No. 602 **April 2019**

Quasi-best approximation in optimization with PDE constraints

F. Gaspoz, C. Kreuzer, A. Veeser, W. Wollner

ISSN: 2190-1767

QUASI-BEST APPROXIMATION IN OPTIMIZATION WITH PDE CONSTRAINTS

FERNANDO GASPOZ, CHRISTIAN KREUZER, ANDREAS VEESER, AND WINNIFRIED WOLLNER

Abstract. We consider finite element solutions to quadratic optimization problems, where the state depends on the control via a well-posed linear partial differential equation. Exploiting the structure of a suitably reduced optimality system, we prove that the combined error in the state and adjoint state of the variational discretization is bounded by the best approximation error in the underlying discrete spaces. The constant in this bound depends on the inverse square-root of the Tikhonov regularization parameter. Furthermore, if the operators of control-action and observation are compact, this quasi-best-approximation constant becomes independent of the Tikhonov parameter as the meshsize tends to 0 and we give quantitative relationships between meshsize and Tikhonov parameter ensuring this independence. We also derive generalizations of these results when the control variable is discretized or when it is taken from a convex set.

1. Introduction

Optimization problems with PDE constraints are ubiquitous. A basic, and regularly considered, example is

$$\min_{(q,u)\in L^2\times H^1_0} \frac{1}{2}\,|u-u_d|_0^2+\frac{\alpha}{2}\,|q|_0^2 \quad \text{subject to} \quad -\Delta u = q \tag{1.1}$$

where $|\cdot|_0$ denotes the L^2-norm over some underlying domain, u_d is the desired state and $\alpha>0$ scales the cost of the control. Additionally, constraints on the control q and/or the state u can be added, and the error due to a discretization of the state equation, and possibly the control, have been analyzed. For piecewise constant discretizations of the control this has been done in [9, 12] including possible box-constraints on the control variable, see also the summary of obtainable convergence orders including Neumann-control in [16]. The consideration of element wise linear functions for the control has been done in [3, 21] in the presence of control constraints.

In [14] it was observed, that the minimization problem could be solved without prescribing a discretization of the control since the control can be recovered from the optimality condition and thus a discretization of the control is induced by the discretization for the state equation. With this $O(h^2)$ convergence for the control in

(Fernando Gaspoz) Technische Universität Dortmund, Fakultät für Mathematik, Vogelpothsweg 87, 44227 Dortmund, Germany.

(Christian Kreuzer) Technische Universität Dortmund, Fakultät für Mathematik, Vogelpothsweg 87, 44227 Dortmund, Germany.

(Andreas Veeser) Dipartimento di Matematica 'F. Enriques', Università degli Studi di Milano, Via C. Saldini, 50, 20133 Milano, Italy.

(Winnifried Wollner) Technische Universität Darmstadt, Fachbereich Mathematik, Dolivostr. 15, 64293 Darmstadt, Germany.

E-mail addresses: fernando.gaspoz@tu-dortmund.de, christian.kreuzer@tu-dortmund.de, andreas.veeser@unimi.it, wollner@mathematik.tu-darmstadt.de.

Date: April 15, 2019.

L^2 could be shown even in the presence of box control-constraints. It was observed by [18] that the same convergence order can be obtained if a discretized control is used and a post-processing step based upon the optimality conditions is applied.

Due to the structure of the objective in (1.1) these above mentioned estimates make use of the 'natural norm'

$$|u|_0 + \sqrt{\alpha}\,|q|_0\,.$$

Although this norm is natural due to the functional, it induces a scaling $\sqrt{\alpha}$ in all estimates involving the control. Further estimates, for instance of H^1-norms of the state thereby also contain this scaling. Moreover, the above 'natural norm' is not balanced in terms of approximation accuracy, i.e., the error of the state in L^2 will typically decay at least as fast as the error of the control.

The later effect, however, is invisible as long as the approximation accuracy of both terms is limited by the selected discrete spaces, and not by the regularity of the solutions, as it is typically the case for the model (1.1). However, in the presence of pointwise constraints on the state, see, e.g., [2, 7, 8, 17, 19] or the gradient of the state [6, 13, 20, 25] optimal order estimates can only be obtained for the control variable; while numerics shows a faster convergence of the error in the state variable in L^2.

As an alternative to the aforementioned works, one may combine the error in the state with error in the (suitably rescaled) adjoint state, measuring both in the norms that are given by the functional analytic set-up of the PDE constraint. For problem (1.1), this leads to the norm

$$\|x\|^2 := |u|_1^2 + \frac{1}{\alpha}\,|p|_1^2\,, \quad x = (u, z), \tag{1.2}$$

where $|\cdot|_1$ denotes the H_0^1-norm. For respective counterparts of (1.2), Chrysafinos and Karatzas [4, 5] prove so-called symmetric error estimates or quasi-best approximation results. The growth of the quasi-best-approximation constant is limited by α^{-2} and $\alpha^{-3/2}$, respectively.

In this article, we prove abstract quasi-best approximation results, where the discretization error is measured in a counterpart of (1.2). In order to illustrate our results, assume that the underlying domain is convex, let $(V_h)_h$ be a sequence of conforming finite-dimensional spaces that approximates H_0^1, and consider the variational discretization of (1.1). If we denote by $x_h = (u_h, p_h)$ the pairs of approximate primal and dual states, our results yield (cf. Theorem 3.2 and Example 3.8)

$$\|x - x_h\| \le \nu_h \inf_{v_h \in V_h \times V_h} \|x - v_h\|$$

with

$$\nu_h \le \kappa_\alpha := 2\left(1 + C_F\left(1 + \frac{2C_F}{\sqrt{\alpha}}\right)\right) \quad \text{and} \quad |\nu_h - 1| \le C_{\mathcal{I}}\kappa_\alpha h \quad \text{as } h \to 0.$$

Here C_F is the constant in the Friedrichs inequality and $C_{\mathcal{I}}$ is an interpolation constant depending on the shape regularity on the underlying meshes. In contrast to the first, non-asymptotic relationship, the second, asymptotic one exploits the compactness of the observation and control-action operators and elliptic regularity theory. Notably, the latter reveals that Céa's lemma, which holds for the constraint discretization, is recovered as $h \to 0$ and, in particular, ensures an approximation quality independent of α for $h = O(\sqrt{\alpha})$.

The rest of the paper proceeds as follows. In Section 2, we state precisely the considered problem class, allowing for any linear, bounded, and inf-sup-stable operator in the constraint. Furthermore, we reduce the optimality system by eliminating the control, and we lay the groundwork for our results by a careful discussion of the continuity and nondegeneracy properties of the associated bilinear form.

Section 3 constitutes the core of this work and establishes quasi-best approximation for the variational discretization. To this end, the variational discretization is viewed as a Petrov-Galerkin method and we employ the formula for the quasi-best-approximation constant in Tantardini and Veeser [23]. For the asymptotic behavior of the quasi-best-approximation constant, we additionally invoke a duality argument, which is similar to, but simpler than, Schatz [22].

The last two sections center on generalizations of these results. In Section 4, we consider approximate control-action operators, covering in particular the discretization of the control variable. Finally, Section 5, deals with nonlinear optimality systems arising from additional convex constraints for the control. The derived results complement those of the linear case and the simplification of Schatz' argument comes in quite useful.

2. Model optimization problem and reduced optimality system

We introduce our model optimization problem. Assume that the control variable q is taken from a real Hilbert space Q with scalar product $(\cdot,\cdot)_Q$ and induced norm $\|\cdot\|_Q$. Its corresponding state $u \in V_1$ is determined by solving a linear boundary value problem of the form

$$Au = Cq \tag{2.1}$$

with the following setting:

- The *state space* V_1 is a Hilbert space with scalar product $(\cdot,\cdot)_1$ and induced norm $\|\cdot\|_1$. Its dual and the corresponding duality pairing are indicated with V_1^* and $\langle\cdot,\cdot\rangle_1$, respectively.
- The *differential operator* A is induced by bilinear form $a\colon V_1 \times V_2 \to \mathbb{R}$, where V_2 is a second Hilbert space with scalar product $(\cdot,\cdot)_2$, induced norm $\|\cdot\|_2$, dual space V_2^*, and dual pairing $\langle\cdot,\cdot\rangle_2$. We assume that the bilinear form a is bounded and satisfies the following inf-sup conditions:

$$M_a := \sup_{\|v_1\|_1=1,\, \|v_2\|_2=1} a(v_1,v_2) < \infty, \tag{2.2a}$$

$$\forall v_1 \in V_1 \quad \Big(\forall v_2 \in V_1 \; a(v_1,v_2) = 0\Big) \implies v_1 = 0. \tag{2.2b}$$

$$m_a := \inf_{\|v_2\|_2=1} \sup_{\|v_1\|_1=1} a(v_1,v_2) > 0, \tag{2.2c}$$

  Employing well-known inf-sup theory (cf., e.g., Babuška [1]), we see that the operator $A\colon V_1 \to V_2^*, v_1 \mapsto a(v_1,\cdot)$ is linear and boundedly invertible.
- The *control-action operator* $C\colon Q \to V_2^*$ is linear and bounded with constant M_C.

Our goal is then to numerically solve the *constrained optimization problem*

$$\min_{(q,u)\in Q\times V_1} \frac{1}{2}\|Iu - u_d\|_W^2 + \frac{\alpha}{2}\|q\|_Q^2 \quad \text{subject to} \quad Au = Cq \tag{2.3}$$

where we assume in addition:

- The *desired "state"* u_d is an element of a Hilbert space W with scalar product $(\cdot,\cdot)_W$ and induced norm $\|\cdot\|_W$.
- The *observation operator* $I\colon V_1 \to W$ is linear, and bounded with constant M_I.
- The *cost of the control*, which can be viewed as a Tikhonov regularization, is scaled with the parameter $\alpha > 0$.

Problem (2.3) is a quadratic minimization problem with a linear constraint. The objective function is convex with respect to

$$(u, q) \mapsto (\|Iu\|_W^2 + \alpha \|q\|_Q^2)^{1/2} \tag{2.4}$$

and strictly convex in q. Consequently, standard arguments ensure the existence of a unique solution; see, e.g., Lions [15, Theorem 1.1] or Tröltzsch [24, Chapter 2.5].

If $Q = L^2 = W$, $V_1 = V_2 = H_0^1$, $A = -\Delta$ is the (weak) Laplacian, and C and I are the canonical compact immersions $L^2 \to \left(H_0^1\right)^*$ and $H_0^1 \to L^2$, then (2.3) simplifies to the optimization problem (1.1) in the introduction. Notice that, in this case, the operators C and I are related by $C^* = I$.

To formulate the *optimality system* for (2.3), it is useful to define the adjoint operators A^*, C^*, I^* of A, C, I by

$$A^* v_2 = a(\cdot, v_2), \quad (q, C^* v_2)_Q = \langle Cq, v_2 \rangle, \quad \langle I^* w, v_1 \rangle_1 = (Iv_1, w)_W$$

for all $v_1 \in V_1$, $v_2 \in V_2$, $q \in Q$, $w \in W$. Thanks to the convexity of the problem (2.3), a pair $(q, u) \in Q \times V_1$ is a minimum point if and only if there exists $p \in V_2$ such that

$$Au = Cq, \quad A^* p = I^*(Iu - u_d), \quad \alpha q = -C^* p. \tag{2.5}$$

We may eliminate q by inserting the last equation into the first one and multiplying the second equation by $\beta > 0$. We thus obtain the following *reduced optimality system* for the pair $(u, p) \in V_1 \times V_2$:

$$\begin{pmatrix} -\beta I^* I & \beta A^* \\ A & \frac{1}{\alpha} CC^* \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} -\beta I^* u_d \\ 0 \end{pmatrix}. \tag{2.6}$$

Notice that the second row of equations, $Au + \frac{1}{\alpha} CC^* p = 0$, suggests scaling the adjoint state p by the factor $\frac{1}{\alpha}$, while the first row, $-\beta I^* Iu + \beta A^* p = -\beta I^* u_d$, suggests no scaling at all. As a compromise, we propose to use $z = \frac{1}{\sqrt{\alpha}} p$ and $\beta = \frac{1}{\sqrt{\alpha}}$.

We thus transform the optimality system (2.5) into

$$Au = Cq, \quad A^* z = \tfrac{1}{\sqrt{\alpha}} I^*(Iu - u_d), \quad \sqrt{\alpha} q = -C^* z \tag{2.7}$$

and the reduced optimality system (2.6) into

$$\begin{pmatrix} -\frac{1}{\sqrt{\alpha}} I^* I & A^* \\ A & \frac{1}{\sqrt{\alpha}} CC^* \end{pmatrix} \begin{pmatrix} u \\ z \end{pmatrix} = \begin{pmatrix} -\frac{1}{\sqrt{\alpha}} I^* u_d \\ 0 \end{pmatrix}. \tag{2.8}$$

This *rescaled and reduced optimality system* deviates from the usual KKT-formulation, but has an interesting structure. As the KKT-formulation, it is symmetric also for non-symmetric A. The off-diagonal consists of two interrelated invertible operators, while the diagonal entries are (semi-)definite, symmetric operators. Notice that, upon inverting the rows, the roles of the diagonal and off-diagonal can be exchanged. For the optimization problem (1.1), the operator matrix is then diagonally dominant in that CC^* and $I^* I$ are compact operators.

Let us give a weak formulation of the rescaled and reduced optimality system. Its rows are equivalently written as

$$\forall \varphi_1 \in V_1 \qquad a(\varphi_1, z) - \tfrac{1}{\sqrt{\alpha}} (Iu, I\varphi_1)_W = -\tfrac{1}{\sqrt{\alpha}} (u_d, I\varphi_1)_W, \tag{2.9a}$$

$$\forall \varphi_2 \in V_2 \qquad a(u, \varphi_2) + \tfrac{1}{\sqrt{\alpha}} (C^* z, C^* \varphi_2)_Q = 0, \tag{2.9b}$$

and so we are led to introduce the Hilbert space

$$V := V_1 \times V_2 \quad \text{with} \quad \|v\| := \left(\|v_1\|_1^2 + \|v_2\|_2^2 \right)^{1/2}, \quad v = (v_1, v_2) \in V,$$

and the bilinear form $b\colon V \times V \to \mathbb{R}$ given by

$$b(v,\varphi) := a(v,\varphi) + \tfrac{1}{\sqrt{\alpha}} c(v,\varphi) \tag{2.10a}$$

with

$$a(v,\varphi) := a(v_1,\varphi_2) + a(\varphi_1, v_2), \tag{2.10b}$$

$$c(v,\varphi) := (C^* v_2, C^* \varphi_2)_Q - (I v_1, I \varphi_1)_W \tag{2.10c}$$

for $v = (v_1, v_2), \varphi = (\varphi_1, \varphi_2) \in V$. Note that we use the same letter a for the bilinear form inducing the operator A and for the one in (2.10b); this "operator overloading" should not cause confusion when the domain is clear. If not, we shall distinguish the two forms by writing $a_{|V_1 \times V_2}$ or $a_{|V \times V}$. In this notation, the variational formulation of the rescaled and reduced optimality system (2.8) simply reads

$$\text{find } x \in V \text{ such that} \quad \forall \varphi \in V \;\; b(x,\varphi) = -\tfrac{1}{\sqrt{\alpha}} (u_d, I\varphi_1)_W \,. \tag{2.11}$$

A pair $x = (u,z) \in X$ is a solution of (2.11) if and only if (u,z) is a solution of (2.9) if and only if the triple $(u, z, -\frac{1}{\sqrt{\alpha}} C^* z) \in V \times Q$ verifies the rescaled optimality system (2.7). Consequently, thanks to the convexity of (2.3), if $x = (u,z) \in V$ is a solution of (2.11), then $(-\frac{1}{\sqrt{\alpha}} C^* z, u) \in Q \times V_1$ is a solution of the original optimization problem (2.3).

Let us analyze the bilinear form $b = a + \frac{1}{\sqrt{\alpha}} c$. We readily see that

$$a_{|V \times V},\ c, \text{ and so } b \text{ are symmetric,} \tag{2.12}$$

but b is not coercive in general. Consider, for example, a set-up where there exists $v = (v_1, v_2) \in V$ such that $\|I v_1\|_W > \|C^* v_2\|_Q$. Then c is not coercive and so, even for a coercive, also b is not coercive for $\alpha > 0$ sufficiently small.

In order to obtain further properties, let us first consider the contributions a and c separately. The bilinear form c is closely related to the original minimization problem (2.3) and its "energy seminorm" (2.4). To see this, observe that, if $(u,z) \in V$ and $\sqrt{\alpha} q = -C^* z$, we have the correspondence

$$\|Iu\|_W^2 + \|C^* z\|_Q^2 = \|Iu\|_W^2 + \alpha \|q\|_Q^2 \,,$$

which motivates to introduce the seminorm

$$|v| := \Big(\|I v_1\|_W^2 + \|C^* v_2\|_Q^2 \Big)^{1/2} \tag{2.13}$$

on V. Thus, denoting by Z the kernel of $|\cdot|$ and realizing that the bilinear form c is well-defined on the quotient space V/Z, we see that

$$\sup_{|v|=1,\,|\varphi|=1} |c(v,\varphi)| = 1 = \inf_{|v|=1} \sup_{|\varphi|=1} c(v,\varphi), \tag{2.14}$$

where the second identity relies on

$$c\big((v_1,v_2),(-v_1,v_2)\big) = \|C^* v_2\|_Q^2 + \|I v_1\|_W^2 = |v|^2 \,. \tag{2.15}$$

Since

$$\forall v \in V \quad |v| \le M \|v\| \tag{2.16}$$

with

$$M := \max\{M_I, M_C\},$$

the form c is also continuous in V, with constant M.

The bilinear form $a_{|V \times V}$ inherits its continuity and nondegeneracy properties from $a_{|V_1 \times V_2}$. More precisely, we have

$$\sup_{\|v\|=1,\,\|\varphi\|=1} |a(v,\varphi)| = M_a \quad \text{and} \quad \inf_{\|v\|=1} \sup_{\|\varphi\|=1} a(v,\varphi) = m_a \tag{2.17}$$

with M_a and m_a from (2.2). While the first identity is straight-forward, the second one hinges on the inf-sup-duality (cf. Babuška [1])

$$\inf_{\|v_1\|_1=1} \sup_{\|\varphi_2\|_2=1} a(v_1,\varphi_2) = \inf_{\|v_2\|_2=1} \sup_{\|\varphi_1\|_1=1} a(\varphi_1,v_2) \tag{2.18}$$

for a with domain $V_1 \times V_2$.

Turning to the complete bilinear form b, we may sum up the continuity properties as follows: for all $v, \varphi \in V$, we have

$$|b(v,\varphi)| \le M_a \|v\| \|\varphi\| + \frac{M}{\sqrt{\alpha}} \|v\| |\varphi| \le \|v\| \|\varphi\|_\alpha \tag{2.19}$$

with

$$\|\varphi\|_\alpha := M_a \|\varphi\| + \frac{M}{\sqrt{\alpha}} |\varphi| . \tag{2.20}$$

Here we have equipped V as trial space with $\|\cdot\|$ and as test space with $\|\cdot\|_\alpha$. The former is in accordance with our scopes in the error analyses below and the latter avoids in particular a dependence on $M/\sqrt{\alpha}$ of the continuity constant of b and in the following bound for the right-hand side in (2.11): for all $\varphi = (\varphi_1, \varphi_2) \in V$,

$$\left| \tfrac{1}{\sqrt{\alpha}} (u_d, I\varphi_1)_W \right| \le \frac{M_I}{\sqrt{\alpha}} \|u_d\|_W \|\varphi_1\|_1 \le \|u_d\|_W \|\varphi\|_\alpha . \tag{2.21}$$

The derivation of the nondegeneracy properties of the bilinear form b is more subtle. In order to establish the crucial inf-sup condition (2.2c), let $\varphi = (\varphi_1, \varphi_2) \in V$ be given.

In order to find a suitable $v = (v_1, v_2) \in V$, we combine the nondegeneracy properties of a and c in the ansatz

$$v = (w_1, w_2) + \gamma(-\varphi_1, \varphi_2), \tag{2.22a}$$

where $\gamma \ge 0$ and $w = (w_1, w_2) \in V$ is chosen with the help of (2.17) such that $\|w\| = \|\varphi\|$ and $a(w,\varphi) \ge m_a \|\varphi\|^2$. We then have

$$\|v\| \le \|w\| + \gamma \|\varphi\| \le (1+\gamma) \|\varphi\| \tag{2.22b}$$

and

$$\begin{aligned} b(v,\varphi) &\ge m_a \|\varphi\|^2 + \frac{\gamma}{\sqrt{\alpha}} |\varphi|^2 - \frac{M}{\sqrt{\alpha}} |\varphi| \|\varphi\| \\ &\ge m_a \left(\|\varphi\| + \frac{M}{M_a \sqrt{\alpha}} |\varphi| \right) \|\varphi\| + \frac{\gamma}{\sqrt{\alpha}} |\varphi|^2 - \frac{2M}{\sqrt{\alpha}} |\varphi| \|\varphi\| . \end{aligned} \tag{2.22c}$$

thanks the continuity (2.14) of c and $m_a \le M_a$. Using the inequality $2st \le \epsilon s^2 + t^2/\epsilon$ with $\epsilon = \frac{L}{1+2L} m_a > 0$ and

$$L := M/\sqrt{\alpha}, \tag{2.23a}$$

we may bound the critical term by

$$\frac{2M}{\sqrt{\alpha}} |\varphi|^2 \le \frac{L}{1+2L} m_a \|\varphi\|^2 + \frac{1+2L}{L} \frac{M^2}{m_a \alpha} |\varphi|^2 .$$

Thus, if we define

$$\gamma := \frac{M}{m_a} \left(1 + \frac{2M}{\sqrt{\alpha}} \right) \tag{2.23b}$$

by the coefficient of $|\varphi|^2$ divided by $\sqrt{\alpha}$, set

$$\kappa := \frac{1+2L}{1+L}(1+\gamma) = \frac{1+2L}{1+L} \left(1 + \frac{M}{m_a} \left(1 + \frac{2M}{\sqrt{\alpha}} \right) \right), \tag{2.23c}$$

and recall (2.22b), we arrive at

$$b(v,\varphi) \geq \frac{1+L}{1+2L}\frac{m_a}{M_a}\,\|\varphi\|_\alpha\,\|\varphi\| \geq \frac{1}{\kappa}\frac{m_a}{M_a}\,\|v\|\,\|\varphi\|_\alpha\,, \tag{2.24}$$

where the norms on the right-hand side coincide with those in the continuity bound (2.19). We therefore have the following basic result.

Theorem 2.1 (Bilinear form of reduced optimality system)**.** *If we equip V as trial space with $\|\cdot\|$ and as test space with $\|\cdot\|_\alpha$, then the inf-sup constant m_b and the continuity constant M_b of the bilinear form* (2.10) *satisfy*

$$0 < \frac{1}{\kappa}\frac{m_a}{M_a} \leq m_b \leq M_b \leq 1,$$

where κ is defined by the relations (2.23)*.*

The inequalities of Theorem 2.1 yield for the condition number of the bilinear form b (i.e., the ratio of its continuity constant to its inf-sup constant)

$$\frac{M_b}{m_b} \leq \kappa\frac{M_a}{m_a}.$$

The second factor, the condition number of the bilinear form a associated with the constraint, is expected to be a kind of lower bound. In this vein, we may view the first factor κ as a bound for the possible amplification of the constraint conditioning, resulting from the interplay of constraint and the objective in the constrained optimization problem (2.3). Inspecting (2.23), we see that κ is a function of the parameters α, M, m_a, and M_a. The next three remarks discuss asymptotic behaviors of κ that will play major roles in what follows or are of independent interest.

Remark 2.2 (Amplification for pure constraint case). Consider the special case $C = 0$ and $I = 0$. Then the rescaled and reduced optimality system (2.8) is a well-posed 'double' boundary value problem. Its condition number with respect to $(V,\|\cdot\|)\times(V,\|\cdot\|)$ is M_a/m_a; cf. (2.17). As $C = 0$ and $I = 0$ imply $M = 0$, $L = 0$, and so $\gamma = 0$ and $\kappa = 1$, this is reproduced by Theorem 2.1.

It is worth mentioning that this limiting case of "pure constraint" is attained in a continuous manner:

$$\kappa - 1 = \big(1+o(1)\big)\frac{M}{\sqrt{\alpha}} \quad \text{as} \quad M \to 0,$$

where $L = M/\sqrt{\alpha}$ is essentially the operator norm of the perturbation.

Remark 2.3 (Amplification for degenerating constraint). While the continuity constant M_a of the bilinear form a does not enter κ, its inf-sup constant m_a does, in a critical manner. More precisely, we have

$$\kappa = \left(\frac{1+2L}{1+L}\left(1+\frac{2M}{\sqrt{\alpha}}\right)M + o(1)\right)\frac{1}{m_a} \quad \text{as} \quad m_a \to 0.$$

Notice that the fraction involving L has only values in the interval $[1,2]$.

Remark 2.4 (Amplification for vanishing regularization). Consider the limit $\alpha \to 0$ of the Tikhonov regularization parameter (while I and C are fixed). Then $L \to \infty$ so that

$$\kappa = \left(\frac{4M^2}{m_a} + o(1)\right)\frac{1}{\sqrt{\alpha}} \quad \text{as} \quad \alpha \to 0. \tag{2.25}$$

Let us see with a simple example that the inf-sup constant m_b in Theorem 2.1 can blow up with this rate and so the lower bound therein cannot be improved for small α without further assumptions on the structure of b.

Consider $V_1 = V_2 = \mathbb{R}^2$, where $\|\cdot\|_1$ and $\|\cdot\|_2$ are the Euclidean norm in $\mathbb{R}^2$,

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \text{ and } \quad \mathbf{C} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

and $\alpha > 0$. The symmetric bilinear form b of the optimality system is then given by the matrix

$$\mathbf{B} = \begin{pmatrix} -\frac{1}{\sqrt{\alpha}} & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & \frac{1}{\sqrt{\alpha}} \end{pmatrix}.$$

For $\varphi_0 = (\sqrt{\alpha}, 0, 1, 0) \in V = \mathbb{R}^4$, we have $\|\varphi_0\|_\alpha = \sqrt{1+\alpha} + 1$ and

$$\sup_{v \in V} \frac{Bv \cdot \varphi_0}{\|v\|} = \sup_{v \in V} \frac{v \cdot (0, 0, \sqrt{\alpha}, 0)}{\|v\|} = \sqrt{\alpha}$$

so that

$$\inf_{\varphi \in V} \sup_{v \in V} \frac{Bv \cdot \varphi}{\|v\| \, \|\varphi\|_\alpha} \leq \sqrt{\frac{\alpha}{2}}. \tag{2.26}$$

Hence, the asymptotic behavior of α in (2.25) is attained.

The chosen norms for V as trial and test space are not always the most convenient ones. This follows from the following remark considering a special case.

Remark 2.5 (Coercive constraints with $C^* = I$). Suppose that $V_1 = V_2$ and $Q = W$ with coinciding scalar products and norms and that the bilinear form $a_{|V_1 \times V_1}$ is coercive with constant $\tilde{m}_a$ and $C^* = I$. It is worth noting that, as $a_{|V_1 \times V_1}$ is not necessarily symmetric, the best coercivity constant $\tilde{m}_a$ may be much smaller than the inf-sup constant m_a. Given $\varphi \in V$, we proceed as in (2.22) taking $w = \varphi$, $\gamma = 0$, and obtain

$$b(v, \varphi) \geq \tilde{m}_a \left(\|\varphi\|^2 + \frac{1}{\sqrt{\alpha}} |\varphi|^2 \right) \tag{2.27a}$$

because of $c(\varphi, \varphi) = 0$. This fits well to the following variant of the continuity bound (2.19):

$$|b(v, \varphi)| \leq \max\{M_a, 1\} \left(\|v\|^2 + \frac{1}{\sqrt{\alpha}} |v|^2 \right)^{1/2} \left(\|\varphi\|^2 + \frac{1}{\sqrt{\alpha}} |\varphi|^2 \right)^{1/2}. \tag{2.27b}$$

Hence, in this case, the condition number of b with respect to the norms in (2.27) is independent of the Tikhonov regularization parameter α. Nevertheless, if $C^* \neq I$, also this choice of norms cannot offer in general an asymptotic behavior better than $1/\sqrt{\alpha}$ as $\alpha \to 0$. In fact, re-computing the example in Remark 2.4 with the norms in (2.27) does not change the behavior of its inf-sup constant.

Let us conclude this section with the following side product of our discussion of the bilinear form b.

Corollary 2.6 (Existence and uniqueness)**.** *The rescaled and reduced optimality system* (2.11) *and thus* (2.5) *has a unique solution.*

Proof. Inequality (2.24) ensures (2.2c) for the bilinear form b and, thanks to the algebraic symmetry of b, also (2.2b). □

3. Analysis for variational discretization

In this section, we analyze the error of the variational discretization of the optimization problem (2.3) according to Hinze [14]. Our key tool is the rescaled and reduced optimality system (2.8), whose Galerkin solution coincides with the approximate solution of the variational discretization.

3.1. Variational discretization and reduced optimality system.

We start by discretizing the PDE constraint (2.1) of the optimization problem (2.3). Recalling its variational formulation

$$\text{find } u \in V_1 \quad \text{such that} \quad \forall \varphi_2 \in V_2 \;\; a(u, \varphi_2) = \langle Cq, \varphi_2 \rangle \,,$$

we choose some conforming finite-dimensional spaces $V_{h,i} \subset V_i$, $i = 1, 2$, such that the restriction of the bilinear form a on $V_{h,1} \times V_{h,2}$ is nondegenerate. The corresponding Petrov-Galerkin method then reads

$$\text{find } u_h \in V_{h,1} \quad \text{such that} \quad \forall \varphi_{h,2} \in V_{h,2} \;\; a(u_h, \varphi_{h,2}) = \langle Cq, \varphi_{h,2} \rangle \,.$$

Using this for the constraint in (2.3), we arrive at the (semi-)discrete optimization problem

$$\begin{aligned} \min_{(\tilde{q}, u_h) \in Q \times V_{h,1}} \; & \frac{1}{2} \|Iu_h - u_d\|_W^2 + \frac{\alpha}{2} \|\tilde{q}\|_Q^2 \\ \text{subject to} \quad & \forall \varphi_{h,2} \in V_{h,2} \quad a(u_h, \varphi_{h,2}) = (\tilde{q}, C^* \varphi_{h,2})_Q \,, \end{aligned} \tag{3.1}$$

where we, in addition, assume that I can be exactly evaluated for any function from $V_{h,1}$. As in the continuous case, $(\tilde{q}, u_h) \in Q \times V_{h,1}$ is the unique solution of (3.1) if and only if there exists $z_h \in V_{h,2}$ such that

$$\begin{aligned} &\forall \varphi_{h,2} \in V_{h,2} \quad a(u_h, \varphi_{h,2}) = (\tilde{q}, C^* \varphi_{h,2})_Q \,, \\ &\forall \varphi_{h,1} \in V_{h,1} \quad a(\varphi_{h,1}, z_h) = \tfrac{1}{\sqrt{\alpha}} \left(Iu_h - u_d, I\varphi_{h,1}\right)_W \,, \\ &\qquad\qquad \sqrt{\alpha} \tilde{q} = -C^* z_h. \end{aligned} \tag{3.2}$$

Also here, we may eliminate the approximate control $\tilde{q}$ by inserting the third equation into the first one. Setting $V_h := V_{h,1} \times V_{h,2}$, the variational formulation of the ensuing discrete rescaled and reduced optimality system is

$$\begin{aligned} &\text{find } x_h = (u_h, z_h) \in V_h \text{ such that} \\ &\qquad \forall \varphi_h = (\varphi_{h,1}, \varphi_{h,2}) \in V_h \quad b(x_h, \varphi_h) = -\tfrac{1}{\sqrt{\alpha}} \left(u_d, I\varphi_{h,1}\right)_W . \end{aligned} \tag{3.3}$$

Its solution x_h is the Galerkin approximation in V_h to the solution x of the variational formulation (2.11) of the rescaled and reduced optimality system. Applying Corollary 2.6 to the discrete spaces therefore yields the following approach to uniqueness and existence of the variational discretization of (2.11).

Lemma 3.1 (Discrete well-posedness)**.** *The discrete reduced optimality system* (3.3) *has a unique variational solution $x_h = (u_h, z_h) \in V_h$. Consequently, the pair $(\tilde{q}, u_h)$ with $\tilde{q} = -\frac{1}{\sqrt{\alpha}} C^* z_h$ is the unique solution of the semidiscrete optimization problem* (3.1)*.*

Remarkably, the approximate solutions $(\tilde{q}, u_h, z_h)$ of the variational discretization (3.2) are computable whenever $C^*_{|V_{h,2}}$ and $I_{|V_{h,1}}$ can be evaluated exactly.

3.2. Non-asymptotic quasi-best approximation.

We shall assess the quality of the Galerkin approximation $x_h = (u_h, z_h) \in V_h$ from (3.3), assuming that we are interested particularly in the $\|\cdot\|_1$-error of the approximate state u_h. For this purpose, we compare it with a suitable best error in V_h.

Let us first recall some basic results in Petrov-Galerkin approximation, which we already formulate for the discretization of the constraint. Let $R_{h,1} v_1 \in V_{h,1}$ be the

generalized Ritz projection of $v_1 \in V_1$ given by $a(R_{h,1}v_1, \varphi_{h,2}) = a(v_1, \varphi_{h,2})$ for all $\varphi_{h,2} \in V_{h,2}$. Since $a_{|V_1 \times V_2}$ satisfies (2.2) and is nondegenerate on $V_{h,1} \times V_{h,2}$, there exists a constant $\mu_h \geq 1$ such that

$$\|v_1 - R_{h,1}v_1\|_1 \leq \mu_h \inf_{v_{h,1} \in V_{h,1}} \|v_1 - v_{h,1}\|_1 ;$$

see, e.g., Babuška [1]. We refer to the smallest possible choice of μ_h as the *quasi-best-approximation constant of the constraint discretization*. Xu and Zikatanov [26] show the identities

$$\mu_h = \|I - R_{h,1}\|_{L(V_{h,1})} = \|R_{h,1}\|_{L(V_{h,1})} \tag{3.4}$$

and Tantardini and Veeser [23, Theorem 2.1] give the formula

$$\mu_h = \sup_{\varphi_{h,2} \in V_{h,2}} \frac{\sup_{\|v_1\|_1 = 1} a(v_1, \varphi_{h,2})}{\sup_{\|v_{h,1}\|_1 = 1} a(v_{h,1}, \varphi_{h,2})}, \tag{3.5}$$

where v_1 varies in V_1 and $v_{h,1}$ varies in $V_{h,1}$ and, for the sake of notational simplicity, a tedious $\varphi_{h,2} \neq 0$ is avoided.

A perhaps striking feature of these formulas is that they are not affected by the choices of the norms in the test spaces $V_{h,2}$ and V_2. This comes in quite useful in our context, as the adjoint state is an auxiliary variable and, in the original approximation problem (2.3), the norm $\|\cdot\|_2$ is free as long as (2.2) continues to hold with $\|\cdot\|_1$. Exploiting this freedom, we propose to (possibly) redefine the norm on the space V_2 by

$$\|v_2\|_2 := \sup_{\varphi_1 \in V_1, \|\varphi_1\|_1 = 1} a(\varphi_1, v_2) \tag{3.6}$$

and so, in particular, to measure the error of the approximate adjoint state z_h in this norm. This redefinition affects the constants that we associated with the constrained optimization problem (2.3). The new continuity and inf-sup constants of the bilinear forms $a_{|V_1 \times V_2}$ are

$$M_a = 1 = m_a. \tag{3.7}$$

The constant M_I is not affected, while we have

$$M_{a,\mathrm{old}}^{-1} \leq \frac{M_C}{M_{C,\mathrm{old}}} \leq m_{a,\mathrm{old}}^{-1}. \tag{3.8}$$

where we indicate quantities before the redefinition by an additional index "old". As in addition

$$m_{a,\mathrm{old}} \|\cdot\|_{2,\mathrm{old}} \leq \|\cdot\|_2 \leq M_{a,\mathrm{old}} \|\cdot\|_{2,\mathrm{old}} ,$$

the results below hold also with the original norm in V_2, but the constants have to be revisited.

The convenience of the choice (3.6) lies in the following consequences of (3.7). The numerator in (3.5) is $\|\varphi_{h,2}\|_2$, which, together with the inf-sup-duality, cf. (2.18), yields

$$\inf_{v_{h,1} \in V_{h,1}} \sup_{\varphi_{h,2} \in V_{h,2}} \frac{a(v_{h,1}, \varphi_{h,2})}{\|v_{h,1}\|_1 \|\varphi_{h,2}\|_2} = \frac{1}{\mu_h} = \inf_{v_{h,2} \in V_{h,2}} \sup_{\varphi_{h,1} \in V_{h,1}} \frac{a(\varphi_{h,1}, v_{h,2})}{\|v_{h,2}\|_2 \|\varphi_{h,1}\|_1} \tag{3.9}$$

for the inf-sup constant of $a_{|V_{h,1} \times V_{h,2}}$. Accordingly, the generalized Ritz projection $R_{h,2}v_2 \in V_{h,2}$ of $v_2 \in V_2$ given by $a(\varphi_{h,1}, R_{h,2}v_2) = a(\varphi_{h,1}, v_2)$ for all $\varphi_{h,1} \in V_{h,1}$ verifies

$$\|v_2 - R_{h,2}v_2\|_2 \leq \mu_h \inf_{v_{h,2} \in V_{h,2}} \|v_2 - v_{h,2}\|_2 .$$

Setting $R_h = (R_{h,1}, R_{h,2})$, we also have

$$\|v - R_h v\| \leq \mu_h \inf_{v_h \in V_h} \|v - v_h\| . \tag{3.10}$$

After these preparations, we are ready to derive a first result about quasi-best approximation of the variational discretization (3.1).

Theorem 3.2 (Non-asymptotic quasi-best approximation)**.** *Let $x = (u, z)$ be any solution of the optimality system* (2.11) *and choose* (3.6) *as norm in V_2. The combined error in the corresponding approximate state u_h and its adjoint z_h of the variational discretization is quasi-best in V_h with*

$$\|x - x_h\| \leq \kappa_h \mu_h \inf_{v_h \in V_h} \|x - v_h\|.$$

Here

$$\kappa_h = \frac{1+2L}{1+L}\left(1 + M\left(1 + \frac{2M}{\sqrt{\alpha}}\right)\mu_h\right) \quad \text{with} \quad L = \frac{M}{\sqrt{\alpha}},$$

and μ_h is the quasi-best-approximation constant of the constraint discretization.

Proof. Thanks to Theorem 2.1 and Lemma 3.1, we can use the counterpart of (3.5) for the characterization (3.3) of the variational discretization. Let $\varphi_h \in V_h$. The continuity bound (2.19) and (3.7) give for the numerator

$$\sup_{\|v\|=1} b(v, \varphi) \leq \|\varphi_h\|_\alpha .$$

For the denominator, we use (2.22), where V is replaced by V_h and, therefore, with $1/\mu_h$ in place of m_a in view of (3.9). We thus obtain

$$\sup_{v_h \in V_h, \|v_h\|=1} b(v_h, \varphi_h) \geq \frac{1}{\kappa_h \mu_h} \|\varphi_h\|_\alpha \tag{3.11}$$

and the proof is finished. □

In the special situation of Remark 2.5, we can obtain the following quasi-best approximation result.

Remark 3.3 (Quasi-best approximation for coercive constraints and $C^* = I$). Suppose that $V_1 = V_2$ and $Q = W$ with coinciding scalar products and norms and that the bilinear form a is V_1-coercive with constant $\tilde{m}_a$ and $C^* = I$. Exploiting the coercivity and continuity properties of Remark 2.5, we derive for the error of the variational discretization (2.11)

$$\|x - x_h\|^2 + \frac{1}{\sqrt{\alpha}} |x - x_h|^2 \leq \frac{\max\{M_a^2, 1\}}{\tilde{m}_a^2} \inf_{v_h \in V_h} \left(\|x - v_h\|^2 + \frac{1}{\sqrt{\alpha}} |x - v_h|^2 \right).$$

The quasi-best approximation constant in the preceding Remark 3.3 does not blow up for vanishing regularization. Nonetheless, when measuring the error merely with $\|\cdot\|$, it does not exclude an $\alpha^{-1/4}$-blow up of the quasi-best approximation constant even in the special case $C^* = I$ considered in Remark 2.4 and, in the light of the example therein, it does not exclude an $\alpha^{-3/4}$-blow up for general operators I and C. As we shall see, the α-dependence in Theorem 3.2 is less severe.

Remark 3.4 (Vanishing regularization and quasi-best approximation). As in Remark 2.4, we consider the limit $\alpha \to 0$ for the Tikhonov regularization parameter. Similarly to there, we have

$$\kappa_h = \left(\frac{4M^2}{\mu_h} + o(1)\right) \frac{1}{\sqrt{\alpha}} \quad \text{as} \quad \alpha \to 0. \tag{3.12}$$

This blow up arises from the lower bound of the inf-sup constant in Theorem 2.1, which cannot be improved because of (2.26). Note however, that the equivalence of the norms $\|\cdot\|_\alpha$ and $\sup_{\|v\|=1} b(v, \cdot)$ is not uniform in α. In the light of (3.5), it is therefore conceivable that (3.12) could be improved by using the latter as test space norm. However, the determination of the discrete inf-sup constant with respect to

this abstract norm appears to be much more involved than the approach (2.22), which directly carries over to discrete spaces.

In any case, we shall show below that, under refinement, the α-dependence disappears for many instances of the optimality system (2.7).

3.3. **Asymptotic quasi-best approximation.** In this section, we complement Theorem 3.2. To be more precise, let ν_h be the *quasi-best-approximation constant of the variational discretization* therein and consider a sequence $(V_h)_h$ of discrete spaces leading to a uniform stable constraint discretization in that

$$\exists \bar{\mu} \geq 1 \quad \forall h > 0 \quad \mu_h \leq \bar{\mu}, \tag{3.13}$$

which is equivalent to discrete inf-sup stability in view of (3.9). Theorem 3.2 then ensures the existence of a constant $\bar{\nu}$ such that

$$\forall h > 0 \quad \nu_h \leq \bar{\nu}. \tag{3.14}$$

This upper bound may be pessimistic. To motivate this assessment, represent the bilinear form b by the operator matrix

$$\begin{pmatrix} A & \frac{1}{\sqrt{\alpha}} CC^* \\ -\frac{1}{\sqrt{\alpha}} I^* I & A^* \end{pmatrix},$$

which is the one in (2.8) with inverted rows. If C and I are compact, this matrix is diagonally dominant in an operator sense and can be viewed as a compact perturbation of the diagonal matrix with the entries A and A^*. Therefore, in order to improve on (3.14), we mimic somewhat the argument in Schatz [22], introducing some new twist.

Let us first observe that, in accordance with Remark 2.2, Theorem 3.2 yields $\nu_h \leq \mu_h$ whenever $M_I = 0 = M_C$. More precisely and generally, we have the following relationship between the two quasi-best-approximation constants.

Lemma 3.5 (Quasi-best-approximation constants)**.** *The quasi-best-approximation constants ν_h and μ_h are related by*

$$|\nu_h - \mu_h| \leq \kappa_h \mu_h \sup_{\|v\|=1} |v - R_h v|,$$

where κ_h is as in Theorem 3.2 and R_h is the generalized Ritz projection in (3.10)*.*

Proof. As in the proof of Theorem 3.2, we will make use of (3.5) with a replaced by b. Given $v \in V$ and $\varphi_h \in V_h$, we can write

$$b(v, \varphi_h) = b(R_h v, \varphi_h) + \frac{1}{\sqrt{\alpha}} c(v - R_h v, \varphi_h)$$

because of $a(v - R_h v, \varphi_h) = 0$. Hence,

$$\left| \sup_{\|v\|=1} b(v, \varphi_h) - \sup_{\|v\|=1} b(R_h v, \varphi_h) \right| \leq \frac{1}{\sqrt{\alpha}} \sup_{\|v\|=1} |c(v - R_h v, \varphi_h)|.$$

As

$$\frac{\sup_{\|v\|=1} b(R_h v, \varphi_h)}{\sup_{v_h \in V_h, \|v_h\|=1} b(v_h, \varphi_h)} \leq \|R_h\|_{L(V)} = \mu_h$$

with equality for some $\varphi_h \in V_h$, we obtain

$$|\nu_h - \mu_h| \leq \sup_{\varphi \in V_h} \frac{\frac{1}{\sqrt{\alpha}} \sup_{\|v\|=1} |c(v - R_h v, \varphi_h)|}{\sup_{v_h \in V_h, \|v_h\|=1} b(v_h, \varphi_h)}.$$

Thanks to (2.14), (2.20), and (3.11) this proves the claimed inequality. □

In order to deploy Lemma 3.5, we need additional assumptions for our optimization problem and its discretization. We shall consider two settings: a "qualitative" and a "quantitative" one. The former assumes in addition

$$I : V_1 \to W \text{ and } C : Q \to V_2^* \text{ are compact} \tag{3.15a}$$

for the optimization problem and

$$\forall v \in V \quad \lim_{h\to 0} \inf_{v_h \in V_h} \|v - v_h\| = 0, \tag{3.15b}$$

for the constraint discretization. Notice that, owing to (3.8), the condition (3.15a) is independent of our choice to equip V_2 with the norm (3.6).

Lemma 3.6 (Qualitative asymptotic quasi-best approximation)**.** *Under the assumptions* (3.13) *and* (3.15)*, the quasi-best-approximation constant* ν_h *satisfies*

$$\nu_h = \mu_h\big(1 + \bar\kappa\, o(1)\big) \quad \textit{as} \quad h \to 0,$$

where

$$\bar\kappa = \frac{1+2L}{1+L}\left(1 + M\left(1 + \frac{2M}{\sqrt{\alpha}}\right)\bar\mu\right) \quad \textit{with} \quad L = \frac{M}{\sqrt{\alpha}}.$$

Proof. In the light of Lemma 3.5 and (3.13), it suffices to verify the uniform convergence

$$\lim_{h\to 0} \sup_{\|v\|=1} |v - R_h v| = 0. \tag{3.16}$$

This follows from a standard argument; we provide details for the sake of completeness. Let $(h_k)_k$ be any sequence with $\lim_{k\to 0} h_k = 0$ and choose v_k such that

$$\forall k \in \mathbb{N} \quad \|v_k\| = 1 \text{ and } \sup_{\|v\|=1} |v - R_k v| \le |v_k - R_k v_k| + \frac{1}{k},$$

where we write k instead h_k whenever the latter is an index. Exploiting (3.13) another time, we see that the sequence given by $d_k := v_k - R_k v_k$ is bounded in the Hilbert space V. Owing to (3.15b), its weak limit $d \in V$ satisfies

$$a(d, \varphi) = a(d - d_k, \varphi) + a(d_k, \varphi - \varphi_k)$$

for any $\varphi \in V$ and $\varphi_k \in V_k$. Choosing φ_k by means of (3.15b), we derive $a(d, \varphi) = 0$ by $k \to \infty$. Consequently, (2.17) yields $d = 0$. Thanks to (3.15a), the operator $I : V_1 \to W$ and the adjoint $C^* : V_2 \to Q$ are compact. This turns the weak convergence $d_k \to 0$ in V into the strong convergence $|d_k| \to 0$ and the proof is finished. □

In order to quantify the convergence in Lemma 3.6, we shall use a duality argument. This requires a second, more specific setting of additional assumptions involving the Sobolev spaces H^s, $s \ge 0$, and their norms $|\cdot|_s$ over some domain. We use $|\cdot|_s$ instead of $\|\cdot\|_s$ in order to avoid confusion with the norms $\|\cdot\|_1$ and $\|\cdot\|_2$ of V_1 and V_2. For $s < 0$, we denote by H^s the (topological) dual space of H^{-s} and $|\cdot|_s$ stands for the dual norm of $|\cdot|_{-s}$.

We suppose that spaces V_1 and V_2 relate to Sobolev spaces in the following way: There are $s_i \in \mathbb{R}$, $i = 1, 2$, and a constant $C_S \ge 1$ such that

$$V_i \text{ is a closed subspace of } H^{s_i} \text{ and } C_S^{-1}\,|\cdot|_{s_i} \le \|\cdot\|_i \le C_S\,|\cdot|_{s_i} \text{ for } i = 1, 2. \tag{3.17a}$$

Furthermore, we suppose that there is $\delta > 0$ such that the following three conditions hold. First, the operators C and I have the boundedness properties

$$C \in L(Q, H^{-s_2+\delta}) \quad \text{and} \quad I \in L(H^{s_1-\delta}, W). \tag{3.17b}$$

Thus, the canonical embeddings $H^{-s_2+\delta} \to H^{-s_2}$ and $H^{s_1} \to H^{s_1-\delta}$ quantify the compactness assumption (3.15a). Second, the differential operator of the constraint

and its adjoint offer the following regularity estimates: there is a constant $C_R > 0$ such that, for all admissible f and g,

$$\left|A^{-1}f\right|_{s_1+\delta} \leq C_R \left|f\right|_{-s_2+\delta} \quad \text{and} \quad \left|A^{-*}g\right|_{s_2+\delta} \leq C_R \left|g\right|_{-s_1+\delta}. \tag{3.17c}$$

Third and last, the approximation spaces V_h verify

$$\inf_{v_h \in V_h} \|v - v_h\| \leq C_{\mathcal{I}} h^{\delta} \left(|v_1|^2_{s_1+\delta} + |v_2|^2_{s_2+\delta} \right)^{1/2} \tag{3.17d}$$

for some constant $C_{\mathcal{I}} > 0$, which quantifies the approximation property (3.15b).

Theorem 3.7 (Quantitative asymptotic best approximation)**.** *Under the assumptions* (3.13) *and* (3.17)*, the quasi-best-approximation constant ν_h satisfies*

$$\nu_h = \mu_h\big(1 + \bar{\kappa} O(h^{\delta})\big) \quad \textit{as} \quad h \to 0,$$

where $\bar{\kappa}$ is as in Lemma 3.6. For the α-dependence of $\bar{\kappa}$, cf. Remark 3.4.

Proof. Similarly as in the first step of the proof of Lemma 3.6, inserting (3.13) and

$$\lim_{h \to 0} \sup_{\|v\|=1} |v - R_h v| = O(h^{\delta}). \tag{3.18}$$

into Lemma 3.5 establishes the claim. To show (3.18), let $v \in V$ with $\|v\| = 1$ and define $\varphi \in V$ as the solution of the following "dual" problem associated with the bilinear form $a_{|V \times V}$:

$$A\varphi_1 = CC^* d_2, \quad A^* \varphi_2 = I^* I d_1,$$

where $d = (d_1, d_2) := v - R_h v$. We thus have

$$\begin{aligned} |v - R_h v|^2 = |d|^2 &= \langle I^* I d_1, d_1 \rangle_1 + \langle CC^* d_2, d_2 \rangle_2 = a(d, \varphi) = a(v - R_h v, \varphi) \\ &= a(v - R_h v, \varphi - \varphi_h) \leq \|v - R_h v\| \, \|\varphi - \varphi_h\|, \end{aligned} \tag{3.19}$$

where $\varphi_h \in V_h$ is arbitrary. For the first factor, (3.10) and (3.13) imply

$$\|v - R_h v\| \leq \mu_h \leq \mu. \tag{3.20}$$

For second factor, we employ (3.17d) with suitable $\varphi_h \in V_h$ to obtain

$$\|\varphi - \varphi_h\| \leq C_{\mathcal{I}} h^{\delta} \left(|\varphi_1|^2_{s_1+\delta} + |\varphi_2|^2_{s_2+\delta} \right)^{1/2}$$

and it remains to show that the norms on the right-hand side are suitably bounded. Let consider the first one. Making use of the regularity estimate (3.17c) and the definition of φ_1, we deduce

$$\begin{aligned} |\varphi_1|_{s_1+\delta} &\leq C_R \left|A\varphi_1\right|_{-s_2+\delta} = C_R \left|CC^* d_2\right|_{-s_2+\delta} \leq C_R \bar{M}_C \left\|C^* d_2\right\|_Q \\ &\leq C_R \bar{M}_C \left|d\right| = C_R \bar{M}_C \left|v - R_h v\right|, \end{aligned}$$

where $\bar{M}_C$ is the operator norm of C from (3.17b). A similar argument yields

$$|\varphi_2|_{s_2+\delta} \leq C_R \bar{M}_I \left|v - R_h v\right|,$$

where $\bar{M}_I$ is the operator norm of I in (3.17b). We insert the previous estimates in the first one and conclude

$$|v - R_h v| \leq \mu C_{\mathcal{I}} C_R \bar{M} h^{\delta}$$

with $\bar{M} := \max\{\bar{M}_I, \bar{M}_C\}$, i.e., (3.18). □

Let us exemplify Theorem 3.7 by two applications. The first one considers the optimization problem (1.1) of the introduction, while the second one is more involved in that the constraint does not allow for a coercive set-up.

Example 3.8 (Simple model optimization). Discretize the optimization problem (1.1) of the introduction with linear finite elements on quasi-uniform meshes with meshsize h. We have $V_1 = H^1_0 = V_2$ and, if we choose $\|\cdot\|_1 = |\nabla\cdot|_0$, we already have $m_a = 1 = M_a$ and (3.6) does not change the norm in V_2. Further, $M_I = C_F = M_C$, where C_F is the constant in the Poincaré-Friedrichs inequality. Moreover, we have $s_1 = 1 = s_2$ and, assuming that the underlying domain is convex, $\delta = 1$. Taking Sobolev seminorm instead of norms in (3.17a), we then have $C_S = 1$ for the relevant cases and $C_R = 1$ thanks to elliptic regularity as well as $\bar{M}_I = 1 = \bar{M}_C$. Standard approximation theory shows (3.17d) with $C_{\mathcal{I}}$ depending on the shape regularity of the underlying meshes. Since $\mu_h = 1$, we conclude

$$|\nu_h - 1| \le 2\left(1 + C_F\left(1 + \frac{2C_F}{\sqrt{\alpha}}\right)\right)h \quad \text{as} \quad h \to 0$$

for the quasi-best-approximation constant of the variational discretization in this case.

Example 3.9 (Point source control). We consider the following modification of the optimization problem (1.1), where the distributed control is replaced by a finite number of point sources:

$$\min_{(q,u)\in\mathbb{R}^\ell\times H^{1-\sigma}_0} \frac{1}{2}|u - u_d|_0^2 + \frac{\alpha}{2}\sum_{j=1}^{\ell} q_j^2 \quad \text{subject to} \quad -\Delta u = \sum_{j=1}^{\ell} q_j \delta_{x_j}, \tag{3.21}$$

where the underlying domain $\Omega \subset \mathbb{R}^2$ is planar, polygonal, Lipschitz, but not necessarily convex, $\{x_j\}_{j=1}^{\ell} \subset \Omega$ are ℓ distinct points, δ_{x_j} denotes the Dirac functional at the point x_j, and $0 < \sigma < \frac{1}{2}$. The bilinear form $a(v,w) = \int_\Omega \nabla v \cdot \nabla w\, dx$, $v, w \in C_0^\infty(\Omega)$, has a continuous and inf-sup-stable extension on $V_1 \times V_2$ with $V_1 = H^{1-\sigma}_0(\Omega)$ and $V_2 = H^{1+\sigma}_0(\Omega)$ and allows for a standard discretization with linear finite elements S_h for both trial and test space; see, e.g., [11]. For the verification of the discrete inf-sup condition, denote by R_h and Λ_h the Ritz projection and the Scott-Zhang interpolation operator, respectively. As

$$|R_h\varphi|_{1+\sigma} \le |\Lambda_h\varphi|_{1+\sigma} + |R_h\varphi - \Lambda_h\varphi|_{1+\sigma} \lesssim |\varphi|_{1+\sigma} + h^{-\sigma}\,|R_h\varphi - \Lambda_h\varphi|_1$$

and

$$\begin{aligned} h^{-\sigma}\,|R_h\varphi - \Lambda_h\varphi|_1 &\le h^{-\sigma}\,|R_h\varphi - \varphi|_1 + h^{-\sigma}\,|\varphi - \Lambda_h\varphi|_1 \\ &\lesssim h^{-\sigma}\,|\varphi - \Lambda_h\varphi|_1 \lesssim |\varphi|_{1+\sigma}\,, \end{aligned}$$

the continuous inf-sup-condition yields, for any $s_h \in S_h$,

$$|s_h|_{1-\sigma} \lesssim \sup_{|\varphi|_{1+\sigma}=1} a(s_h,\varphi) = \sup_{|\varphi|_{1+\sigma}=1} a(s_h, R_h\varphi) \lesssim \sup_{\varphi_h\in S_h, \|\varphi\|_{1+\sigma}=1} a(s_h,\varphi_h),$$

and so

$$|s_h|_{1-\sigma} \le \mu_h \sup_{\varphi_h\in S_h, \|\varphi_h\|_2=1} a(s_h,\varphi_h),$$

where μ_h depends only on continuous inf-sup constant and on the shape regularity of the underlying mesh and we switched to (3.6) for the norm on V_2. To complete the setting, we set $W = L^2(\Omega)$, $Q = \mathbb{R}^\ell$, and let I be the canonical embedding $H^{1-\sigma}(\Omega) \to L^2(\Omega)$ and $C : \mathbb{R}^\ell \to H^{-(1+\sigma)}(\Omega)$ be given by $Cq = \sum_{j=1}^{\ell} q_j\delta_{x_j}$. The continuity constants M_I and M_C are of order 1 and ℓ, respectively. Notice that, for $\sigma = 0$, C is not continuous because functions in $H^1_0(\Omega)$ do not have point values in general. Choosing $\delta \in (0,\sigma)$, we have (3.17) with $s_1 = 1-\sigma$, $s_2 = 1+\sigma$ and therefore

$$\nu_h = \mu_h\left(1 + \frac{O(h^\delta)}{\sqrt{\alpha}}\right) \quad \text{as} \quad h \to 0.$$

4. Analysis with approximate control-action operator

In this section, we shall analyze the approximation properties of a variational discretization, where the control-action operator is approximated. This includes the case of a discretized control space.

4.1. **Approximate variational discretization.** Let $V_{h,i} \subset V_i$, $i = 1, 2$, be the same finite-dimensional conforming spaces introduced in Section 3.1 and assume that the linear operator $C_h^* : V \to Q$ approximates C^*. Then the (semi-)discrete optimization

$$\begin{aligned} \min_{(\tilde{q}_h, u_h) \in Q \times V_{h,1}} \ & \frac{1}{2} \|Iu_h - u_d\|_W^2 + \frac{\alpha}{2} \|\tilde{q}_h\|_Q^2 \\ \text{subject to} \quad & \forall \varphi_{h,2} \in V_{h,2} \quad a(u_h, \varphi_{h,2}) = (\tilde{q}_h, C_h^* \varphi_{h,2})_Q \,, \end{aligned} \tag{4.1}$$

generalizes (3.1). It has the solution $(\tilde{q}_h, \tilde{u}_h) \in Q \times V_{h,1}$ if and only if there exists $\tilde{z}_h \in V_{h,2}$ such that

$$\begin{aligned} \forall \varphi_{h,2} \in V_{h,2} \quad & a(\tilde{u}_h, \varphi_{h,2}) = (\tilde{q}_h, C_h^* \varphi_{h,2})_Q \,, \\ \forall \varphi_{h,1} \in V_{h,1} \quad & a(\varphi_{h,1}, \tilde{z}_h) = \tfrac{1}{\sqrt{\alpha}} \left(I\tilde{u}_h - u_d, I\varphi_{h,1}\right)_W \,, \\ & \sqrt{\alpha}\tilde{q}_h = -C_h^* \tilde{z}_h. \end{aligned} \tag{4.2}$$

As before, we may eliminate $\tilde{q}_h$. If we define

$$b_h(v, \varphi) := a(v, \varphi) + \frac{1}{\sqrt{\alpha}} c_h(v, \varphi)$$

with

$$c_h(v, \varphi) := (C_h^* v_2, C_h^* \varphi_2)_Q - (Iv_1, I\varphi_1)_W$$

for $v, \varphi \in V = V_1 \times V_2$, then the reduced version of (4.2) is the following perturbation of the optimality system (3.3):

$$\begin{aligned} & \text{find } \tilde{x}_h = (\tilde{u}_h, \tilde{z}_h) \in V_h \text{ such that} \\ & \qquad \forall \varphi_h = (\varphi_{h,1}, \varphi_{h,2}) \in V_h \quad b_h(\tilde{x}_h, \varphi_h) = -\tfrac{1}{\sqrt{\alpha}} \left(u_d, I\varphi_{h,1}\right)_W \,, \end{aligned} \tag{4.3}$$

where $V_h = V_{h,1} \times V_{h,2}$. Before we proceed to analyze its discretization error, let us give an important class of examples.

Example 4.1 (Discretized controls). We consider a conforming discretization of the control variable. More precisely, replacing Q in (3.1) with a finite-dimensional subspace $Q_h \subset Q$ leads to the discrete optimality system

$$\begin{aligned} \forall \varphi_{h,2} \in V_{h,2} \quad & a(\tilde{u}_h, \varphi_{h,2}) = (\tilde{q}_h, C^* \varphi_{h,2})_Q \,, \\ \forall \varphi_{h,1} \in V_{h,1} \quad & a(\varphi_{h,1}, \tilde{z}_h) = \tfrac{1}{\sqrt{\alpha}} \left(I\tilde{u}_h - u_d, I\varphi_{h,1}\right)_W \,, \\ \forall p_h \in Q_h \quad & (\sqrt{\alpha}\tilde{q}_h, p_h)_Q = -(C^* \tilde{z}_h, p_h)_Q. \end{aligned} \tag{4.4}$$

If we denote by P_h the Q-orthogonal projection onto Q_h, then the third equations mean

$$\tilde{q}_h = -\tfrac{1}{\sqrt{\alpha}} P_h C^* \tilde{z}_h$$

and, therefore, the right-hand side of the first equation can be rewritten as follows:

$$(\tilde{q}_h, C^* \varphi_{h,2})_Q = -\tfrac{1}{\sqrt{\alpha}} \left(P_h C^* \tilde{z}_h, C^* \varphi_{h,2}\right)_Q = -\tfrac{1}{\sqrt{\alpha}} \left(P_h C^* \tilde{z}_h, P_h C^* \varphi_{h,2}\right)_Q .$$

Hence, the reduced version of (4.4) is a special case of (4.3) with

$$C_h^* = P_h C^*.$$

As the bilinear form b_h coincides with b except for using C_h^* in place of C, the non-asymptotic continuity and nondegeneracy properties of b in Section 2-3 immediately carry over by replacing M_C with the operator norm M_{C_h} of C_h^*. In particular, setting $\tilde{M}_h := \max\{M_I, M_{C_h}\}$ and defining

$$\|\varphi\|_{\alpha,h} := M_a \|\varphi\| + \frac{\tilde{M}_h}{\sqrt{\alpha}} |\varphi|, \tag{4.5}$$

inequality (2.19) yields

$$|b_h(v,\varphi)| \le M_a \|v\| \|\varphi\| + \frac{\tilde{M}_h}{\sqrt{\alpha}} \|v\| |\varphi| \le \|v\| \|\varphi\|_{\alpha,h} \tag{4.6}$$

for all $v, \varphi \in V$. Furthermore, (3.11) and the inf-sup duality (2.18) for $b_{h|V_h \times V_h}$ imply

$$\sup_{\varphi_h \in V_h, \|\varphi_h\|_{\alpha,h}=1} b_h(v_h, \varphi_h) \ge \frac{1}{\tilde{\kappa}_h \mu_h} \|v_h\|, \tag{4.7}$$

for all $v_h \in V_h$, where

$$\tilde{\kappa}_h = \frac{1+2\tilde{L}}{1+\tilde{L}} \left(1 + \tilde{M}_h \mu_h \left(1 + \frac{2\tilde{M}_h}{\sqrt{\alpha}}\right)\right) \quad \text{with} \quad \tilde{L} = \frac{\tilde{M}_h}{\sqrt{\alpha}} \tag{4.8}$$

and μ_h is the quasi-best-approximation constant of the constraint discretization.

Since the structures of the discrete problems (4.3) and (3.3) are the same, well-posedness of (4.3) follows from Lemma 3.1.

4.2. **Approximation.** As in the error analysis of Section 3.2, we adopt the convenient choice

$$\text{(3.6) as norm in } V_2.$$

Here we start our analysis by splitting the error into an approximation part and a consistency part.

Lemma 4.2 (Approximation and consistency error)**.** *Let $x = (u, z)$ be any solution of the optimality system* (2.11) *and let $\tilde{x}_h$ be its approximation from* (4.3)*. Then the error satisfies*

$$\begin{aligned} \|x - \tilde{x}_h\| &\le \tilde{\kappa}_h \mu_h \left(\inf_{v_h \in V_h} \|x - v_h\| + \frac{1}{\sqrt{\alpha}} \sup_{\varphi_h \in V_h} \frac{\left\langle (C_h C_h^* - CC^*) z, \varphi_{h,2} \right\rangle_2}{\|\varphi_h\|_{\alpha,h}} \right) \\ &\le 2\tilde{\kappa}_h \mu_h \|x - \tilde{x}_h\|. \end{aligned}$$

Here $\tilde{\kappa}_h$ is defined by (4.8) *and μ_h is the quasi-best-approximation constant of the constraint discretization from* (3.10)*.*

Proof. Define $x_h^* \in V_h$ by

$$\forall \varphi_h \in V_h \quad b_h(x_h^*, \varphi_h) = b_h(x, \varphi_h).$$

Then Theorem 3.2 with b_h, x_h^*, $\tilde{\kappa}_h$ in place of b, x_h, κ_h gives

$$\|x - x_h^*\| \le \tilde{\kappa}_h \mu_h \inf_{v_h \in V_h} \|x - v_h\|$$

and we have the identities

$$\begin{aligned} b_h(x_h^* - \tilde{x}_h, \varphi_h) &= b_h(x - \tilde{x}_h, \varphi_h) = b_h(x, \varphi_h) - b(x, \varphi_h) \\ &= \frac{1}{\sqrt{\alpha}} \langle C_h C_h^* z - CC^* z, \varphi_{h,2} \rangle_2 \end{aligned}$$

for all $\varphi_h \in V_h$. In view of (4.6) and (4.7), these identities imply

$$\frac{1}{\tilde{\kappa}_h \mu_h} \|x_h^* - \tilde{x}_h\| \leq \frac{1}{\sqrt{\alpha}} \sup_{\varphi_h \in V_h} \frac{\left\langle (C_h C_h^* - CC^*) z, \varphi_{h,2} \right\rangle_2}{\|\varphi_h\|_{\alpha,h}} \leq \|x - \tilde{x}_h\| .$$

The claim follows from the obvious inequalities $\|x - \tilde{x}_h\| \leq \|x - x_h^*\| + \|x_h^* - \tilde{x}_h\|$ and $\inf_{v_h \in V_h} \|x - v_h\| \leq \|x - \tilde{x}_h\|$. □

For the next corollary it is necessary to consider a sufficiently large class of optimization problems, e.g., the class $\mathcal{P}$ of optimization problems, where a constraint can be of the form $Au = Cq + f$ for some $f \in V_2^*$ and I^* may be surjective.

Corollary 4.3 (Necessary condition for quasi-best approximation)**.** *If the approximate variational discretization* (4.3) *is quasi-best in the class $\mathcal{P}$, then*

$$\forall v_{2,h} \in V_{2,h} \quad \|C_h^* v_{2,h}\|_Q = \|C^* v_{2,h}\|_Q .$$

Proof. Let $v_{2,h} \in V_{2,h}$ be arbitrary and take some $v_{1,h} \in V_{1,h}$. Then $v_h = (v_{1,h}, v_{2,h}) \in V_h \subset V$ is a possible solution in the class $\mathcal{P}$. Since (4.3) is quasi-best in $\mathcal{P}$, the discrete solution is exactly $v_h \in V_h$. Hence, by Lemma 4.2 we have $(C_h C_h^* - CC^*) v_{2,h} = 0$, which yields $\|C_h^* v_{h,2}\|_Q = \|C^* v_{h,2}\|_Q$. □

Although possible, it is difficult to imagine that a practical approximation C_h^* satisfies the condition in Corollary 4.3 without coinciding with C. We therefore consider in what follows only assumptions on C_h^* that lead to asymptotic quasi-best approximation. In view of Lemma 4.2, this requires, that the consistency error vanishes at least as fast as the best approximation error, i.e.,

$$\sup_{\varphi_h \in V_h} \frac{\left\langle (C_h C_h^* - CC^*) z, \varphi_{h,2} \right\rangle_2}{\|\varphi_h\|_{\alpha,h}} = o\left(\inf_{v_h \in V_h} \|x - v_h\| \right). \tag{4.9}$$

Moreover, to capture in the limit the compactness of C^* resulting from assumption (3.15a), we assume that

$$d_h \rightharpoonup 0 \text{ weakly in } V_2 \text{ as } h \to 0 \quad \Longrightarrow \quad C_h^* d_h \to 0 \text{ strongly as } h \to 0. \tag{4.10}$$

This implies that the operator norms $\|C_h^*\|_{L(V_2,Q)} = \tilde{M}_h = \max\{M_I, M_{C_h}\}$ are uniformly bounded. Indeed, suppose that $\tilde{M}_h \to \infty$ as $h \to 0$ and, for each $h > 0$, let $\varphi_2^h \in V_2$ be such that $\|C_h^* \varphi_2^h\|_Q = \tilde{M}_h$ and $\|\varphi_2^h\|_2 = 1$. Then $\varphi_2^h / \tilde{M}_h \to 0$ in V_2 as $h \to 0$, which, in view of (4.10), yields a contradiction. Consequently,

$$\tilde{M} := \sup_h \tilde{M}_h = \sup_h \max\{M_I, M_{C_h}\}$$

is finite.

Lemma 4.4 (Qualitative asymptotic quasi-best approximation with approximate control-action)**.** *Let $x = (u, z) \in V$ be a solution to problem* (2.11) *and let $\tilde{x}_h = (\tilde{u}_h, \tilde{z}_h) \in V_h$, $h > 0$, be the corresponding approximations given by* (4.3)*. Furthermore, assume uniform stability* (3.13)*, approximability* (3.15b)*, limiting compactness* (4.10)*, and that $I : V_1 \to W$ is compact. If the exact solution x satisfies* (4.9)*, we have*

$$\|x - \tilde{x}_h\| \leq \mu_h \left(1 + \frac{\tilde{\kappa}}{\sqrt{\alpha}} \, o(1) \right) \inf_{v_h \in V_h} \|x - v_h\| \quad \text{as } h \to 0,$$

where

$$\tilde{\kappa} = \frac{1 + 2\tilde{L}}{1 + \tilde{L}} \left(1 + \tilde{M} \bar{\mu} \left(1 + \frac{2\tilde{M}}{\sqrt{\alpha}} \right) \right) < \infty \quad \text{with} \quad \tilde{L} = \frac{\tilde{M}}{\sqrt{\alpha}}.$$

Proof. As in the proof of Lemma 4.2, define $x_h^* \in V_h$ by

$$\forall \varphi_h \in V_h \quad b_h(x_h^*, \varphi_h) = b_h(x, \varphi_h).$$

We deduce

$$\|x - x_h^*\| \leq \mu_h \big(1 + \tilde{\kappa}_h \, o(1)\big) \inf_{v_h \in V_h} \|x - v_h\| \tag{4.11}$$

by replacing b with b_h and x_h with x_h^* in Lemma 3.5 and using the limiting compactness (4.10) instead of the compactness of $C^* : V_2 \to Q$ in the proof of Lemma 3.6. Next, proceeding as in the proof of Lemma 4.2, assumption (4.9) on the exact solution gives

$$\frac{\sqrt{\alpha}}{\tilde{\kappa}\mu} \|x_h^* - \tilde{x}_h\| = o\left(\inf_{v_h \in V_h} \|x - v_h\| \right).$$

We therefore conclude by inserting the two preceding relationships into the triangle inequality $\|x - \tilde{x}_h\| \leq \|x - x_h^*\| + \|x_h^* - \tilde{x}_h\|$. □

We turn to prove a quantitative quasi-best approximation result. To this end, we need to specify the qualitative assumptions (4.9) and (4.10) by quantitative ones. We shall assume that

$$\sup_{\varphi_h \in V_h} \frac{\big\langle (C_h C_h^* - CC^*) z, \varphi_{h,2} \big\rangle_2}{\|\varphi_h\|_{\alpha,h}} = O(h^\delta) \inf_{v_h \in V_h} \|x - v_h\| \tag{4.12}$$

and that

$$C_h \in L(Q, H^{-s_2+\delta}) \quad \text{is uniformly bounded with respect to } h > 0, \tag{4.13}$$

where $\delta > 0$ is suitably chosen. Note that (4.13) reduces for $C_h = C$ to the part regarding C in the quantitative counterpart (3.17b) of the qualitative compactness (3.15a).

Theorem 4.5 (Quantitative asymptotic quasi-best approximation with approximate control-action)**.** *Let x, $\tilde{x}_h$, $h > 0$, and $\tilde{\kappa}$ be as in Lemma 4.4. In addition, assume uniform stability* (3.13) *and that there exists $\delta > 0$ such that we have* (3.17)*, where* (4.13) *replaces the assumption on C in* (3.17b)*. If the exact solution x satisfies also* (4.12) *with the same δ, we have*

$$\|x - \tilde{x}_h\| \leq \mu_h \left(1 + \frac{\tilde{\kappa}}{\sqrt{\alpha}} O(h^\delta) \right) \inf_{v_h \in V_h} \|x - v_h\| \quad \text{as } h \to 0,$$

Proof. We follow the lines of the proof of Lemma 4.4, but replacing (4.9) with (4.12) and (4.11) with a quantitative argument in the spirit of Theorem 3.7. To this end, it suffices to use (4.13) instead of (3.17b). □

We conclude this section by assessing the key assumptions (4.9) and (4.12) by a remark and an example.

Remark 4.6 (Ensuring dominated consistency error)*.* As

$$\sup_{\varphi_h \in V_h} \frac{\big\langle (C_h C_h^* - CC^*) z, \varphi_{h,2} \big\rangle_2}{\|\varphi_h\|_{\alpha,h}} \leq \|C_h C_h^* - CC^*\|_{L(V_2, V_2^*)}$$

for

$$\|C_h C_h^* - CC^*\|_{L(V_2, V_2^*)} := \sup_{\varphi_h \in V_h} \frac{\big\langle (C_h C_h^* - CC^*) z, \varphi_{h,2} \big\rangle_2}{\|\varphi_h\|_2},$$

we may verify assumptions (4.9) and (4.12) using relationships for $\|C_h C_h^* - CC^*\|_{L(V_2, V_2^*)}$.

Example 4.7 (Simple model optimization and piecewise constant controls). Consider the setting of Example 3.8, but with problem (1.1) with linear finite elements for the constraint and piecewise constants for the control variable. In the light of Example 4.1, this full discretization can be cast into (4.3) with $C_h = P_h C$, where P_h is the L^2-projection onto piecewise constants. By duality, we have

$$\|C_h C_h^* - CC^*\|_{L(V_2,V_2^*)} \leq c_1 h^2,$$

where c_1 depends on the shape regularity of the underlying meshes. Suppose that there is a constant c_2 such that

$$\inf_{v_h \in V_h} \|x - v_h\| \geq c_2 h.$$

This holds for example if the matrix norm of the Hessian of the exact state or its adjoint state are bounded away from 0 in a fixed subdomain. We conclude

$$\|C_h C_h^* - CC^*\|_{L(V_2,V_2^*)} \leq c_1 h^2 \leq \frac{c_1}{c_2} h \inf_{v_h \in V_h} \|x - v_h\|,$$

i.e., (4.12) with $\delta = 1$ and a constant depending on the exact solution under consideration.

5. Analysis with Control Constraints

This section generalize our approach to optimization problems that are nonlinear because of constraints on the control.

5.1. **Control constraints and discretization.** Let $K \subset Q$ be the set of admissible controls. We assume that

$$K \text{ is nonempty, closed, and convex} \tag{5.1}$$

and denote by $\Pi_K : Q \to K$ the projection operator onto K which is characterized by $\|q - \Pi_K q\|_Q = \inf_{p \in K} \|q - p\|_Q$ or, equivalently, by

$$\forall p \in K \quad (q - \Pi_K q, \Pi_K q - p)_Q \geq 0.$$

The latter characterization implies

$$(\Pi_K(q) - \Pi_K(p), q - p)_Q \geq \|\Pi_K(q) - \Pi_K(p)\|_Q^2 \tag{5.2}$$

for all $q, p \in Q$, which in turn shows that the operator Π_K is strongly monotone and Lipschitz continuous, in both cases with constant 1.

The generalization of problem (2.3) incorporating convex control constraints is then the *convex optimization problem*

$$\min_{(q,u) \in K \times V_1} \frac{1}{2} \|Iu - u_d\|_W^2 + \frac{\alpha}{2} \|q\|_Q^2 \quad \text{subject to} \quad Au = Cq. \tag{5.3}$$

Thanks to (5.1), a solution (q, u) is characterized by the existence of $z \in V$ such that the following counterpart of the rescaled optimality system (2.7) is satisfied:

$$Au = Cq, \quad A^* z = \tfrac{1}{\sqrt{\alpha}} I^*(Iu - u_d), \quad q = \Pi_K(-\tfrac{1}{\sqrt{\alpha}} C^* z). \tag{5.4}$$

As in Section 2, we insert the third equation into the first one and consider the corresponding *weak formulation of the rescaled and reduced optimality system*:

$$\text{find } x \in V \text{ such that} \quad \forall \varphi \in V \;\; b_K(x, \varphi) = -\tfrac{1}{\sqrt{\alpha}} (u_d, I\varphi_1)_W, \tag{5.5}$$

where $b_K := a + c_{K,\alpha}$ and

$$c_{K,\alpha}(v, \varphi) := -\left(\Pi_K(-\tfrac{1}{\sqrt{\alpha}} C^* v_2), C^* \varphi_2\right)_Q - \tfrac{1}{\sqrt{\alpha}} (Iv_1, I\varphi_1)_W,$$

which already incorporates the $1/\sqrt{\alpha}$-scaling. In contrast to the previous sections, $c_{K,\alpha}$ and so b_K are in general not linear in the first argument. Nonetheless, if we introduce the pseudometric

$$\delta_{K,\alpha}(v,w)^2 := \alpha \left\| \Pi_K(-\tfrac{1}{\sqrt{\alpha}} C^* v_2) - \Pi_K(-\tfrac{1}{\sqrt{\alpha}} C^* w_2) \right\|_Q^2 + \|I(v_1 - w_1)\|_Q^2 ,$$

inequality (5.2) leads to the following replacement of the properties (2.14) of the bilinear form c: if $v, w \in V$ and $\varphi = \big(-(v_1 - w_1), v_2 - w_2\big)$, then

$$c_{K,\alpha}(v,\varphi) - c_{K,\alpha}(w,\varphi) \geq \frac{1}{\sqrt{\alpha}} \delta_{K,\alpha}(v,w)^2, \tag{5.6a}$$

while, for any $v, w, \varphi \in V$ arbitrary, we have,

$$|c_{K,\alpha}(v,\varphi) - c_{K,\alpha}(w,\varphi)| \leq \frac{1}{\sqrt{\alpha}} \delta_{K,\alpha}(v,w)\, |\varphi| . \tag{5.6b}$$

In addition, we have, for $v, w \in V$,

$$\delta_{K,\alpha}(v,w) \leq |v - w| . \tag{5.7}$$

The continuity bound (5.6b) leads to

$$|b_K(v,\varphi) - b_K(w,\varphi)| \leq d_{K,\alpha}(v,w)\, \|\varphi\| \tag{5.8}$$

with the metric

$$d_{K,\alpha}(v,w) := M_a \|v - w\| + \frac{M}{\sqrt{\alpha}} \delta_{K,\alpha}(v,w), \quad v, w \in V.$$

Notice that the role of the two arguments of c and b_K cannot be interchanged. We adapt (2.22) to this new situation in the following way: given $v, w \in V$, we choose $\varphi = T_K(v - w)$, where $T_K : V \to V$ is the linear operator given by

$$T_K \psi := m_a(A^{-1} J_2 \psi_2, A^{-*} J_1 \psi_1) + \gamma(-\psi_1, \psi_2), \tag{5.9}$$

γ as in (2.23b), and $J_i : V_i \to V_i^*$ is the Riesz map for V_i, $i = 1, 2$. In view of (2.24), we thus obtain the following counterpart of Theorem 2.1.

Theorem 5.1 (Properties of form b_K)**.** *If we equip V as trial space with $d_{K,\alpha}$ and as test space with $\|\cdot\|$, then we have, for any $v, w, \varphi \in V$,*

$$\begin{aligned} b_K\big(v, T_K(v - w)\big) - b_K\big(w, T_K(v - w)\big) &\geq \frac{1 + L}{1 + 2L} \frac{m_a}{M_a} d_{K,\alpha}(v,w) \|v - w\| \\ &\geq \frac{1}{\kappa} \frac{m_a}{M_a} d_{K,\alpha}(v,w) \|T_K(v - w)\| \end{aligned}$$

and

$$|b_K(v,\varphi) - b_K(w,\varphi)| \leq d_{K,\alpha}(v,w) \|\varphi\| ,$$

where κ is defined by (2.23)*.*

Also here, we can conclude existence and uniqueness as a side-product.

Corollary 5.2 (Well-posedness with control constraints)**.** *The optimization problem* (5.5) *has a unique solution.*

Proof. We shall apply the Zarantonello's theorem of strongly monotone operators [27, Theorem 25.B] in the Hilbert space $(V, \|\cdot\|)$. To prepare this, we first observe that

$$T_K \text{ is a linear isomorphism on } (V, \|\cdot\|). \tag{5.10}$$

Indeed, it is continuous with constant $1 + \gamma$ owing to (2.22b) and boundedly invertible on account of the consequence

$$\frac{1 + L}{1 + 2L} \frac{m_a}{M_a} \|v\| \|v\|_\alpha \leq b(T_K v, v) \leq \|T_K v\| \|v\|_\alpha$$

of (2.19) and (2.24) for the bilinear form b. Let us consider the nonlinear operator $\widetilde{B}_K : V \to V^*$ defined by

$$\langle \widetilde{B}_K v,\, \varphi\rangle = b_K(v, T_K\varphi),$$

where $\langle\cdot,\cdot\rangle$ denotes the duality pairing associated with $(V, \|\cdot\|)$. Making use of Theorem 5.1, (2.19) and (5.7), we see that, for all $v, w \in V$,

$$\langle \widetilde{B}_K v - \widetilde{B}_K w, v - w\rangle \geq \frac{1+L}{1+2L} m_a \|v-w\|^2$$

and

$$\langle \widetilde{B}_K v - \widetilde{B}_K w, \varphi\rangle \leq \left(M_a + \frac{M^2}{\sqrt{\alpha}}\right)(1+\gamma)\,\|v-w\|\,\|\varphi\|\,.$$

Hence, $\widetilde{B}_K$ is strongly monotone and Lipschitz continuous and therefore boundedly invertible by [27, Theorem 25.B]. In light of (5.10), we can conclude by noting $T_K^{-*}\widetilde{B}_K v = b_K(v,\cdot)$ for all $v \in V$. □

In order to discretize the optimization problem (5.3) with control constraints, we proceed as in Section 3.1. Introducing the discrete space $V_h = V_{h,1} \times V_{h,2}$ as therein, the variational discretization can be characterized as follows:

$$\text{find } x_h \in V_h \text{ such that} \quad \forall \varphi_h \in V_h \;\; b_K(x_h, \varphi_h) = -\frac{1}{\sqrt{\alpha}}(u_d, I\varphi_{h,1})_W. \tag{5.11}$$

Here we need that $\Pi_K(-C^* v_{h,2}/\sqrt{\alpha})$ can be evaluated exactly for $v_{h,2} \in V_{h,2}$. This occurs, for example, when we consider (1.1) with box constraints and discretize with linear finite elements. If Π_K has to be approximated, the subsequent error analysis involves additional technicalities, similar to those addressed in Section 4.

Existence and uniqueness of solutions to (5.11) can be established in a similar way as Corollary 5.2. Using (3.6) as in norm in V_2, the major change is to replace the operator (5.9) by $T_{K,h} : V_h \to V_h$ given by

$$T_{K,h}\psi_h := \frac{1}{\mu_h}(A_h^{-1} J_{h,2}\psi_2, A_h^{-*} J_{h,1}\psi_{h,1}) + \gamma(-\psi_{h,1}, \psi_{h,2}), \tag{5.12}$$

where $A_h v_{h,1} := a(v_{h,1},\cdot)_{|V_{h,2}}$, $v_{h,1} \in V_{h,1}$, is the discrete counterpart of A, $1/\mu_h$ is its inf-sup-constant, γ is as in (2.23), and $J_{h,i} : V_{h,i} \to V_{h,i}^*$ is the Riesz map for $V_{h,i}$, $i = 1, 2$.

5.2. **Quasi-best approximation.** We analyze the quasi-best-approximation properties of the nonlinear variational discretization (5.11), adopting again

$$\text{(3.6) as norm in } V_2.$$

The following non-asymptotic result draws heavily on Theorem 5.1, which needed an α-dependent error notion for V as trial space.

Theorem 5.3 (Non-asymptotic quasi-best approximation with control constraints)**.** *If x_h is the approximation given by* (5.11) *to an arbitrary solution x of* (5.5)*, then its error is quasi-best in V_h in that*

$$d_{K,\alpha}(x, x_h) \leq (\kappa_h \mu_h + 1) \inf_{v_h \in V_h} d_{K,\alpha}(x, v_h),$$

where κ_h and μ_h are as in Theorem 3.2.

Proof. Given any $v_h \in V_h$, we first write

$$d_{K,\alpha}(x, x_h) \leq d_{K,\alpha}(x, v_h) + d_{K,\alpha}(v_h, x_h). \tag{5.13}$$

To bound the second term, we employ Theorem 5.1 with, respectively, V_h, $T_{K,h}$, $1/\mu_h$, 1, and κ_h in place of V, T_K, m_a, M_a, and κ. Writing $\varphi_h = T_{K,h}(v_h - x_h)$, the definitions of x and x_h thus yield,

$$\begin{aligned}\frac{1}{\kappa_h\mu_h} d_{K,\alpha}(v_h, x_h)\,\|\varphi_h\| &\le b_K(v_h,\varphi_h) - b_K(x_h,\varphi_h)\\ &= b_K(v_h,\varphi_h) - b_K(x,\varphi_h) \le d_{K,\alpha}(x,v_h)\,\|\varphi_h\|\end{aligned}$$

and the claimed inequality is established as $T_{K,h}$ is invertible. □

The "+1" in the bound for the quasi-best-approximation constant in Theorem 5.3 arises from the triangle inequality (5.13), which is avoided in deriving in (3.5). Yet, the following asymptotic quasi-best approximation results involving the generalized Ritz projection from (3.10) are not affected by such an augmentation.

Lemma 5.4 (Nonlinear variational and generalized Ritz approximations)**.** *Let x and x_h be as in Theorem 5.3. The generalized Ritz projection $R_h x$ of x and x_h are related by*

$$d_{K,\alpha}(x_h, R_h x) \le \kappa_h\mu_h \frac{M}{\sqrt{\alpha}}\,|x - R_h x|\,,$$

where κ_h and μ_h are as in Theorem 3.2.

Proof. Applying Theorem 5.1 with the setting as in Theorem 5.3, writing $\varphi_h = T_{K,h}(x_h - R_h x)$, and recalling (5.7), we derive

$$\begin{aligned}\frac{1}{\kappa_h\mu_h} d_{K,\alpha}(x_h, R_h x)\,\|\varphi_h\| &\le b_K(x_h,\varphi_h) - b_K(R_h x,\varphi_h)\\ &= b_K(x,\varphi_h) - b_K(R_h x,\varphi_h)\\ &= c_{K,\alpha}(x,\varphi_h) - c_{K,\alpha}(R_h x,\varphi_h) \le \frac{M}{\sqrt{\alpha}}\,|x - R_h x|\,\|\varphi_h\|\end{aligned}$$

and, again thanks to the invertibility of $T_{K,h}$, the proof is finished. □

Let us sharpen Lemma 5.4 with the help of the additional assumptions and arguments from Section 3.3 regarding the linear optimality system.

Theorem 5.5 (Supercloseness to the generalized Ritz approximation)**.** *Let x, x_h, and $R_h x$ be as in Lemma 5.4. Moreover, assume* (3.13) *and define $\bar\kappa$ as in Lemma 3.6. If* (3.15) *holds, then*

$$d_{K,\alpha}(x_h, R_h x) \le \frac{M}{\sqrt{\alpha}}\bar\kappa\bar\mu\, o(\|x - R_h x\|) \text{ as } h \to 0.$$

More specifically, if (3.17) *holds, then*

$$d_{K,\alpha}(x_h, R_h x) \le \frac{M}{\sqrt{\alpha}}\bar\kappa\bar\mu\, O(h^\delta\,\|x - R_h x\|) \text{ as } h \to 0.$$

For the α-dependence of $\bar\kappa$, cf. Remark 2.4.

Proof. In view of Lemma 5.4, it suffices to show $|x - R_h x| = o(\|x - R_h x\|)$. To this end, we modify the argument in Lemma 3.6 slightly; a similar argument has been used by [10] under weaker assumptions on $(V_h)_h$. Let $(h_k)_k$ be any sequence with $\lim_{k\to\infty} h_k = 0$ and, writing k whenever h_k is an index, consider

$$d_k := \begin{cases} \dfrac{x - R_k x}{\|x - R_k x\|}, & \text{if } x \ne R_k x,\\ 0, & \text{otherwise.}\end{cases}$$

The sequence $(d_k)_k$ is bounded in the Hilbert space V by definition. For its weak limit $d \in V$, we have

$$a(d,\varphi) = a(d - d_k,\varphi) + a(d_k,\varphi - \varphi_k)$$

for arbitrary $\varphi \in V$ and $\varphi_k \in V_h$. Consequently, (3.15b), $k \to \infty$, and (2.17) yield $d = 0$. In view of (3.15a), $d_k \to 0$ weakly in V then implies $|d_k| \to 0$.

For the second statement, we just note that the main step of the proof of Theorem 3.7 with $v = x - R_h x$ leads to $|v - R_h v| = O(h^\delta \, \|x - R_h x\|)$. □

In view of the inverse triangle inequality

$$\big|\, \|x - x_h\| - \|x - R_h x\| \,\big| \leq \|x_h - R_h x\| \leq d_{K,\alpha}(x_h, R_h x),$$

Theorem 5.5 readily yields the following asymptotic quasi-best approximation result.

Corollary 5.6 (Asymptotic quasi-best approximation with control constraints)**.** *Let $\nu_{K,h}$ be the quasi-best-approximation constant for the nonlinear variational discretization* (5.11) *with respect to $\|\cdot\|$. Moreover, assume* (3.13) *and define $\bar\kappa$ as in Lemma 3.6. If* (3.15) *holds, then*

$$\nu_{K,h} \leq \mu_h \left(1 + \frac{M}{\sqrt{\alpha}} \bar\kappa \, o(1)\right) \text{ as } h \to 0.$$

More specifically, if (3.17) *holds, then*

$$\nu_{K,h} \leq \mu_h \left(1 + \frac{M}{\sqrt{\alpha}} \bar\kappa \, O(h^\delta)\right) \text{ as } h \to 0.$$

For the α-dependence of $\bar\kappa$, cf. Remark 2.4.

In comparison with Lemma 3.6 and Theorem 3.7, Corollary 5.6 features an additional $M/\sqrt{\alpha}$-factor. This factor stems from the fact that the derivation we went through used an error notion that also incorporates it.

References

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

[2] E. Casas and M. Mateos, *Uniform convergence of the FEM. Applications to state constrained control problems*, Comput. Appl. Math., 21 (2002), pp. 67–100.

[3] E. Casas and F. Tröltzsch, *Error estimates for linear-quadratic elliptic control problems*, in Analysis and Optimization of Differential Systems (Constanta, 2002), Kluwer Acad. Publ., Boston, MA, 2003, pp. 89–100.

[4] K. Chrysafinos and E. N. Karatzas, *Symmetric error estimates for discontinuous Galerkin approximations for an optimal control problem associated to semilinear parabolic PDE's*, Mar. 2012.

[5] K. Chrysafinos and E. N. Karatzas, *Symmetric error estimates for discontinuous Galerkin time-stepping schemes for optimal control problems constrained to evolutionary Stokes equations*, Comput. Optim. Appl., 60 (2015), pp. 719–751.

[6] K. Deckelnick, A. Günther, and M. Hinze, *Finite element approximation of elliptic control problems with constraints on the gradient*, Numer. Math., 111 (2009), pp. 335–350.

[7] K. Deckelnick and M. Hinze, *Convergence of a finite element approximation to a state-constrained elliptic control problem*, SIAM J. Numer. Anal., 45 (2007), pp. 1937–1953.

[8] ———, *Numerical analysis of a control and state constrained elliptic control problem with piecewise constant control approximations*, in Numerical Mathematics and Advanced Applications, 2008, pp. 597–604.

[9] R. S. Falk, *Approximation of a class of optimal control problems with order of convergence estimates*, J. Math. Anal. Appl., 44 (1973), pp. 28–47.

[10] M. Feischl, T. Führer, and D. Praetorius, *Adaptive FEM with optimal convergence rates for a certain class of nonsymmetric and possibly nonlinear problems*, SIAM J. Numer. Anal., 52 (2014), pp. 601–625.

[11] F. D. Gaspoz, P. Morin, and A. Veeser, *A posteriori error estimates with point sources in fractional Sobolev spaces*, Numer. Methods Partial Differential Equations, 33 (2017), pp. 1018–1042.

[12] T. Geveci, *On the approximation of the solution of an optimal control problem governed by an elliptic equation*, RAIRO Anal. Numér., 13 (1979), pp. 313–328.

[13] A. Günther and M. Hinze, *Elliptic control problems with gradient constraints - variational discrete versus piecewise constant controls*, Comput. Optim. Appl., 49 (2011), pp. 549–566.

[14] M. Hinze, *A variational discretization concept in control constrained optimization: The linear-quadratic case*, Comp. Optim. Appl., 30 (2005), pp. 45–61.
[15] J.-L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Die Grundlehren der mathematischen Wissenschaften, Springer, Berlin – Heidelberg – New York, 1. ed., 1971.
[16] K. Malanowski, *Convergence of approximations vs. regularity of solutions for convex, control-constrained optimal-control problems*, Appl. Math. Optim., 8 (1982), pp. 69–95.
[17] C. Meyer, *Error estimates for the finite-element approximation of an elliptic control problem with pointwise state and control constraints*, Control Cybernet., 37 (2008), pp. 51–85.
[18] C. Meyer and A. Rösch, *Superconvergence properties of optimal control problems*, SIAM J. Control Optim., 43 (2004), pp. 970–985.
[19] I. Neitzel and W. Wollner, *A priori L^2-discretization error estimates for the state in elliptic optimization problems with pointwise inequality state constraints*, Numer. Math., 138 (2018), pp. 273–299.
[20] C. Ortner and W. Wollner, *A priori error estimates for optimal control problems with pointwise constraints on the gradient of the state*, Numer. Math., 118 (2011), pp. 587–600.
[21] A. Rösch, *Error estimates for linear-quadratic control problems with control constraints*, Optim. Methods Softw., 21 (2006), pp. 121–134.
[22] A. H. Schatz, *An observation concerning Ritz-Galerkin methods with indefinite bilinear forms*, Math. Comp., 28 (1974), pp. 959–962.
[23] F. Tantardini and A. Veeser, *The L^2-projection and quasi-optimality of Galerkin methods for parabolic equations*, SIAM J. Numer. Anal., 54 (2016), pp. 317–340.
[24] F. Tröltzsch, *Optimale Steuerung partieller Differentialgleichungen*, Vieweg, 1. ed., 2005.
[25] W. Wollner, *A priori error estimates for optimal control problems with constraints on the gradient of the state on nonsmooth polygonal domains*, in Control and Optimization with PDE Constraints, K. Bredies, C. Clason, K. Kunisch, and G. von Winckel, eds., vol. 164 of International Series of Numerical Mathematics, Birkhäuser, 2013, pp. 193–215.
[26] J. Xu and L. Zikatanov, *Some observations on Babuška and Brezzi theories*, Numer. Math., 94 (2003), pp. 195–202.
[27] E. Zeidler, *Nonlinear functional analysis and its applications. II/B*, Springer-Verlag, New York, 1990.