|_{\mathscr{N}}^2.$$

However, in (2.8), the objective is not to match the primal, but the dual variable of the state equation. For example, in the stationary heat (or any diffusion type) problem the objective would not be to generate a desired temperature field, but the respective flux. Similarly, in linear elasticity, the objective would be to generate a certain stress field, not displacements. Such objective fields σ^d may be, e.g., measurement data. The likely lack of regularity can be easily solved as follows.

Accepted Manuscript

Cost functional of type

$$J_2(v) := \|\Lambda y(v) - \sigma^d\|^2_{\mathscr{A}} + \|v - u^d\|^2_{\mathscr{N}}$$

can be shifted using a projection: Find $y^d \in \mathsf{V}_0$ such that

$$(\mathscr{A}(\Lambda y^d - \sigma^d), \Lambda q)_{\mathsf{W}} = 0, \quad \forall q \in \mathsf{V}_0.$$

Then, $J(v) = J_2(v) - \|\Lambda y^d - \sigma^d\|^2_{\mathscr{A}}$.

Example 2.3. The Dirichlet problem presented in Example 2.2 is not of the standard form. Typically the cost functional is of the form

$$\widehat{J}(v) := \|y(v) - \hat{y}^d\|^2_{L_2(\Omega,\mathbb{R}^d)} + \alpha\|v - u^d\|^2_{L_2(\Omega)}.$$

However, we select

$$J(v) := \|\nabla\left(y(v) - y^d\right)\|^2_{L_2(\Omega,\mathbb{R}^d)} + \alpha\|v - v^d\|^2_{L_2(\Omega)},$$

i.e, our aim is to match the flux of the state variable $y(v)$ with the objective field ∇y^d. If the objective field σ^d is not of the form ∇y^d (it is generated by measurements or otherwise lacks regularity), then the respective cost functional would be

$$J_2(v) := \|\nabla y(v) - \sigma^d\|^2_{L_2(\Omega,\mathbb{R}^d)} + \alpha\|v - v^d\|^2_{L_2(\Omega)}.$$

Accepted Manuscript

It can be easily shifted to the desired form by solving (in practise numerically) the projection problem: Find y^d such that

$$\int_\Omega \left(\nabla y^d - \sigma^d\right) \cdot \nabla q \, \mathrm{d}x = 0, \quad \forall q \in H_0^1(\Omega).$$

Then the objective function values simply are shifted as follows:

$$J(v) = J_2(v) - \|\nabla y^d - \sigma^d\|^2_{L_2(\Omega, \mathbb{R}^d)}.$$

The derivative of J at v is

$$\begin{aligned}\langle J'(v), w\rangle_{\mathsf{U}} &= \lim_{t\to 0_+} \frac{1}{t}\left(J(v+tw) - J(v)\right) = 2\langle \mathscr{B}w, y(v) - y^d\rangle_{\mathsf{V}_0} + (v - u^d, w)_{\mathscr{N}} \\ &= 2(\mathscr{I}_{\mathsf{U}}^{-1}\mathscr{B}^*(y(v) - y^d) + \mathscr{N}(v - u^d), w)_{\mathsf{U}}. \end{aligned} \tag{2.10}$$

The necessary conditions for the optimal control problem (2.9) are (2.7) and

$$\langle J'(u), v - u\rangle_{\mathsf{U}} \geq 0, \quad \forall v \in \mathsf{U}_{\mathrm{ad}} \tag{2.11}$$

(see, e.g., (1, Ch. I, Th. 1.3), (4, Le. 2.21)), i.e.,

$$(\mathscr{I}_{\mathsf{U}}^{-1}\mathscr{B}^*(y(u) - y^d) + \mathscr{N}(u - u^d), v - u)_{\mathsf{U}} \geq 0, \quad \forall v \in \mathsf{U}_{\mathrm{ad}}. \tag{2.12}$$

Note that for the cost functional of type (2.8), there is no need to define an adjoint state to present the necessary conditions (compare (1, Chap. II, Th. 1.4)).

Accepted Manuscript

The following proposition (dating back to (11), see, e.g., (12, Chap. I, Pr. 2.2) or (13, Chap. 7, Pr. 7.4)) allows to write (2.12) in a different form.

Proposition 2.1. *Including the earlier assumptions, let $x \in \mathsf{U}$. Then, the following conditions are equivalent,*

(i) $(u - x, v - u)_{\mathcal{N}} \geq 0, \quad \forall v \in \mathsf{U}_{\mathrm{ad}}$,

(ii) $\|x - u\|_{\mathcal{N}} = \inf_{v \in \mathsf{U}_{\mathrm{ad}}} \|x - v\|_{\mathcal{N}}$,

(iii) $u = \Pi_{\mathrm{ad}}^{\mathcal{N}} x$, *where* $\Pi_{\mathrm{ad}}^{\mathcal{N}} : \mathsf{U} \to \mathsf{U}_{\mathrm{ad}}$ *is a projection.*

Proposition 2.1 and (2.12) yield the well known projection condition

$$u = \Pi_{\mathrm{ad}}^{\mathcal{N}} \left(u^d - \mathcal{N}^{-1} \mathcal{I}_{\mathsf{U}}^{-1} \mathcal{B}^* (y(u) - y^d) \right). \tag{2.13}$$

Remark 2.2. Typical choice is $\mathcal{N} = \alpha \mathrm{Id}$, where $\alpha > 0$ and Id denotes the identity mapping. Then (2.13) becomes

$$u = \Pi_{\mathrm{ad}} \left(u^d - \frac{1}{\alpha} \mathcal{I}_{\mathsf{U}}^{-1} \mathcal{B}^* (y(u) - y^d) \right).$$

Remark 2.3. If $\mathsf{U}_{\mathrm{ad}} = \mathsf{U}$, then $\Pi_{\mathrm{ad}}^{\mathcal{N}} = \mathrm{Id}$ and (2.13) reduces to

$$u = u^d - \mathcal{N}^{-1} \mathcal{I}_{\mathsf{U}}^{-1} \mathcal{B}^* (y(u) - y^d). \tag{2.14}$$

Accepted Manuscript

Requiring that u satisfies also the state equation, i.e. substituting u from (2.14) to (2.7) instead of v yields a following linear problem: Find $y(u) \in \mathsf{V}_0$ satisfying

$$a(y(u), z) + \langle \mathscr{B}\mathscr{N}^{-1}\mathscr{I}_{\mathsf{U}}^{-1}\mathscr{B}^* y(u), z\rangle_{\mathsf{V}_0}$$
$$= \langle \ell + \mathscr{B}u^d, z\rangle_{\mathsf{V}_0} + \langle \mathscr{B}\mathscr{N}^{-1}\mathscr{I}_{\mathsf{U}}^{-1}\mathscr{B}^* y^d, z\rangle_{\mathsf{V}_0} \quad \forall z \in \mathsf{V}_0. \tag{2.15}$$

3. ESTIMATES

3.1. Estimates for the state equation

The solution $y(v) \in \mathsf{V}_0$ of (2.7) minimizes a quadratic energy functional (see, e.g., (1, Chapter I, Theorem 1.2 and Remark 1.5)), i.e.,

$$E(y(v)) \leq E(q) := |||q|||^2 - 2\langle \ell + \mathscr{B}v, q\rangle_{\mathsf{V}_0}, \quad \forall q \in \mathsf{V}_0. \tag{3.1}$$

The benefit for measuring $y(v) - y^d$ in the $|||\cdot|||$-norm in (2.8) (instead of, e.g., $\|\cdot\|_{\mathsf{W}}$-norm) is due to the following results. Theorem 3.1 is due to (14) and generalized in (6). Theorem 3.2 is proved, e.g., in (9, Chap. 6, (6.2.3)), (6, Chap. 7, (7.1.19)). Upper bounds of more general type have been presented already in (15; 16). For the convenience, we repeat both the proofs.

Theorem 3.1. *Let $y(v)$ be the solution of* (3.1) *and $z \in \mathsf{V}_0$ be arbitrary, then*

$$|||y(v) - z|||^2 = E(z) - E(y(v)). \tag{3.2}$$

Accepted Manuscript

Proof. By (2.7),

$$
\begin{aligned}
|||y(v) - z|||^2 &= |||y(v)|||^2 - 2a(y(v), z) + |||z|||^2 \\
&\qquad - 2\left(a(y(v), y(v)) - \langle \ell + \mathscr{B}v, y(v)\rangle_{\mathsf{V}_0}\right) \\
&= -|||y(v)|||^2 - 2\langle \ell + \mathscr{B}v, z\rangle_{\mathsf{V}_0} + |||z|||^2 + 2\langle \ell + \mathscr{B}v, y(v)\rangle_{\mathsf{V}_0} \\
&= E(z) - E(y(v)).
\end{aligned}
$$

□

Theorem 3.2. *Let $y(v)$ be the solution of* (3.1) *and $z \in \mathsf{V}_0$ be arbitrary, then*

$$
\sup_{q \in \mathsf{V}_0} \underline{M}^2(z, q, v) = |||y(v) - z|||^2 = \inf_{\substack{\tau \in \mathsf{Q} \\ \beta > 0}} \overline{M}^2(z, \tau, \beta, v),
$$

where

$$
\underline{M}^2(z, q, v) := E(z) - E(q) \tag{3.3}
$$

and

$$
\overline{M}^2(z, \tau, \beta, v) := (1 + \beta)\|\mathscr{A}\Lambda z - \tau\|^2_{\mathscr{A}^{-1}} + \frac{1 + \beta}{\beta} \| - \Lambda^* \tau + \mathscr{B}v + \ell \|^2. \tag{3.4}
$$

Proof. By (3.1) and (3.2), we have for all $q \in \mathsf{V}_0$

$$
\underline{M}^2(z, q, v) := E(z) - E(q) \le E(z) - E(y(v)) = |||z - y(v)|||^2,
$$

Accepted Manuscript

which proves the lower bound. In order to derive the upper bound, we subtract $a(z, q)$ from both sides of (2.7) and add zero by (2.2)

$$a(y(v) - z, q) = -a(z, q) + \langle \ell + \mathscr{B}v, q\rangle_{\mathsf{V}_0} - \langle \Lambda^*\tau, q\rangle_{\mathsf{V}} + (\tau, \Lambda q)_{\mathsf{H}}.$$

Then, we recall (2.5), and reorganize

$$(\mathscr{A}\Lambda(y(v) - z) - \tau, \Lambda q)_{\mathsf{H}} = -(\mathscr{A}\Lambda z - \tau, \Lambda q)_{\mathsf{H}} + \langle -\Lambda^*\tau + \ell + \mathscr{B}v, q\rangle_{\mathsf{V}_0}$$

To the first term, we apply Cauchy-Schwartz type inequality[1] and to the second term (2.4) to obtain

$$(\mathscr{A}\Lambda(y(v) - z) - \tau, \Lambda q)_{\mathsf{H}} \leq (\|\mathscr{A}\Lambda z - \tau\|_{\mathscr{A}^{-1}} + \| - \Lambda^*\tau + \ell + \mathscr{B}v\|)\, \|\Lambda q\|_{\mathscr{A}}$$

Next, we substitute $q = y(v) - z$, which yields

$$\|\Lambda(y(v) - z)\|_{\mathscr{A}} \leq \|\mathscr{A}\Lambda z - \tau\|_{\mathscr{A}^{-1}} + \| - \Lambda^*\tau + \ell + \mathscr{B}v\|.$$

The form (3.4) follows from squaring both sides and using Young's inequality[2] and (2.6).

□

Remark 3.1. The functionals (3.3) and (3.4) have superscript two, since they are lower and upper bound for the square of the error.

[1] $(\tau, \mu)_{\mathsf{H}} \leq \|\tau\|_{\mathscr{A}}\|\mu\|_{\mathscr{A}^{-1}}, \quad \forall \tau, \mu \in \mathsf{H}$

[2] $(a - b)^2 \leq (1 + \beta)a^2 + \frac{1+\beta}{\beta}b^2, \quad \forall a, b \in \mathbb{R},\ \beta > 0.$

Accepted Manuscript

Note that by (3.2)

$$\underline{M}^2(z, y(v), v) = E(z) - E(y(v)) = |||z - y(v)|||^2.$$

Thus the supremum over $\underline{M}^2$ is obtained at $q = y(v)$.

Substituting $\tau = \mathscr{A}\Lambda y(v)$ to (3.4) causes the second term to vanish, since $y(v)$ is the exact solution of the state equation, i.e., $-\Lambda^*\mathscr{A}\Lambda y(v) + \ell + \mathscr{B}v = 0$. Tending $\beta \to 0$ yields,

$$\lim_{\beta \to 0_+} \overline{M}^2(z, \mathscr{A}\Lambda y(v), \beta, v) = \|\mathscr{A}\Lambda(z - y(v))\|^2_{\mathscr{A}^{-1}} = \|\Lambda(z - y(v))\|^2_{\mathscr{A}} = |||z - y(v)|||^2.$$

The functionals (3.3) and (3.4) are typically used to estimate the error between the exact solution (in our special case related to the state equation) $y(v)$ and an arbitrary conforming (approximation) $z \in \mathsf{V}$. Note that they do not depend on the exact solution and they do not assume any additional information of a particular numerical method generating z. Instead, for example, the upper bound is valid for any $\tau \in \mathsf{Q}$, but in order it to be accurate τ has to be a good approximation of the flux $A\Lambda y(v)$. Thus all algorithms generating approximations of the flux can be used to compute a posteriori error estimates. For more detailed discussion of the performance and implementation of different algorithms, see, e.g., (10).

3.2. Estimates for the cost functional

Applying Theorem 3.2 to the first term of (2.8), leads to two-sided bounds for $J(v)$. Indeed, let y^d play the role of the approximate solution of the state equation. Then, the first term of

Accepted Manuscript

(2.8) represents the "approximation error", which can be estimated from both sides using (3.3) and (3.4). These bounds are guaranteed, have no gap, and do not depend on $y(v)$, i.e., they do not require the solution of the state equation.

Theorem 3.3. *For any* $v \in \mathsf{U}$,

$$\sup_{q \in \mathsf{V}_0} \underline{J}(v, q) = J(v) = \inf_{\substack{\tau \in \mathsf{Q} \\ \beta > 0}} \overline{J}(v, \tau, \beta), \tag{3.5}$$

where

$$\underline{J}(v, q) := \underline{M}^2(y^d, q, v) + \|v - u^d\|_{\mathcal{N}}^2 \tag{3.6}$$

and

$$\overline{J}(v, \tau, \beta) := \overline{M}^2(y^d, \tau, \beta, v) + \|v - u^d\|_{\mathcal{N}}^2. \tag{3.7}$$

Theorem 3.3 can be used to estimate $J(u)$. By (2.9) and (3.5),

$$\inf_{v \in \mathsf{U}_{\mathrm{ad}}} \underline{J}(v, q) \le J(u) \le \overline{J}(v, \tau, \beta), \quad \forall\, q \in \mathsf{V}_0,\ v \in \mathsf{U},\ \tau \in \mathsf{H},\ \beta > 0, \tag{3.8}$$

where all inequalities hold as equalities if and only if $v = u$, $q = y(u)$, $\tau = \mathcal{A}\Lambda y(u)$, and $\beta \to 0$. In view of (3.8), it is very important that the minimizer of $\underline{J}(v, q)$ over $v \in \mathsf{U}_{\mathrm{ad}}$ can

Accepted Manuscript

be explicitly computed. Computation of the minimizers of $\overline{J}$ require further assumptions of the structure of the problem (cf. Propositions 4.1 and 4.2).

Proposition 3.1. *For all $v \in \mathsf{U}_{\mathrm{ad}}$ and $q \in \mathsf{V}_0$,*

$$\underline{J}(\hat{v}(q), q) = \inf_{v \in \mathsf{U}_{\mathrm{ad}}} \underline{J}(v, q), \tag{3.9}$$

$$\underline{J}(v, \hat{q}(v)) = \sup_{q \in \mathsf{V}_0} \underline{J}(v, q),$$

where $\hat{q}(v) = y(v)$ (from (2.7)) and

$$\hat{v}(q) := \Pi_{\mathrm{ad}}^{\mathcal{N}} \left(u^d + \mathcal{N}^{-1} \mathcal{F}_{\mathsf{U}}^{-1} \mathcal{B}^*(y^d - q) \right). \tag{3.10}$$

Proof. The condition $\hat{q}(v) = y(v)$ follows directly from Remark 3.1.

By (3.1), (3.3), and (3.6), $\underline{J}$ has the following form

$$\begin{aligned}\underline{J}(v, q) &= |||y^d|||^2 - 2\langle \ell, y^d \rangle - |||q|||^2 + 2\langle \ell, q \rangle + 2\langle Bv, q - y^d \rangle_{\mathsf{V}_0} + \|v - u^d\|_{\mathcal{N}}^2 \\ &= \|v\|_{\mathcal{N}}^2 - 2(v, u^d)_{\mathcal{N}} - 2\langle \mathcal{B}v, y^d \rangle_{\mathsf{V}_0} - |||q|||^2 + 2\langle \ell, q \rangle + 2\langle \mathcal{B}v, q \rangle_{\mathsf{V}_0} + \text{const.}\end{aligned}$$

Clearly, it is quadratic w.r.t v and the minimizer $\hat{v} \in \mathsf{U}_{\mathrm{ad}}$ is identified by the following variational inequality (see, e.g., (1, Chap. I, Th. 1.2)):

$$(\hat{v}, v - \hat{v})_{\mathcal{N}} \geq (v - \hat{v}, u^d)_{\mathcal{N}} + \langle \mathcal{B}(v - \hat{v}), y^d - q \rangle_{\mathsf{V}_0}, \quad \forall v \in \mathsf{U}_{\mathrm{ad}}.$$

Accepted Manuscript

Applying (2.3) and adding $\mathrm{Id} = \mathcal{N}\mathcal{N}^{-1}$ leads at

$$(\hat{v}, v - \hat{v})_{\mathcal{N}} \geq (v - \hat{v}, u^d)_{\mathcal{N}} + \left((v - \hat{v}), \mathcal{N}\mathcal{N}^{-1}\mathcal{I}_{\mathsf{U}}^{-1}\mathcal{B}^*(y^d - q)\right)_{\mathsf{U}}, \quad \forall v \in \mathsf{U}_{\mathrm{ad}}$$

and reorganizing yields

$$\left(\hat{v} - u^d + \mathcal{N}^{-1}\mathcal{I}_{\mathsf{U}}^{-1}\mathcal{B}^*(q - y^d), v - \hat{v}\right)_{\mathcal{N}} \geq 0, \quad \forall v \in \mathsf{U}_{\mathrm{ad}}.$$

Finally, Proposition 2.1 leads at (3.10). □

Remark 3.2. By (3.5) and (3.8), $\overline{J}(v, \tau, \beta)$ is an upper bound of $J(u)$ for all $v \in \mathsf{U}_{ad}$, $\tau \in \mathsf{Q}$, and $\beta > 0$ and $\underline{J}(v, q)$ is a lower bound for $J(v)$ for all $q \in \mathsf{U}_{ad}$, but it is a lower bound of $J(u)$ only if $v = \hat{v}(q)$ (see (3.10)).

Remark 3.3. Lower bound $\underline{J}$ generates a saddle point formulation for the original optimal control problem (2.9). Find $(\tilde{v}, \tilde{q})$ satisfying

$$\underline{J}(\tilde{v}, q) \leq \underline{J}(\tilde{v}, \tilde{q}) \leq \underline{J}(v, \tilde{q}), \quad \forall v \in \mathsf{U}_{\mathrm{ad}}, q \in \mathsf{V}_0. \tag{3.11}$$

Note that $\underline{J}$ is convex, lower semi-continuous, and coercive w.r.t. v and concave, upper semi-continuous, and anti-coercive w.r.t q, U_{ad} is convex, closed, and non-empty, and V_0 is convex, closed, and non-empty. Thus, the solution of (3.11) exists and is unique (see, e.g., (12, Chap. VI, Pr. 2.4)). By Remark 3.1, $\tilde{v} = u$ and $\tilde{q} = y(u)$. Moreover, $\hat{v}(y(u)) = u$, where $\hat{v}$ is defined in (3.10). The left and right-hand-side of (3.11) yield (3.1) and (2.12) (i.e., necessary conditions (2.7) and (2.11)), respectively.

Accepted Manuscript

Remark 3.4. By (3.8), $J(u) \leq J(v) \leq \overline{J}(v, \tau, \beta)$ and it is easy to see that $J(u) = \lim_{\beta \to 0} \overline{J}(u, \mathscr{A}\Lambda y(u), \beta)$. Thus, the upper bound generates a minimization problem

$$\overline{J}(u, \mathscr{A}\Lambda y(u), 0) = \min_{\substack{v \in \mathsf{U}_{\mathrm{ad}}, \tau \in \mathsf{Q} \\ \beta > 0}} \overline{J}(v, \tau, \beta),$$

where the constraint related to (2.7) does not appear.

3.3. Estimates for an error quantity

The following identity can be viewed as an analog of (3.1) for the optimal control problem. It defines the proper error measure generated by the respective the objective functional. We use it to measure the difference between $v \in \mathsf{U}_{\mathrm{ad}}$ (any admissible control) and u (the optimal control). Moreover, using earlier results, we will estimate this difference from below and above without knowing the optimal control.

Theorem 3.4. *For any* $v \in \mathsf{U}_{\mathrm{ad}}$,

$$|||y(v) - y(u)|||^2 + \|v - u\|_{\mathcal{N}}^2 + \langle J'(u), v - u \rangle_{\mathsf{U}} = J(v) - J(u). \tag{3.12}$$

Proof. We have,

$$\begin{aligned} J(v) - J(u) &= |||y(v) - y^d|||^2 + \|v - u^d\|_{\mathcal{N}}^2 - |||y(u) - y^d|||^2 + \|u - u^d\|_{\mathcal{N}}^2 \\ &= |||y(v)|||^2 - |||y(u)|||^2 + 2a(y(u) - y(v), u^d) \\ &\quad + \|v\|_{\mathcal{N}}^2 - \|u\|_{\mathcal{N}}^2 + 2(u - v, u^d)_{\mathcal{N}}. \end{aligned}$$

Accepted Manuscript

Then, we add and subtract $2|||y(u)|||^2 - 2a(y(u), y(v)) = 2a(y(u) - y(v), y(u))$ and $2\|u\|_{\mathcal{N}}^2 - 2(u, v)_{\mathcal{N}} = 2(u - v, u)_{\mathcal{N}}$ to generate squares and have

$$J(v) - J(u) = |||y(v) - y(u)|||^2 + 2a(y(u) - y(v), y^d) - 2a(y(u) - y(v), y(u)) \\ + \|v - u\|_{\mathcal{N}}^2 + 2(u - v, u^d)_{\mathcal{N}} - 2(u - v, u)_{\mathcal{N}}.$$

Finally reorganizing yields,

$$J(v) - J(u) = |||y(v) - y(u)|||^2 + 2a(y(v) - y(u), y(u) - y^d) \\ + \|v - u\|_{\mathcal{N}}^2 + 2(v - u, u - u^d)_{\mathcal{N}}.$$

By (2.7) and (2.10),

$$a(y(v) - y(u), y(u) - y^d) = \langle \mathscr{B}(v - u), y(u) - y^d \rangle_{\mathsf{V}_0}$$

and

$$2a(y(v) - y(u), y(u) - y^d) + 2(v - u, u - u^d)_{\mathcal{N}} = \langle J'(u), v - u \rangle_{\mathsf{U}}.$$

□

Equality (3.12) shows that it is reasonable to include $\langle J'(u), v - u \rangle_{\mathsf{U}}$ to the applied error measure. Obviously, $\langle J'(u), v - u \rangle_{\mathsf{U}}$ is positive for any $v \in \mathsf{U}_{\mathrm{ad}}$ by (2.11), it is convex and vanishes if $v = u$. Thus, the error measure is

$$\mathrm{err}^2(v) := |||y(v) - y(u)|||^2 + \|v - u\|_{\mathcal{N}}^2 + \langle J'(u), v - u \rangle_{\mathsf{U}}. \tag{3.13}$$

Accepted Manuscript

The "derivative weight" guarantees that the sensitivity of the cost functional at the optimal control is taken into account. Most importantly, $\mathrm{err}(v)$ can be estimated from both sides by computable functionals, which do not require the knowledge of the optimal control u, the respective state $y(u)$, or the exact state $y(v)$. Indeed, applying (3.5), (3.8), and (3.9) to the right hand side of (3.12) yields the following theorem:

Theorem 3.5. *For any* $v \in \mathsf{U}_{\mathrm{ad}}$,

$$\sup_{\substack{q \in \mathsf{V}_0, v_2 \in \mathsf{U}_{\mathrm{ad}}, \\ \tau \in \mathsf{Q}, \beta > 0}} \underline{\mathrm{err}}^2(v, q, v_2, \tau, \beta) = \mathrm{err}^2(v) = \inf_{\substack{\tau_2 \in \mathsf{Q}, \beta_2 > 0, \\ q_2 \in \mathsf{V}_0}} \overline{\mathrm{err}}^2(v, \tau_2, \beta_2, q_2), \tag{3.14}$$

where

$$\underline{\mathrm{err}}^2(v, q, v_2, \tau, \beta) := \underline{J}(v, q) - \overline{J}(v_2, \tau, \beta)$$

and

$$\overline{\mathrm{err}}^2(v, \tau_2, \beta_2, q_2) := \overline{J}(v, \tau_2, \beta_2) - \underline{J}(\hat{v}(q_2), q_2).$$

Theorem 3.5 is of practical value. Assume that after some computations (or even guessing, the estimates do not depend on the applied method) we have obtained $v \in \mathsf{U}_{\mathrm{ad}}$. Naturally, we want to find out how close it is to the optimal control u and select (3.13) as the error measure. The right hand side of (3.14) allows us to estimate

$$\mathrm{err}^2(v) \leq \overline{\mathrm{err}}^2(v, \tau_2, \beta_2, q_2), \quad \forall \tau_2 \in \mathsf{Q}, \beta_2 > 0, q_2 \in \mathsf{V}_0.$$

Accepted Manuscript

It is important that the estimate is guaranteed, i.e., it does not matter which τ_2, β_2, or q_2 we select. The bound is always an upper bound, but particular choices result in sharper bounds. One alternative is to minimize $\overline{\mathrm{err}}^2(v, \tau_2, \beta_2, q_2)$ numerically w.r.t. τ_2, β_2, or q_2. Other options include, e.g., post-processing of v. The lower bound for the error can be treated using similar ideas. The resulting algorithms are discussed in more detail in Sect. 4.

Remark 3.5. By Remark 3.2, (3.6), (3.7), and (3.12), the equality (3.14) is attained at

$$\underline{\mathrm{err}}^2(v, y(v), u, \mathscr{A}\Lambda y(u), 0) = \mathrm{err}^2(v) = \overline{\mathrm{err}}^2(v, \mathscr{A}\Lambda y(v), 0, y(u)).$$

Remark 3.6. Obviously $J(v)$ and $\mathrm{err}^2(v)$ are positive. However, e.g., the lower bound $\underline{J}(\hat{v}(q_2), q_2)$ for $J(u)$ may be negative if q_2 is not close enough to $y(u)$ and $\underline{\mathrm{err}}^2(v, q, v_2, \tau, \beta)$ may be negative value if v_2 is not "good enough" in comparison with v, or the upper bound $\overline{J}(v_2, \tau, \beta)$ is not "sharp enough".

4. EXAMPLES, ALGORITHMS AND NUMERICAL TESTS

4.1. Examples

In this Section, we present some optimal control problems for which the presented error methods can be used, i.e., they belong to the general class of problems defined in Sect. 2. In the following examples, the domain $\Omega \subset \mathbb{R}^d$ is open, bounded, simply-connected, Lipschitz domain with boundary Γ. Spaces are $\mathsf{W} = L^2(\Omega)$, $\mathsf{V} = H^1(\Omega)$, $\mathsf{H} = L^2(\Omega, \mathbb{R}^d)$, and $\mathsf{Q} = H(\mathrm{div}, \Omega)$. Operators are $\Lambda = \nabla$, $\Lambda^* = -\mathrm{div}$, $\mathscr{A} = \mathrm{Id}$, and $N = \alpha\mathrm{Id}$ ($\alpha > 0$). Then $a(q, z) :=$

Accepted Manuscript

$(\nabla q, \nabla z)_{L^2(\Omega,\mathbb{R}^d)}$ and $|||w||| = \|\nabla w\|_{L^2(\Omega,\mathbb{R}^d)}$. The examples differ only by the selection of V_0, U, $\mathscr{B}$, and ℓ.

4.1.1. Dirichlet problem, distributed control

Let $\mathsf{U} := L^2(\Omega)$, $\mathsf{V}_0 := H_0^1(\Omega)$, and $\langle \ell, w\rangle = (f, w)_{L^2(\Omega)}$, where $f \in L^2(\Omega)$. Here, $H_0^1(\Omega)$ denotes the subspace of $H^1(\Omega)$, where functions have zero trace. Moreover, $B = \mathrm{Id}$, i.e., $\langle Bv, q\rangle = (v, q)_{L^2(\Omega)}$. The analog of (2.1) is the Friedrichs inequality

$$\|q\|_{L^2(\Omega)} \le c_\Omega \|\nabla q\|_{L^2(\Omega,\mathbb{R}^d)}, \quad \forall q \in H_0^1(\Omega).$$

The cost functional (2.8) is

$$J(v) := \|\nabla(y(v) - y^d)\|^2_{L^2(\Omega,\mathbb{R}^d)} + \alpha\|v - u^d\|^2_{L^2(\Omega)}. \tag{4.1}$$

The state equation (2.7) is

$$(\nabla y(v), \nabla z)_{L^2(\Omega,\mathbb{R}^d)} = (f + v, z)_{L^2(\Omega)}, \quad \forall z \in H_0^1(\Omega) \tag{4.2}$$

and it has the classical form

$$\begin{cases} -\Delta y(v) = f + v & \text{a.e. in } \Omega, \\ y(v) = 0 & \text{on } \Gamma. \end{cases}$$

The majorant (3.4) is

$$\overline{M}^2(q, \tau, \beta, v) = (1 + \beta)\|\tau - \nabla z\|^2_{L^2(\Omega,\mathbb{R}^d)} + \frac{1+\beta}{\beta} c_\Omega^2 \|\mathrm{div}\tau + f + v\|^2_{L^2(\Omega)}.$$

Accepted Manuscript

The Proposition 3.1 shows how minimizers and maximizers for $\underline{J}(v, q)$ can be computed. Analogously, the following proposition characterizes the minimizers of $\overline{J}(v, \tau, \beta)$. These identities are of practical importance. Assume that $v \in \mathsf{U}_{\mathrm{ad}}$ is given. By Theorem 3.3, we can estimate the exact value of $J(v)$ from above by $\overline{J}(v, \tau, \beta)$, where $\tau \in H(\mathrm{div}, \Omega)$ and $\beta > 0$ can be any. However, in order to obtain as accurate upper bound as possible, we want to identify the particular τ and β for which the upper bound is minimized. Then, we can apply numerical methods to compute approximations of minimizing τ and β (The resulting upper bound is still guaranteed upper bound and the approximation error merely reduces the accuracy of the upper bound).

Proposition 4.1. *For all* $v \in \mathsf{U}_{\mathrm{ad}}$, $\tau \in H(\mathrm{div}, \Omega)$, *and* $\beta > 0$

$$\overline{J}(\hat{v}(\tau, \beta), \tau, \beta) = \inf_{v \in \mathsf{U}_{\mathrm{ad}}} \overline{J}(v, \tau, \beta),$$

$$\overline{J}(v, \hat{\tau}(v, \beta), \beta) = \inf_{\tau \in H(\mathrm{div}, \Omega)} \overline{J}(v, \tau, \beta),$$

$$\overline{J}(v, \tau, \hat{\beta}(v, \tau)) = \inf_{\beta > 0} \overline{J}(v, \tau, \beta),$$

where

$$\hat{v}(\tau, \beta) = \Pi_{\mathrm{ad}} \left(\frac{\alpha\beta}{(1+\beta)c_\Omega^2} u^d - \mathrm{div}\tau - f \right), \tag{4.3}$$

$\hat{\tau} := \hat{\tau}(v, \beta)$ *satisfies*

$$\begin{aligned} &\beta(\hat{\tau}, \xi)_{L^2(\Omega, \mathbb{R}^d)} + c_\Omega^2 (\mathrm{div}\hat{\tau}, \mathrm{div}\xi)_{L^2(\Omega)} \\ &\quad = \beta(\nabla y^d, \xi)_{L^2(\Omega, \mathbb{R}^d)} + c_\Omega^2 (f + v, \mathrm{div}\xi)_{L^2(\Omega)}, \quad \forall \xi \in H(\mathrm{div}, \Omega), \end{aligned} \tag{4.4}$$

Accepted Manuscript

and

$$\hat{\beta}(v,\tau) = \frac{c_\Omega \|\mathrm{div}\tau + f + v\|_{L^2(\Omega)}}{\|\tau - \nabla y^d\|_{L^2(\Omega,\mathbb{R}^d)}}. \tag{4.5}$$

Proof. The upper bound $\overline{J}$ can be rewritten as follows,

$$\begin{aligned}\overline{J}(v,\tau,\beta) &= (1+\beta)\|\tau - \nabla z\|^2_{L^2(\Omega,\mathbb{R}^d)} + \frac{1+\beta}{\beta} c_\Omega^2 \|\mathrm{div}\tau + f + v\|^2_{L^2(\Omega)} + \alpha\|v - u^d\|^2_{L^2(\Omega)} \\ &= \left(\frac{1+\beta}{\beta} c_\Omega^2 + \alpha\right)\|v\|^2_{L^2(\Omega)} - 2\left(\alpha u^d - \frac{1+\beta}{\beta} c_\Omega^2 (\mathrm{div}\tau + f), v\right)_{L^2(\Omega)} \\ &\quad + \text{const w.r.t } v.\end{aligned}$$

Thus, the minimizer $\hat{v} \in \mathsf{U}_{\mathrm{ad}}$ satisfies

$$\left(\frac{1+\beta}{\beta} c_\Omega^2 + \alpha\right)(\hat{v}, w - \hat{v})_{L^2(\Omega)} \geq \left(\alpha u^d - \frac{1+\beta}{\beta} c_\Omega^2 (\mathrm{div}\tau + f), w - \hat{v}\right)_{L^2(\Omega)}, \ \forall w \in \mathsf{U}_{\mathrm{ad}}.$$

Reorganizing leads at

$$(\hat{v} - \frac{\alpha\beta}{(1+\beta)c_\Omega^2} u^d + \mathrm{div}\tau + f, w - \hat{v})_{L^2(\Omega)}, \quad \forall w \in \mathsf{U}_{\mathrm{ad}},$$

and Proposition 2.1 yields (4.3).

Condition (4.4) can be easily derived, since $\overline{M}^2$ is quadratic w.r.t. $\tau \in H(\mathrm{div},\Omega)$ and (4.5) results from solving a one-dimensional minimization problem. □

The relation (3.10) becomes

$$\hat{v}(q) = \Pi_{\mathrm{ad}}\left(u^d + \frac{1}{\alpha}(y^d - q)\right), \tag{4.6}$$

Accepted Manuscript

where $\Pi_{\text{ad}} : L^2(\Omega) \to \mathsf{U}_{\text{ad}}$ is a projection.

Example 4.1. If we do not have any restrictions for the control, i.e., $\mathsf{U}_{\text{ad}} = L^2(\Omega)$, then by (2.15) $y(u) \in H_0^1(\Omega)$ satisfies

$$(\nabla y(u), \nabla z)_{L^2(\Omega,\mathbb{R}^d)} + \frac{1}{\alpha}(y(u), z)_{L^2(\Omega)} = (f + \frac{1}{\alpha}y^d + u^d, z)_{L^2(\Omega)}, \quad \forall z \in H_0^1(\Omega) \tag{4.7}$$

and $\Pi_{\text{ad}} = \text{Id}$ in (4.3) and (4.6).

Moreover, in this (unlikely) special case, we are able to estimate a difference between a given state $z \in H_0^1(\Omega)$ and the state related to the optimal control $y(u)$ as follows:

$$\|\nabla(y(u) - z)\|^2_{L^2(\Omega,\mathbb{R}^d)} + \frac{1}{\alpha}\|y(u) - q\|^2_{L^2(\Omega)} = \inf_{\substack{\tau\in H(\text{div},\Omega),\beta>0,\\ v\in L^2(\Omega,[0,1])}} \overline{M}(z, \tau, \beta, v),$$

where

$$\begin{aligned}\overline{M}(z, \tau, \beta, v) := (1 + \beta)\|\nabla z - \tau\|^2_{L^2(\Omega,\mathbb{R}^d)} \\ + \frac{1+\beta}{\beta}c_\Omega^2\|v\mathscr{R}(z, \tau)\|^2_{L^2(\Omega)} + \alpha\|(1 - v)\mathscr{R}(z, \tau)\|^2_{L^2(\Omega)}\end{aligned}$$

and

$$\mathscr{R}(z, \tau) = \text{div}\tau - \frac{1}{\alpha}z + f + \frac{1}{\alpha}y^d + u^d.$$

This result is mainly a curiosity and the proof (which follows closely the lines of Theorem 3.2, see, e.g., (6, Ch. 4.2), and (10, Ch. 3.2)) is omitted.

Accepted Manuscript

Example 4.2. Let

$$\mathsf{U}_{\text{ad}} = \{v \in L^2(\Omega) \mid \psi_- \leq v \leq \psi_+ \quad \text{a.e in } \Omega\}, \tag{4.8}$$

where $\psi_-, \psi_+ \in L^\infty(\Omega)$. Then the projection operator $\Pi_{\text{ad}} : L^2(\Omega) \to \mathsf{U}_{\text{ad}}$ is

$$\Pi_{\text{ad}} v = \min\left\{\psi_+, \max\{\psi_-, v\}\right\}.$$

Example 4.3. Let

$$\mathsf{U}_{\text{ad}} = \{v \in L^2(\Omega) \mid \|v\|_{L^2(\Omega)} \leq M\},$$

where M is a positive constant. Then the projection operator $\Pi_{\text{ad}} : L^2(\Omega) \to \mathsf{U}_{\text{ad}}$ is

$$\Pi_{\text{ad}} v = \begin{cases} \frac{Mv}{\|v\|_{L^2(\Omega)}} & \text{if } \|v\|_{L^2(\Omega)} > M, \\ v & \text{else} \end{cases}$$

4.1.2. Neumann problem, boundary control

Here, we briefly present the analogous definitions and results for the Neumann problem. The boundary Γ consists of two parts $\Gamma_N \cup \Gamma_D$, where Γ_D has a positive measure. By the trace theorem there exists a bounded linear mapping $\gamma : H^{-1}(\Omega) \to L^2(\Gamma_N)$,

$$\|\gamma q\|_{L^2(\Gamma_N)} \leq c\|q\|_{H^1(\Omega)},$$

Accepted Manuscript

such that $\gamma v = v_{|\Gamma}$ for all $v \in C^1(\bar{\Omega})$. Let $\mathsf{U} := L^2(\Gamma_N)$ and

$\mathsf{V}_0 := V_0 := \{w \in H^1(\bar{\Omega}) \mid w \text{ has zero trace on } \Gamma_D\}$.

Moreover, $\langle Bv, q\rangle = (v, \gamma q)_{L^2(\Gamma_N)}$ and $\langle \ell, q\rangle = (f, q)_{L^2(\Omega)} - (g, \gamma q)_{L^2(\Gamma_N)}$, where $f \in L^2(\Omega)$ and $g \in L^2(\Gamma_N)$.

The cost functional (2.8) is

$$J(v) := \|\nabla(y(v) - y^d)\|^2_{L^2(\Omega)} + \alpha\|v - u^d\|^2_{L^2(\Gamma_N)},$$

and the state equation (3.1) is

$$(\nabla y(v), \nabla q)_{L^2(\Omega,\mathbb{R}^d)} = (f, q)_{L^2(\Omega)} + (g + v, \gamma q)_{L^2(\Gamma_N)}, \quad \forall q \in V_0.$$

It has the classical form

$$\begin{cases} -\Delta y(v) = f & \text{a.e. in } \Omega, \\ y(v) = 0 & \text{on } \Gamma_D, \\ \dfrac{\partial y(v)}{\partial n} = g + v & \text{on } \Gamma_N. \end{cases}$$

The majorant (3.4) has the form (see, e.g., (6, Sect. 4.1) for details)

$$\overline{M}^2(q, \tau, \beta) = (1 + \beta)\|\tau - \nabla q\|^2_{L^2(\Omega,\mathbb{R}^d)} + \frac{1+\beta}{\beta}\left(c^2_{\Omega,2}\|\mathrm{div}\tau + f\|^2_{L^2(\Omega)} + c^2_{\Gamma_N}\|\frac{\partial \tau}{\partial n} + g + v\|^2_{L^2(\Gamma_N)}\right),$$

Accepted Manuscript

where constants satisfy

$$\|q\|_{L^2(\Omega)} \le c_{\Omega,2}\|\nabla q\|_{L^2(\Omega,\mathbb{R}^2)} \quad \text{and} \quad \|q\|_{L^2(\Gamma_N)} \le c_{\Gamma_N}\|\nabla q\|_{L^2(\Omega,\mathbb{R}^2)}, \quad \forall q \in V_0.$$

Proposition 4.2. *For all* $q \in H_0^1(\Omega)$, $\tau \in H(\mathrm{div}, \Omega)$, *and* $\beta > 0$

$$\overline{J}(q, \hat{\tau}, \beta) = \inf_{\tau \in H(\mathrm{div},\Omega)} \overline{J}(q, \tau, \beta),$$

$$\overline{J}(q, \tau, \hat{\beta}) = \inf_{\beta > 0} \overline{J}(q, \tau, \beta),$$

where $\hat{\tau}$ *satisfies*

$$\beta(\hat{\tau}, \xi)_{L^2(\Omega,\mathbb{R}^d)} + c_\Omega^2(\mathrm{div}\hat{\tau}, \mathrm{div}\xi)_{L^2(\Omega)} + c_{\Gamma_N}^2(\frac{\partial\hat{\tau}}{\partial n}, \frac{\partial\xi}{\partial n})_{L^2(\Gamma_N)}$$
$$= \beta(\nabla q, \xi)_{L^2(\Omega,\mathbb{R}^d)} + c_\Omega^2(f + v, \mathrm{div}\xi)_{L^2(\Omega)} + c_{\Gamma_N}^2(g + v, \frac{\partial\xi}{\partial n})_{L^2(\Gamma_N)},$$
$$\forall \xi \in H(\mathrm{div}, \Omega)$$

and

$$\hat{\beta} = \frac{\left(c_{\Omega,2}^2\|\mathrm{div}\tau + f\|_{L^2(\Omega)}^2 + c_{\Gamma_N}^2\|\frac{\partial\tau}{\partial n} + g + v\|_{L^2(\Gamma_N)}^2\right)^{1/2}}{\|\tau - \nabla q\|_{L^2(\Omega,\mathbb{R}^d)}}.$$

4.2. Algorithms

The results of Sect. 3 give grounds for several error estimation Algorithms. Note that the estimates in Theorems 3.3 and 3.5 are valid for any approximations from U_{ad}. There is no need for Galerkin orthogonality, extra regularity, or mesh dependent data. Thus they can be combined with any existing numerical scheme, which generates approximations of the

Accepted Manuscript

optimal control (and/or state). Computation of the derived estimates requires some finite dimensional subspaces. Hereafter, assume that $\mathsf{U}_{\mathrm{ad}}^h \subset \mathsf{U}_{\mathrm{ad}}$ $\mathsf{V}_0^h \subset \mathsf{V}_0$ and $\mathsf{Q}^h \subset \mathsf{Q}$ are given. They can be generated, e.g., by finite elements or Fourier series. The approximate solution of (2.7) is denoted by $y^h(v) \in \mathsf{V}_0^h \subset \mathsf{V}_0$. It satisfies

$$a(y^h(v), z) = \langle Bv + \ell, z\rangle_{\mathsf{V}_0}, \quad \forall z \in \mathsf{V}_0^h. \tag{4.9}$$

Remark 4.1. By Remark 3.2, the evaluation of (the approximation of) $J(v)$ by computing $y^h(v)$ from (4.9) and $J_h(v) := |||y^h(v) - y^d|||^2 + \|v - u^d\|_{\mathcal{N}}^2$ coincides with the lower bound $\underline{J}(v, y^h(v)) = \max_{y \in \mathsf{V}_0^h} \underline{J}(v, y)$.

The generation of the estimates for the cost function value $J(v)$ for a given approximation $v \in \mathsf{U}_{\mathrm{ad}}$ is depicted as Algorithm 1. It takes into account the fact that the approximation error of solving the state equation is encompassed into the computation $J(v)$. As a practical "side product"', Algorithm 1generates a lower bound for the optimal cost functional value $J(u)$.

Accepted Manuscript

Algorithm 1 Generation of bounds for the cost functional value

input: $v \in \mathsf{U}_{\mathrm{ad}}$ {approximation of the control} V_0^h {subspace for state}, Q^h {subspace for the flux of state}, $I_{\max}$ {maximum number of iterations}, ε {stopping criteria}

$y^h = \underset{y \in \mathsf{V}_0^h}{\operatorname{argmax}}\, \underline{J}(v, y)$ {compute $y^h(v)$ from (4.9)}
$\hat{v}^h = \underset{v \in \mathsf{U}_{\mathrm{ad}}}{\operatorname{argmin}}\, \underline{J}(v, y^h)$ {compute $\hat{v}(y^h)$ by (3.10)}
$\beta^0 = 1$
for $k = 1$ **to** $I_{\max}$ **do**
  $\tau^k = \underset{\tau \in \mathsf{Q}^h}{\operatorname{argmin}}\, \overline{J}(v, \tau, \beta^{k-1})$
  $\beta^k = \underset{\beta > 0}{\operatorname{argmin}}\, \overline{J}(v, \tau^k, \beta)$
  if $\frac{\overline{J}(v,\tau^{k-1},\beta^{k-1}) - \overline{J}(v,\tau^k,\beta^k)}{\overline{J}(v,\tau^k,\beta^k)} < \varepsilon$ **then**
    break
  end if
end for
$\underline{J}_h(v) = \underline{J}(v, y^h)$
$\overline{J}_h(v) = \overline{J}(v, \tau^k, \beta^k)$
$\underline{J}_h(u) = \underline{J}(\hat{v}^h, y^h)$

output: $\underline{J}_h(u)$ {lower bound for $J(v)$}, $\overline{J}_h(v)$ {upper bound for $J(v)$}, $\underline{J}_h(u)$ {lower bound for $J(u)$},

First step in Algorithm 1 is related to the computation of lower bounds for $J(v)$, which requires a numerical solution to the state equation (4.9). This means solving a single boundary value problem numerically, e.g., by the finite element method. The second step is produces a lower bound for $J(u)$ and requires computation of the projection (3.10). The upper bound for $J(v)$ is computed iteratively by minimizing $\overline{J}(v, \tau, \beta)$ over τ and keeping β fixed, then switching the roles and minimizing, and repeating until maximum amount of iteration steps are taken or $\overline{J}$ converges. The minimization over β can be done algebraically, but the minimization over τ leads to solving a linear PDE using suitable numerical methods. Examples of the minimization steps for the Dirichlet problem are (4.4) (the weak form of the respective PDE resulting from the minimization w.r.t. τ) and (4.5) (algebraic expression for the minimizing β).

Accepted Manuscript

In order to test the presented error estimates, a projected gradient method (see, e.g., (17; 18)) is applied to generate a sequence approximations. Method consists of line searches along (anti)gradient directions, where all evaluated points are first projected to the admissible set. The method was chosen for its simplicity to highlight the use of derived error estimates. A projected gradient method with error estimates is depicted as Algorithm 2. At the beginning of every projected gradient step Algorithm 1is used to generate estimates of the cost functional values of the respective approximation. After the execution of Algorithm 2(N iteration steps taken), cost estimates are recalled to generate two-sided estimates for $\mathrm{err}(v)$ (By (3.12) and (3.13) we practically estimate the difference $J(v) - J(u)$) at each iteration step ($k = 1, \ldots, N$) as follows:

$$\mathrm{err}^2(v^k) \geq \underline{J}_h^2(v^k) - \overline{J}_h(v^N)$$

$$\mathrm{err}^2(v^k) \leq \overline{J}_h(v^k) - \underline{J}_h^N(u)$$

Note that the iterate of the last step (N'th step) is used to generate as accurate bounds as possible for $J(u)$.

Accepted Manuscript

Algorithm 2 Projected gradient method with guaranteed cost estimates

input: $v^0 \in \mathsf{U}_{\mathrm{ad}}$ {initial approximation of the control} V_0^h {subspace for state}, Q^h {subspace for the flux of state}, $I_{\max}^{PG}$ {maximum number of iterations (projected gradient)}, ε^{PG} {stopping criteria (projected gradient)} $I_{\max}$ {maximum number of iterations ($\overline{J}$ minimization)}, ε {stopping criteria ($\overline{J}$ minimization)}

for $k = 0$ **to** $I_{\max}^{\mathrm{PG}}$ **do**
  $\left\{\underline{J}_h(v^k),\, \overline{J}_h(v^k),\, \underline{J}_h^k(u)\right\} = \mathrm{GenerateCostEstimates}(v^k, \mathsf{V}_0^h, \mathsf{Q}^h, I_{\max}, \varepsilon)$
  $d^k = 2\left(B^*(y^d - y(v^k)) + \mathcal{N}(u^d - v^k)\right)$ {search direction}
  $s^k = \underset{\lambda \in [0, \lambda_{\max}]}{\operatorname{argmin}}\ J_h\left(\Pi_{\mathrm{ad}}\left(v^k + \lambda d(v^k)\right)\right)$ {step length (golden section method)}
  $v^{k+1} = v^k + s^k d^k$ {update approximation}
  if $\frac{\|v^k - v^{k-1}\|}{\|v^{k-1}\|} < \varepsilon^{\mathrm{PG}}$ **then**
    break
  end if
end for

output: $\{v^k)\}_{k=1}^N$ {sequence of approximations}, $\{\underline{J}_h(v^k)\}_{k=1}^N$ {lower bounds for $J(v^n)$}, $\{\overline{J}_h(v^k)\}_{k=1}^N$ {upper bounds for $J(v^n)$}, $\underline{J}_h^N(u)$ {lower bound for $J(u)$}

4.3. Numerical tests

Finite dimensional subspaces are generated by the finite element method (see, e.g., (19)). In these tests, $\mathsf{U} = L^2(\Omega)$, $\mathsf{V}_0 = H_0^1(\Omega)$, and $\mathsf{Q} = H(\mathrm{div}, \Omega)$. Subspaces $DG_h^p \subset L^2(\Omega)$, $V_h^p \subset H_0^1(\Omega)$, and $\mathrm{RT}^p \subset H(\mathrm{div}, \Omega)$ are generated by Discontinous Galerkin elements, Lagrange elements, and Raviart-Thomas elements, respectively. Superscripts p denote the order of basis functions. All the numerical tests were performed using FEniCS (see (20, Ch. 3) for detailed descriptions of the applied elements and for additional references).

Example 4.4. Let $\Omega = (0, 1)^2$. Consider the optimal control problem generated by (4.1), (4.2), and U_{ad} defined by (4.8), where $\psi_-(x_1, x_2) = -3$ and $\psi_+(x_1, x_2) = 3$. Select

$$y(x_1, x_2) = \sin(k_1 \pi x_1)\sin(k_1 \pi x_2),$$

$$y^d(x_1, x_2) = \sin(k_1 \pi x_1)\sin(k_1 \pi x_2) + \beta \sin(m_1 \pi x_1)\sin(m_1 \pi x_2),$$

Accepted Manuscript

$$u^d(x_1, x_2) = 0$$

$$u(x_1, x_2) = \max\left\{\psi_-(x_1, x_2), \min\left\{\psi_+(x_1, x_2), \frac{\beta}{\alpha}\sin(m_1\pi x_1)\sin(m_1\pi x_2)\right\}\right\}$$

$$f(x_1, x_2) = \pi^2(k_1^2 + k_2^2)\sin(k_1\pi x_1)\sin(k_1\pi x_2) - u(x_1, x_2),$$

where $k_1, k_2, m_1, m_2 \in \mathbb{Z}$ and $\beta \in \mathbb{R}$.

In Example 4.4, select $k_1 = 1$, $k_2 = 1$, $m_1 = 2$, $m_2 = 1$, $\beta = 0.5$, and $\alpha = 0.05$. A mesh of 50×50 cells divided to triangular elements is being used. Consider first linear elements, i.e., $p_1 = p_2 = p_3 = 1$, the amount of corresponding global degrees of freedom are $\dim(\mathrm{DG}_h^1) = 15000$, $\dim(V_h^1) = 2601$, and $\dim(\mathrm{RT}_h^1) = 7600$. The bounds generated by Algorithm 2($I_{\max}^{PG} = 10$) are depicted in Fig. 1. In Figure 1 (a), it is shown how the estimates for the cost functional values behave during the projected gradient iteration and in Figure 1 (b), we can observe how the error quantity $J(v^k) - J(u)$ decreases during the iteration and how the error estimates bound it from both sides.

If the order of approximation for state and flux are increased, i.e., subspaces V_h and Q_h are enhanced, then the error bounds improves significantly. Comparison of Fig. 1 (b) and Fig. 2 (b) shows clearly how the accuracy of error bounds improves. The approximation error is same in both pictures, since the approximation accuracy (applied finite element space) for control was not changed. Here $\dim(V_h^2) = 10201$ and $\dim(\mathrm{RT}_h^2) = 25200$. As expected, more computational effort (more degrees of freedom for flux and state) resulted in more accurate bounds. This methodology lets numerical analyst to balance between

Accepted Manuscript

the accuracy of the error estimates and the computational cost. Moreover, it may not be sensible to compute error estimates on every iteration (in our case, projected gradient method) step, since the computation of error estimates (as done here) requires additional numerical solutions to boundary value problems. However, this cost can be reduced by using (local) post-processing methods (see, e.g., (10)). In previous examples, $J(v)$ and $J(u)$ (i.e. evaluations of integrals involving the analytical solution, which provided the reference values for the exact error) were computed using a uniformly refined mesh and 121 integration points in each triangle.

Obviously, the negative lower bound for the error could be rejected immediately. Sharp lower bound requires a very good approximation of the optimal control $v \approx u$ and the corresponding flux of the respective state $\tau \approx \nabla y(u)$. Then the upper bound $J(u) \leq J(v) \leq \overline{J}(v, \tau, \beta)$ would be very efficient. These numerical experiments show the main property of functional type estimates, they are guaranteed regardless of the accuracy of the approximate solution. This research yields grounds for rich development, e.g., combination of the estimates with subspaces generated by isogeometric elements (see (21)) or the use of estimates as indicators to generate adaptive methods.

REFERENCES

[1] J.-L. Lions (1971). *Optimal Control of Systems Governed by Partial Differential Equations*. Translated from the French by S. K. Mitter. Die Grundlehren der mathematischen Wissenschaften, Band 170. Springer-Verlag, New York.

Accepted Manuscript

[2] K. Ito and K. Kunisch (2008). *Lagrange Multiplier Approach to Variational Problems and Applications*. Advances in Design and Control, Vol. 15. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

[3] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich (2009). *Optimization with PDE Constraints*. Mathematical Modelling: Theory and Applications, Vol. 23. Springer, New York.

[4] F. Tröltzsch (2010). *Optimal Control of Partial Differential Equations*. Graduate Studies in Mathematics, Vol. 112. American Mathematical Society, Providence, RI. Theory, Methods and Applications, Translated from the 2005 German Original by Jürgen Sprekels.

[5] A. Gaevskaya, R. W. H. Hoppe, and S. Repin (2007). A posteriori error estimation for elliptic optimal control problems with distributed control. *J. Math. Sci. (N. Y.)* 144:4535–4547.

[6] S. Repin (2008). *A Posteriori Estimates for Partial Differential Equations*. Radon Series on Computational and Applied Mathematics, Vol. 4. Walter de Gruyter GmbH & Co. KG, Berlin.

[7] M. Wolfmayr (2015). Functional a posteriori estimates for elliptic optimal control problems. *Proc. Appl. Math. Mech.* 15:621–622.

[8] U. Langer, S. Repin, and M. Wolfmayr (2015). Functional a posteriori error estimates for time periodic parabolic optimal control problems. Technical Report 38, Johann Radon Institute for Computational and Applied Mathematics.

Accepted Manuscript

[9] P. Neittaanmäki and S. Repin (2004). *Reliable Methods for Computer Simulation, Error Control and a Posteriori Estimates*. Elsevier, New York.

[10] O. Mali, P. Neittaanmäki, and S. Repin (2014). *Accuracy Verification Methods*. Computational Methods in Applied Sciences, Vol. 32. Springer, Dordrecht. Theory and Algorithms.

[11] J.-J. Moreau (1965). Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France* 93:273–299.

[12] I. Ekeland and R. Temam (1976). *Convex Analysis and Variational Problems*. North–Holland, New York.

[13] F. Clarke (2013). *Functional Analysis, Calculus of Variations and Optimal Control*. Graduate Texts in Mathematics, Vol. 264. Springer, London.

[14] S. G. Mikhlin (1964). *Variational Methods in Mathematical Physics*. Translated by T. Boddington; editorial introduction by L. I. G. Chambers. A Pergamon Press Book. The Macmillan Co., New York.

[15] S. Repin (1997). A posteriori estimates for approximate solutions of variational problems with strongly convex functionals. *Problems of Mathematical Analysis* 17:199–226.

[16] S. Repin (2000). A posteriori error estimation for variational problems with uniformly convex functionals. *Math. Comp.* 69(230):481–500.

[17] W. A. Gruver and E. Sachs (1981). *Algorithmic Methods in Optimal Control*. Research Notes in Mathematics, Vol. 47. Pitman (Advanced Publishing Program), Boston, Mass.-London.

Accepted Manuscript

[18] C. T. Kelley (1999). *Iterative Methods for Optimization*. Frontiers in Applied Mathematics, Vol. 18. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

[19] P. G. Ciarlet (1978). *The Finite Element Method for Elliptic Problems*. Studies in Mathematics and Its Applications, Vol. 4. North-Holland Publishing Co., Amsterdam.

[20] A. Logg, K.-A. Mardal, G. N. Wells, et al. (2012). *Automated Solution of Differential Equations by the Finite Element Method*. Springer.

[21] S. K. Kleiss and S. K. Tomar (2015). Guaranteed and sharp a posteriori error estimates in isogeometric analysis. *Comput. Math. Appl.* 70(3):167–190.

Accepted Manuscript

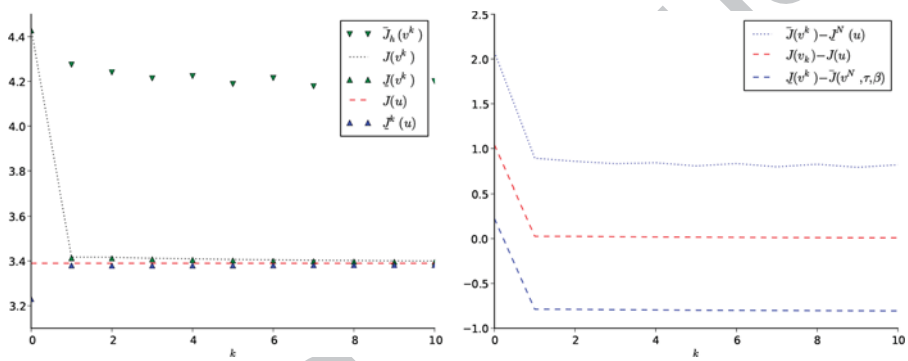

Figure 1: Estimates for the cost function value (a) and the error quantity (b), where subspaces for control, state, and flux are DG_h^1, V_h^1, and RT_h^1, respectively.

Accepted Manuscript

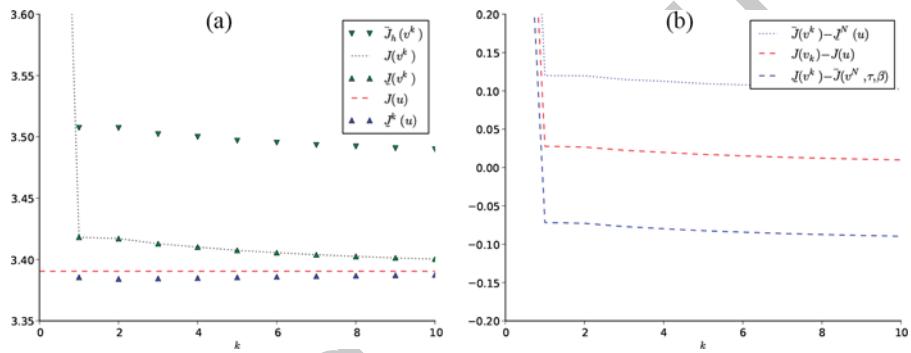

Figure 2: Estimates for the cost function value (a) and the error quantity (b), where subspaces for control, state, and flux are DG_h^1, V_h^2, and RT_h^2, respectively.

Accepted Manuscript