Proper Orthogonal Decomposition for Parameter Estimation of bilinear elliptic problems

Martin Kahlbacher

December 5, 2006

Contents

1 Introduction

In this work we are dealing with estimation problems for a scalar parameter in elliptic partial differential equations (PDEs).
In many applications, the parameter that has to be identified, typically satisfies certain inequality constraints, for physical or technical reasons. Therefore, this parameter identification problem can be formulated as an optimal control problem for a partial differential equation with inequality constraints.
This problem can be solved, for instance, by an augmented Lagrange technique, or by a semi-smooth Newton method, see, e.g., [38]. In our work we concentrate on an augmented Lagrange algorithm, combined with a globalized SQP method, as investigated in [23], [28], and [40], for instance. For the presentation of a numerically inexpensive globalization strategy let us refer to [18]. SQP methods for parameter estimation problems are investigated, e.g., in [16] and [22].
The inequality constraint in the corresponding optimal control problem is handled by the augmented Lagrange function. At each level of the augmented Lagrange method a globalized SQP algorithm is used to solve the equality constrained problem.
Parameter estimation often requires repeated, reliable and real-time prediction of the parameter. Thus, we apply a model reduction of the optimal control problem in order to save computing time. Reduced-basis element methods for parameter dependent elliptic systems are discussed in [5], [33], and [34], for instance. In our work we apply a technique called proper orthogonal decomposition (POD) to derive a reduced-order model of the optimal control problem.
POD is a method for deriving low order models for linear and non-linear systems of differential equations. It is based on projecting the system onto subspaces consisting of basis elements that contain characteristics of the expected solution. In this work the POD basis is derived from solutions to the underlying PDE for different parameter values (we call these solutions the 'snapshots'). Another application of POD is in the field of time-dependent PDEs, where the snapshots are taken on a certain grid of time-instances.
Let us comment on further literature containing applications of POD. It is successfully used in different fields including signal analysis and pattern recognition (see, e.g., [15]), fluid dynamics and coherent structures (see, e.g., [21] and [41]) and more recently in control theory (see, e.g., [2], [4], [30], [31],

and [32]) and inverse problems (see [3], for instance).

The work is organized in the following manner: The setting of the optimal control problem is introduced in Section 2. Here we also define the elliptic partial differential equation and its weak formulation. For the spatial discretization we apply a POD Galerkin approximation which is investigated in detail in Section 3. Some error estimates between the POD-based solution and the exact solution to a given elliptic problem are given in Section 3.4 (see also [27] and [29]). Moreover, we show a few test examples in Section 3.5 which illustrate our theoretical results. We observe that the POD method yields good approximation results for the solution to the partial differential equation. In Section 4 we have a detailed introduction to the parameter estimation problem in infinite-dimensional functional spaces. Finally we show how the augmented Lagrange algorithm (including the SQP-method) can be discretized by a Galerkin ansatz, for instance by POD. In Section 4.5 we show some applications of the algorithms introduced in Section 4.3 for concrete numerical examples. It turns out that the POD-discretization of our proposed solving technique indeed gives us satisfactory results. Compared to the FE-discretized problem, the estimated parameter as well as the optimal state are very similar. Yet, computing times are much smaller than when we are applying the FE method for discretizing the optimal control problem.

Some applications where a POD approximation appears to be useful, include vibroacoustic problems. Another aspect that remains to be investigated is the problem of parameters varying on subdomains.

2 Functional analysis

Let us begin with a brief introduction to functional analysis. In Section 2.1 we will define some spaces that will be useful in this work. They belong to the class of Sobolev spaces. For more details on Sobolev spaces we refer the reader to [14]. In Section 2.2 a few properties of elliptic differential equations are presented. Finally, in Section 2.3 we give a brief introduction to the method of finite elements.
Throughout this work, we are regarding functions that map elements from an open, bounded domain $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, into $\mathbb{R}$.

2.1 Sobolev spaces

Let us define the following function space:

Definition 2.1 *The space $C_c^\infty(\Omega) \subset C^\infty(\Omega)$ contains all smooth functions ϕ with compact support, that means, for a compact subset $A \subset \Omega$*

$$\phi^{(k)}(x) = 0 \text{ for all } x \notin A$$

holds for all $k \in \mathbb{N}$.

We shall write $A \subset\subset \Omega$ provided the closure $\bar{A}$ of A satisfies $\bar{A} \subset \Omega$ and $\bar{A}$ is a compact subset of $\mathbb{R}^d$.
Now we are able to give the definition of a weak partial derivative.

Definition 2.2 *Suppose that*

$$u, v_i \in L^1_{loc}(\Omega) = \left\{ \varphi : \Omega \to \mathbb{R} \;\middle|\; \int_V |\varphi(x)| \, \mathrm{d}x < \infty \text{ for each } V \subset\subset \Omega \right\}$$

and $1 \leq i \leq d$. For $i \in \{1, \ldots, d\}$ we call v_i the weak partial derivative with respect to the i^{th} component of u if

$$\int_\Omega u \frac{\partial \phi}{\partial x_i} \mathrm{d}x = -\int_\Omega v_i \phi \mathrm{d}x$$

for all test functions $\phi \in C_c^\infty(\Omega)$.
The vector of all weak partial derivatives $(v_1, ..., v_d)^T$ is denoted by ∇u and called the gradient of u.

We will operate in the following function spaces which belong to the class of Sobolev spaces:

Definition 2.3 *The functional space $L^2(\Omega)$ is the space of all measurable functions u (see Definition A.11) which satisfy*

$$\int_\Omega |u(x)|^2 \mathrm{d}x < \infty.$$

Remark 2.4 *Supplied with the inner product*

$$\langle u, v\rangle_{L^2(\Omega)} = \int_\Omega u(x)v(x)\mathrm{d}x \text{ for } u, v \in L^2(\Omega)$$

and its induced norm

$$\|u\|_{L^2(\Omega)} = \sqrt{\langle u, u\rangle_{L^2(\Omega)}} \text{ for } u \in L^2(\Omega)$$

$L^2(\Omega)$ is a Hilbert space.

Let $\varphi_i, \phi_i \in L^2(\Omega)$ for $i \in \{1, ..., d\}$, and let $\varphi = (\varphi_1, ..., \varphi_d)^T$ and $\phi = (\phi_1, ..., \phi_d)^T$ be vectors of dimension d. Then the inner product of φ and ϕ is given by the common product topology, i.e.,

$$\langle \varphi, \phi\rangle_{L^2(\Omega)^d} = \sum_{i=1}^{d} \langle \varphi_i, \phi_i\rangle_{L^2(\Omega)}.$$

We next introduce a subspace of $L^2(\Omega)$ which contains smoother functions.

Definition 2.5 *The functional space $H^1(\Omega)$ is the space of all measurable functions u which satisfy*

$$\int_\Omega \Big(|u(x)|^2 + \|\nabla u(x)\|^2_{\mathbb{R}^d}\Big)\mathrm{d}x < \infty,$$

where $\|\nabla u(x)\|_{\mathbb{R}^d}$ stands for the Euclidean norm of the vector $\nabla u(x) \in \mathbb{R}^d$.

Remark 2.6 *Supplied with the inner product*

$$\langle u, v\rangle_{H^1(\Omega)} = \int_\Omega \Big(u(x)v(x) + \nabla u(x) \cdot \nabla v(x)\Big)\mathrm{d}x$$
$$\Big(= \langle u, v\rangle_{L^2(\Omega)} + \langle \nabla u, \nabla v\rangle_{L^2(\Omega)^d} \Big) \text{ for } u, v \in H^1(\Omega)$$

and its induced norm

$$\|u\|_{H^1(\Omega)} = \sqrt{\langle u, u\rangle_{H^1(\Omega)}} \text{ for } u \in H^1(\Omega)$$

$H^1(\Omega)$ *is a Hilbert space.*

Definition 2.7 (Dual space) *The dual space of a Banach space V is the set of all linear and continuous mappings from V into $\mathbb{R}$. It is denoted by V'.*

By $\langle \cdot\,, \cdot \rangle_{V',V}$ we denote the dual pairing between V and its dual space.

2.2 Elliptic systems

The partial differential equations that we regard in this work are in the class of elliptic equations. More precisely, we are dealing with systems of the form

$$-c\Delta u + \beta \cdot \nabla u + qu = f \quad \text{in } \Omega, \tag{2.1a}$$

$$c\frac{\partial u}{\partial n} + \sigma u = g \quad \text{on } \Gamma = \partial\Omega, \tag{2.1b}$$

where $f \in L^2(\Omega)$ and $g \in L^2(\Gamma)$ are bounded inhomogeneities, $c > 0$, $\beta \in \mathbb{R}^d$, $q > 0$ and $\sigma \in \mathbb{R}$ hold, and the function u belongs to $H^1(\Omega)$. Conditions to the parameters c, q, β, and σ in order to ensure the existence of a unique (weak) solution to (2.1) will be given in Proposition 2.10.
In the Robin boundary condition (2.1b), n denotes the outward normal vector.

We introduce the linear operator $L : H^1(\Omega) \to H^1(\Omega)'$ by

$$\langle Lu, \varphi\rangle_{H^1(\Omega)',H^1(\Omega)} = c\int_\Omega \nabla u\cdot\nabla\varphi\,\mathrm{d}x + \sigma\int_\Gamma u\varphi\,\mathrm{d}s + q\int_\Omega u\varphi\mathrm{d}x + \int_\Omega \beta\cdot\nabla u\varphi\mathrm{d}x$$

for $u, \varphi \in H^1(\Omega)$. It follows directly from the proof of Proposition 2.10 that L is also bounded and therefore continuous.
Moreover, let us introduce the linear mapping $F : H^1(\Omega) \to \mathbb{R}$ by

$$\langle F, \varphi\rangle_{H^1(\Omega)',H^1(\Omega)} = \int_\Omega f\varphi\mathrm{d}x + \int_\Gamma g\varphi\mathrm{d}s \text{ for } \varphi \in H^1(\Omega). \tag{2.2}$$

Since $f \in L^2(\Omega)$ and $g \in L^2(\Gamma)$, F is also bounded and therefore, $F \in H^1(\Omega)'$.

Definition 2.8 *The function u is called a weak solution to (2.1) if*

$$Lu = F \text{ in } H^1(\Omega)'. \tag{2.3}$$

In other words,

$$\langle Lu, \varphi \rangle_{H^1(\Omega)', H^1(\Omega)} = \langle F, \varphi \rangle_{H^1(\Omega)', H^1(\Omega)} \text{ for all } \varphi \in H^1(\Omega). \tag{2.4}$$

An important tool in functional analysis is the Lax-Milgram lemma (see [14]) which gives us conditions for the existence and the uniqueness of a weak solution to (2.1).

Theorem 2.9 (Lax-Milgram lemma) *Assume that V is a Hilbert space and*

$$B : V \times V \to \mathbb{R}$$

is a bilinear mapping for which there exist constants $\alpha_1, \alpha_2 > 0$ such that the continuity condition

$$|B(u, v)| \leq \alpha_1 \|u\|_V \|v\|_V \text{ for all } u, v \in V$$

and the coercivity condition

$$\alpha_2 \|u\|_V^2 \leq B(u, u) \text{ for all } u \in V$$

hold. Finally, let $F \in V'$.
Then there exists a unique element $u \in V$ such that

$$B(u, v) = \langle F, v \rangle_{V', V} \text{ for all } v \in V.$$

Let $0 < q_l \leq q_u$ be fixed and let $V = H^1(\Omega)$. For every parameter $q \in \mathcal{I} = [q_l, q_u]$ we introduce the (parametrized) bilinear form $B(\cdot\,, \cdot\,; q) : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ that corresponds to our elliptic differential equation by

$$\begin{aligned} B(u, \phi; q) &= \int_\Omega c \nabla u \cdot \nabla \phi + q u \phi + \beta \cdot \nabla u \phi \, \mathrm{d}x + \sigma \int_\Gamma u \phi \, \mathrm{d}s \\ &= \langle Lu, \phi \rangle_{H^1(\Omega)', H^1(\Omega)} \end{aligned} \tag{2.5}$$

for $u, \phi \in H^1(\Omega)$ and the functional $F \in H^1(\Omega)'$ by (2.2).

Proposition 2.10 *Let $q \in \mathcal{I}$. If*

$$\alpha_2 := \min\left\{\frac{c}{2}, q - \frac{\|\beta\|_{\mathbb{R}^d}^2}{2c}\right\} + \min\{0, \sigma C_\Gamma^2\} > 0, \tag{2.6}$$

holds (with C_Γ denoting the trace constant from Lemma A.14), then there exists a unique weak solution u to (2.1).

Proof. We must check whether the continuity condition and the coercivity condition of the Lax-Milgram lemma hold for the bilinear form $B(\cdot\,,\cdot\,;q)$ with an arbitrary $q \in \mathcal{I}$. Let us first regard the continuity condition. Applying the Cauchy-Schwarz inequality (see Lemma A.13) several times and the trace theorem (see Lemma A.14) twice we have

$$\begin{aligned}
|B(u,v;q)| =& \Big| c\langle\nabla u, \nabla v\rangle_{L^2(\Omega)^d} + \sigma\langle u, v\rangle_{L^2(\Gamma)} + q\langle u, v\rangle_{L^2(\Omega)} \\
&\quad + \int_\Omega \langle\beta, \nabla u(x)\rangle_{\mathbb{R}^d} v(x)\,\mathrm{d}x \Big| \\
\le& c\|\nabla u\|_{L^2(\Omega)^d}\|\nabla v\|_{L^2(\Omega)^d} + |\sigma|\,\|u\|_{L^2(\Gamma)}\|v\|_{L^2(\Gamma)} \\
&\quad + q\|u\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \int_\Omega \|\beta\|_{\mathbb{R}^d}\|\nabla u(x)\|_{\mathbb{R}^d}|v(x)|\,\mathrm{d}x \\
\le& c\|\nabla u\|_{L^2(\Omega)^d}\|\nabla v\|_{L^2(\Omega)^d} + |\sigma|\,C_\Gamma^2\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)} \\
&\quad + q\|u\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \|\beta\|_{\mathbb{R}^d}\big\langle\|\nabla u\|_{\mathbb{R}^d}, |v|\big\rangle_{L^2(\Omega)} \\
\le& c\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)} + |\sigma|C_\Gamma^2\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)} \\
&\quad + q\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)} + \|\beta\|_{\mathbb{R}^d}\|\nabla u\|_{L^2(\Omega)^d}\|v\|_{L^2(\Omega)} \\
\le& c\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)} + |\sigma|C_\Gamma^2\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)} \\
&\quad + q\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)} + \|\beta\|_{\mathbb{R}^d}\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)} \\
=& \big(c + |\sigma|C_\Gamma^2 + q + \|\beta\|_{\mathbb{R}^d}\big)\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)}.
\end{aligned}$$

Thus, by setting $\alpha_1 := c + |\sigma|C_\Gamma^2 + q + \|\beta\|_{\mathbb{R}^d} > 0$, the continuity condition of the Lax-Milgram lemma is satisfied.
Next we aim to find conditions to the parameters in the elliptic equation such that the coercivity condition is satisfied. Let $u \in H^1(\Omega)$ and $q \in \mathcal{I}$.

By applying Young's inequality (Lemma A.2), we find

$$\begin{aligned}
B(u,u;q) =& c\|\nabla u\|^2_{L^2(\Omega)^d} + \sigma\|u\|^2_{L^2(\Gamma)} + q\|u\|^2_{L^2(\Omega)} + \int_\Omega \langle \beta, \nabla u(x)\rangle_{\mathbb{R}^d} u(x)\,\mathrm{d}x \\
\geq& c\|\nabla u\|^2_{L^2(\Omega)^d} - \int_\Omega \|\beta\|_{\mathbb{R}^d}\|\nabla u(x)\|_{\mathbb{R}^d}|u(x)|\,\mathrm{d}x + q\|u\|^2_{L^2(\Omega)} \\
&+ \sigma\|u\|^2_{L^2(\Gamma)} \\
\geq& c\|\nabla u\|^2_{L^2(\Omega)^d} - \|\beta\|_{\mathbb{R}^d}\big\langle \|\nabla u\|_{\mathbb{R}^d}, |u|\big\rangle_{L^2(\Omega)} + q\|u\|^2_{L^2(\Omega)} \\
&+ \sigma\|u\|^2_{L^2(\Gamma)} \\
\geq& c\|\nabla u\|^2_{L^2(\Omega)^d} - \|\beta\|_{\mathbb{R}^d}\left(\frac{c}{2\|\beta\|_{\mathbb{R}^d}}\|\nabla u\|^2_{L^2(\Omega)^d} + \frac{\|\beta\|_{\mathbb{R}^d}}{2c}\|u\|^2_{L^2(\Omega)}\right) \\
&+ q\|u\|^2_{L^2(\Omega)} + \sigma\|u\|^2_{L^2(\Gamma)}.
\end{aligned}$$

If $\sigma \geq 0$ holds, then we have

$$\begin{aligned}
\langle F, u\rangle_{H^1(\Omega)',H^1(\Omega)} = B(u,u;q) &\geq \frac{c}{2}\|\nabla u\|^2_{L^2(\Omega)^d} + \left(q - \frac{\|\beta\|^2_{\mathbb{R}^d}}{2c}\right)\|u\|^2_{L^2(\Omega)} \\
&\geq \min\left\{\frac{c}{2}, q - \frac{\|\beta\|^2_{\mathbb{R}^d}}{2c}\right\}\left(\|\nabla u\|^2_{L^2(\Omega)^d} + \|u\|^2_{L^2(\Omega)}\right) \\
&= \min\left\{\frac{c}{2}, q - \frac{\|\beta\|^2_{\mathbb{R}^d}}{2c}\right\}\|u\|^2_{H^1(\Omega)}.
\end{aligned}$$

Because we presumed that $c > 0$, we only have to ensure that $q - \frac{\|\beta\|^2_{\mathbb{R}^d}}{2c} > 0$. If $\sigma < 0$, however, we deduce from the trace theorem

$$\sigma\|u\|^2_{L^2(\Gamma)} \geq \sigma C_\Gamma^2\|u\|^2_{H^1(\Omega)}$$

and therefore

$$\begin{aligned}
B(u,u;q) &\geq \min\left\{\frac{c}{2}, q - \frac{\|\beta\|^2_{\mathbb{R}^d}}{2c}\right\}\left(\|\nabla u\|^2_{L^2(\Omega)^d} + \|u\|^2_{L^2(\Omega)}\right) + \sigma C_\Gamma^2\|u\|^2_{H^1(\Omega)} \\
&= \left(\min\left\{\frac{c}{2}, q - \frac{\|\beta\|^2_{\mathbb{R}^d}}{2c}\right\} + \sigma C_\Gamma^2\right)\|u\|^2_{H^1(\Omega)}.
\end{aligned}$$

Thus, if we have parameters $c > 0, q > 0, \beta \in \mathbb{R}^d, \sigma \in \mathbb{R}$ such that (2.6) holds, the conditions of the Lax-Milgram lemma are satisfied and (2.1) has a unique weak solution u. ∎

Remark 2.11 *It follows from the proof of Proposition 2.10 that for any* $q \in \mathcal{I}$

$$B(u,u;q) \geq \alpha(q)\|u\|^2_{H^1(\Omega)} \text{ for all } u \in H^1(\Omega)$$

with

$$\alpha(q) = \min\left\{\frac{c}{2}, q - \frac{\|\beta\|^2_{\mathbb{R}^d}}{2c}\right\} + \min\{0, \sigma C_\Gamma^2\} > 0.$$

Corollary 2.12 *Let all assumptions of Proposition 2.10 be satisfied. Then*

$$\|u\|_{H^1(\Omega)} \leq \frac{1}{\alpha_2}\|F\|_{H^1(\Omega)'}$$

holds, where α_2 *is introduced in* (2.6).

Proof. From the proof of Proposition 2.10 we find that

$$\begin{aligned}\alpha_2\|u\|^2_{H^1(\Omega)} &\leq B(u,u;q) = \langle F, u\rangle_{H^1(\Omega)',H^1(\Omega)} \leq \|F\|_{H^1(\Omega)'}\|u\|_{H^1(\Omega)} \\ &\leq \frac{1}{2\alpha_2}\|F\|^2_{H^1(\Omega)'} + \frac{\alpha_2}{2}\|u\|^2_{H^1(\Omega)},\end{aligned}$$

thus we obtain the a-priori estimate for u in the H^1-norm as given in the claim. ∎

2.3 The finite element method

We want to give a brief introduction to the finite element (FE) method in this section. The reader is referred to [10], [11], or [17], for instance, to find a detailed description of the method.
Beside, e.g., the method of finite differences, the FE method is a widely-spread and well-investigated procedure for solving boundary value problems, in particular of the form (2.1).

If the boundary value problem is formulated as a variational equation, this variational equation can, in turn, be discretized and then solved by FE approximations.
We view a function $u : \Omega \to \mathbb{R}$ as a linear combination of a given finite set of linearly independent ansatz functions, the so called 'finite elements'. Therefore we make the Galerkin ansatz

$$u^h(x) = \sum_{i=1}^{n_{FE}} \mathrm{u}_i \varphi_i(x), \tag{2.7}$$

where n_{FE} is the number of finite elements, $\{\varphi_i\}_{i=1}^{n_{FE}}$ is the set of the finite element ansatz functions, and $\{\mathrm{u}_i\}_{i=1}^{n_{FE}}$ is the set of the respective coefficients. These coefficients are the unknowns in the discretized variational equation and have to be computed (numerically).

We define the set V^h as the set of all functions which can be written in the form (2.7), i.e., $V^h = \mathrm{span}\{\varphi_i\}_{i=1}^{n_{FE}}$. Here we choose piecewise linear ansatz functions $\{\varphi_i\}_{i=1}^{n_{FE}}$, so that $V^h \subset H^1(\Omega)$ holds.
Of course, the number of finite elements should be chosen high enough to approximate the exact solution sufficiently well. If we take only a few FE ansatz functions, we often find unsatisfactory results for the approximation u^h. Another possibility to improve the FE approximation is to apply adaptive grids. This is not the focus of this research, though.

The linearly independent finite elements are characterized in the way that each element is unequal zero only in an area around one grid point of the domain Ω. Moreover, every point x in the domain is covered by exactly one FE basis function that has a function value larger than zero at x.
The FE ansatz functions should be chosen as simple as possible in order to prevent high computation times.

The discretization of the domain can be done in numerous ways. Some commonly used variants for two-dimensional problems apply a triangular or a rectangular grid of the domain.

From the boundary value problem (2.1) we obtain the variational equation

$$B(u, v; q) = \langle F, v\rangle_{H^1(\Omega)', H^1(\Omega)} \text{ for all } v \in H^1(\Omega) \text{ and } q \in \mathcal{I}.$$

Now we look at the discrete variational problem

$$B(u^h, v^h; q) = \langle F, v^h\rangle_{H^1(\Omega)', H^1(\Omega)} \text{ for all } v^h \in V^h(\Omega) \text{ and } q \in \mathcal{I}, \tag{2.8}$$

where u^h is as in (2.7) and v^h is an arbitrary linear combination of the finite elements, i.e.,

$$v^h(x) = \sum_{i=1}^{n_{FE}} \mathrm{v}_i \varphi_i(x).$$

The existence of a unique solution u^h to (2.8) follows again from the Lax-Milgram lemma, where we require the same hypothesis for the constants c,

q, σ, and β, because we operate in a closed subset of $H^1(\Omega)$.

Of course, problem (2.8) is linear in v^h and the finite elements themselves are in the space of linear combinations of the ansatz functions $\{\varphi_i\}_{i=1}^{n_{FE}}$. Thus, (2.8) is equivalent to

$$B(u^h, \varphi_j; q) = \langle F, \varphi_j \rangle_{H^1(\Omega)', H^1(\Omega)}, \quad 1 \leq j \leq n_{FE}.$$

Inserting (2.7) yields

$$B\Big(\sum_{i=1}^{n_{FE}} \mathrm{u}_i \varphi_i(x), \varphi_j; q \Big) = \langle F, \varphi_j \rangle_{H^1(\Omega)', H^1(\Omega)}, \quad 1 \leq j \leq n_{FE}.$$

Because $B(\cdot, \cdot; q)$ is a bilinear form for every $q \in \mathcal{I}$ we deduce

$$\sum_{i=1}^{n_{FE}} B(\varphi_i, \varphi_j; q) \mathrm{u}_i = \langle F, \varphi_j \rangle_{H^1(\Omega)', H^1(\Omega)}, \quad 1 \leq j \leq n_{FE}. \tag{2.9}$$

For a given bilinear form $B(\cdot, \cdot; q)$ and a given functional F we are able to compute the terms

$$\mathrm{B}_{ij} = B(\varphi_i, \varphi_j; q) \text{ for } 1 \leq i, j \leq n_{FE} \tag{2.10}$$

and

$$\mathrm{F}_j = \langle F, \varphi_j \rangle_{H^1(\Omega)', H^1(\Omega)} \text{ for } 1 \leq j \leq n_{FE} \tag{2.11}$$

for the chosen FE functions $\{\varphi_k\}_{k=1}^{n_{FE}}$. Therefore we obtain n_{FE} equations for the n_{FE} unknowns $\mathrm{u}_1, ..., \mathrm{u}_{n_{FE}}$ and we can solve the linear equation system $\mathrm{B}^T \mathrm{u} = \mathrm{F}$ uniquely, with $\mathrm{B} = ((\mathrm{B}_{ij})) \in \mathbb{R}^{n_{FE} \times n_{FE}}$, $\mathrm{u} = (\mathrm{u}_i) \in \mathbb{R}^{n_{FE}}$, and $\mathrm{F} = (\mathrm{F}_j) \in \mathbb{R}^{n_{FE}}$ for u, provided B is regular. This follows from the Lax-Milgram lemma.

3 The POD method

The problem which we are facing in this work (see Section 4) is an infinite-dimensional optimization problem. Discretizing the problem by, for instance, the finite element or finite difference method, we usually obtain a large-scale discretized problem. This might lead to troubles regarding computational aspects such as storage limits or extremely high computing times. Our goal is to approximate the high-dimensional model by a low-dimensional model. Several model reduction methods are known so far. Some of them are based on the singular value decomposition (SVD), like balanced truncation, Hankel-norm approximation, singular perturbation, or proper orthogonal decomposition (POD). Compared to balanced truncation, POD can also be applied to nonlinear systems. Another prominent class of model reduction methods uses 'Approximation by moment matching', which includes, for instance, the Lanczos algorithm, the Arnoldi algorithm, and the cross grammian. SVD based methods usually show a globally better approximation, while moment matching methods approximate the system locally better. For a survey of many of these methods see [1].

In this section we focus on the method of POD. It is a powerful technique for model reduction of linear and non-linear systems. It is based on a Galerkin type discretization with basis elements created from the system itself. This is in contrast to, e.g., finite element techniques, where the elements of the subspaces are uncorrelated to the physical properties of the system that they approximate.
The approach of the POD method is that we calculate some snapshots of specific solutions to our system. These might be solutions of a non-stationary differential equation at certain time instances, or, as we have in this work, solutions of our parameter-dependent elliptic problem for certain parameter values. These snapshots are used to compute basis functions that approximate the system we want to solve in an efficient and reliable way. In many cases we only need a small number of basis functions (relative to the precision of the posed problem) for computing POD solutions that are really close to the corresponding FE solutions. The advantage is that the linear and non-linear systems which arise when using POD are much smaller than those when using only the FE method. Hence, computing times are often significantly smaller in the POD method, which makes this method very practicable in applications where equation systems have to be solved fast

and repeatedly. Another positive aspect of the POD method is the saving of memory.

In Section 3.1 we introduce the POD method in finite dimensional vector spaces. This is followed by an introduction to the POD method in general Hilbert spaces in Section 3.2. Afterwards we will investigate the POD method in discretized functional spaces, in particular we apply the FE method for the discretization of the Hilbert space $H^1(\Omega)$. In Section 3.4 some results concerning error estimates between the POD solution and the exact solution are presented. Finally, some numerical examples that confirm our theoretical results are presented in Section 3.5, including a strategy called the Experimental Order of Decay (EOD) – see [19].

3.1 POD in vector spaces

Let $Y = [y_1, ..., y_n] \in \mathbb{R}^{m \times n}$ be a given matrix, where the vectors y_i denote the so called snapshots for $i = 1, ..., n$.
First we recall the singular value decomposition (SVD) of a matrix:

Theorem 3.1 (SVD) *For a (real valued) matrix $Y \in \mathbb{R}^{m \times n}$ with*

$$r := \operatorname{rank}(Y) \leq \min\{m, n\}$$

there exist r positive singular values $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0$ together with orthogonal matrices $U = [u_1, ..., u_m] \in \mathbb{R}^{m \times m}$ and $V = [v_1, ..., v_n] \in \mathbb{R}^{n \times n}$, which satisfy

$$U^T Y V = \Sigma \in \mathbb{R}^{m \times n}.$$

The matrix Σ has the entries $\Sigma(i, i) = \sigma_i$ for $1 \leq i \leq r$, all other components are zero. We define the matrix R by $R = \operatorname{diag}(\sigma_1, ..., \sigma_r) \in \mathbb{R}^{r \times r}$. Moreover,

$$Y v_i = \sigma_i u_i \text{ and } Y^T u_i = \sigma_i v_i \text{ hold for } 1 \leq i \leq r.$$

For a proof of this theorem, let us refer to [13], for instance.

Because of

$$Y^T Y v_i = Y^T \sigma_i u_i = \sigma_i^2 v_i$$

and

$$Y Y^T u_i = Y \sigma_i v_i = \sigma_i^2 u_i$$

the singular vectors $\{v_i\}_{i=1}^r$ and $\{u_i\}_{i=1}^r$ are eigenvectors for Y^TY and YY^T, respectively. Their associated eigenvalues satisfy $\lambda_i = \sigma_i^2$.
We have $Y = U\Sigma V^T$, and equivalently $Y = U^r R(V^r)^T$, where U^r and V^r denote those matrices which consist of the first r columns of U and V, respectively.

Moreover, we have $U^r R(V^r)^T = U^r B^r$ with $B^r = R(V^r)^T \in \mathbb{R}^{r\times n}$. We now express the column vector y_j ($j \in \{1, ..., n\}$) as a linear combination of the r vectors $u_1, ..., u_r$:

$$\begin{aligned} y_j &= \sum_{i=1}^r B_{ij}^r u_i = \sum_{i=1}^r (R(V^r)^T)_{ij} u_i = \sum_{i=1}^r ((U^r)^T U^r R(V^r)^T)_{ij} u_i \\ &= \sum_{i=1}^r ((U^r)^T Y)_{ij} u_i = \sum_{i=1}^r \Big(\sum_{k=1}^m U_{ki}^r Y_{kj} \Big) u_i = \sum_{i=1}^r \langle u_i, y_j \rangle_{\mathbb{R}^m} u_i. \end{aligned}$$

It follows that $B_{ij}^r = \langle u_i, y_j \rangle_{\mathbb{R}^m}$.
We will use this equality later when we aim to approximate the vectors y_j in an optimal manner by an orthonormal set of basis vectors that is of a smaller size than r.

Next we fix $\ell \in \{1, ..., r\}$. The POD basis of rank ℓ is given by the solution to the maximization problem

$$\max_{u_1,...,u_\ell} \sum_{i=1}^{\ell} \sum_{j=1}^{n} |\langle y_j, u_i \rangle_{\mathbb{R}^m}|^2 \text{ subject to (s.t.) } \langle u_i, u_j \rangle_{\mathbb{R}^m} = \delta_{ij}, \tag{3.1}$$

i.e., we look for ℓ orthonormal vectors vectors $u_1, ..., u_\ell$ which maximize the projection of the columns $\{y_i\}_{i=1}^n$ in the mean. It turns out that SVD can be used to solve (3.1).

First of all, let us consider the optimization problem for $\ell = 1$, that means we look for only one POD basis function:

$$\max_{u} \sum_{j=1}^{n} |\langle y_j, u \rangle_{\mathbb{R}^m}|^2 \text{ s.t. } \|u\|_{\mathbb{R}^m}^2 = 1. \tag{3.2}$$

We define the equality constraint $e : \mathbb{R}^m \to \mathbb{R}$ by

$$e(u) = 1 - \|u\|_{\mathbb{R}^m}^2.$$

Lemma 3.2 *Problem* (3.2) *admits a local solution* $u^* \in \mathbb{R}^m$.

Proof. The set of feasible solutions is non-empty because there are vectors that satisfy the equality condition $\|u\|^2_{\mathbb{R}^m} = 1$ (for example, $u = e_1 = (1, 0, ..., 0)^T \in \mathbb{R}^m$). We deduce that there is a maximizing sequence $\{u^k\}_{k\in\mathbb{N}}$ in $\mathbb{R}^m$ in the (bounded) set of feasible solutions with

$$\lim_{k\to\infty} \sum_{j=1}^{n} |\langle y_j, u^k\rangle_{\mathbb{R}^m}|^2 = \sup_{u\in\mathbb{R}^m} \left\{ \sum_{j=1}^{n} |\langle y_j, u\rangle_{\mathbb{R}^m}|^2 \middle| \|u\|^2_{\mathbb{R}^m} = 1 \right\}.$$

Due to the Bolzano-Weierstrass theorem every bounded real sequence contains a convergent subsequence $\{u^{k_i}\}_{i\in\mathbb{N}}$. Thus, there exists a $u^* \in \mathbb{R}^m$ such that

$$\lim_{i\to\infty} u^{k_i} = u^* \text{ in } \mathbb{R}^m,$$

which implies

$$\lim_{i\to\infty} \langle y_j, u^{k_i}\rangle_{\mathbb{R}^m} = \langle y_j, u^*\rangle_{\mathbb{R}^m} \text{ for any } j \in \{1, ..., n\}.$$

Therefore, the sum of the squared norms over all indices j also converges as $i \to \infty$:

$$\lim_{i\to\infty} \sum_{j=1}^{n} |\langle y_j, u^{k_i}\rangle_{\mathbb{R}^m}|^2 = \sum_{j=1}^{n} |\langle y_j, u^*\rangle_{\mathbb{R}^m}|^2.$$

Since $\|u^*\|^2_{\mathbb{R}^m} = 1$ holds, we have proved that u^* is a solution to the maximization problem (3.2). ∎

Next we prove that a solution u^* to (3.2) is a regular point – see Definition A.18. Thus, we need to show that the gradient of the equality constraint $e(u^*) = 0$ is surjective.

Lemma 3.3 *The operator* $\nabla e(u)$ *is surjective.*

Proof. We have $\nabla e(u^*) = -2u^*$. Since any feasible solution satisfies $\|u^*\|_{\mathbb{R}^m} = 1$, $u^* \neq 0$ and therefore $\nabla e(u^*) = -2u^* \neq 0$. Obviously, $\nabla e(u^*)$ is surjective. ∎

It follows from Lemma 3.3 and Definition A.18 that any solution u^* to (3.2) is a regular point.

The Lagrange function associated to (3.2) is given by

$$\mathcal{L}(u,\lambda) = \sum_{j=1}^{n} |\langle y_j, u\rangle_{\mathbb{R}^m}|^2 + \lambda(1 - \|u\|_{\mathbb{R}^m}^2)$$

for $u \in \mathbb{R}^m$ and $\lambda \in \mathbb{R}$ (see Theorem A.20). In order to obtain first-order necessary optimality conditions we build the derivatives of the Lagrangian with respect to an arbitrary component u_i, $i \in \{1, ..., m\}$, of the vector u:

$$\begin{aligned}
\nabla_{u_i}\mathcal{L}(u,\lambda) = \frac{\partial \mathcal{L}(u,\lambda)}{\partial u_i} &= \frac{\partial}{\partial u_i}\Bigg(\sum_{j=1}^{n} |\langle y_j, u\rangle_{\mathbb{R}^m}|^2 + \lambda\Big(1 - \|u\|_{\mathbb{R}^m}^2\Big)\Bigg) \\
&= \frac{\partial}{\partial u_i}\Bigg(\sum_{j=1}^{n}\Big|\sum_{k=1}^{m} Y_{kj}u_k\Big|^2 + \lambda\Big(1 - \sum_{k=1}^{m} u_k^2\Big)\Bigg) \\
&= 2\sum_{j=1}^{n}\Bigg(\sum_{k=1}^{m} Y_{kj}u_k\Bigg)Y_{ij} - 2\lambda u_i = 2\Bigg(\sum_{k=1}^{m}\sum_{j=1}^{n} Y_{ij}(Y^T)_{jk}u_k - \lambda u_i\Bigg) \\
&= 2\Bigg(\sum_{k=1}^{m}(YY^T)_{ik}u_k - \lambda u_i\Bigg) = 2((YY^Tu)_i - \lambda u_i).
\end{aligned}$$

Regarding these equations for i from 1 to m together, setting $\nabla_{u_i}\mathcal{L}(u,\lambda) = 0$ for each i from 1 to m, and finally multiplying the equations with one half we obtain the eigenvalue problem:

$$YY^Tu = \lambda u. \tag{3.3}$$

We observe that the matrix $YY^T \in \mathbb{R}^{m\times m}$ is symmetric and positive semi-definite because of

$$(YY^T)^T = (Y^T)^TY^T = YY^T$$

and

$$x^T(YY^T)x = (Y^Tx)^T(Y^Tx) = \|Y^Tx\|_{\mathbb{R}^n}^2 \geq 0 \text{ for any } x \in \mathbb{R}^m.$$

The symmetry of YY^T yields that all eigenvalues are real, and because of the positive semi-definiteness we observe for any eigenvector v

$$0 \leq v^TYY^Tv = v^T(\lambda v) = \lambda\|v\|_{\mathbb{R}^m}^2,$$

and therefore $\lambda \geq 0$ holds for any eigenvalue λ.

In fact, any eigenvector of YY^T satisfies the first-order necessary optimality condition (3.3) together with its corresponding eigenvalue as the Lagrange multiplier. We can prove though that the eigenvector u_1 corresponding to the largest eigenvalue λ_1 of the matrix YY^T actually solves the maximization problem (3.2).

Theorem 3.4 *The eigenvector corresponding to the largest eigenvalue of the matrix YY^T is a global solution to* (3.2).

Proof. We evaluate the cost function for the eigenvector u_1 corresponding to the largest eigenvalue of YY^T:

$$\begin{aligned}
\sum_{j=1}^{n} |\langle y_j, u_1\rangle_{\mathbb{R}^m}|^2 &= \sum_{j=1}^{n} \langle y_j, u_1\rangle_{\mathbb{R}^m} \langle y_j, u_1\rangle_{\mathbb{R}^m} \\
&= \left\langle \sum_{j=1}^{n} \langle y_j, u_1\rangle_{\mathbb{R}^m} y_j, u_1 \right\rangle_{\mathbb{R}^m} \\
&= \left\langle \sum_{j=1}^{n}\sum_{k=1}^{m} Y_{kj}(u_1)_k Y_{\cdot,j}, u_1 \right\rangle_{\mathbb{R}^m} \\
&= \left\langle \sum_{k=1}^{m}\sum_{j=1}^{n} Y_{\cdot,j}(Y^T)_{jk}(u_1)_k, u_1 \right\rangle_{\mathbb{R}^m} \\
&= \langle YY^T u_1, u_1\rangle_{\mathbb{R}^m} \\
&= \langle \lambda_1 u_1, u_1\rangle_{\mathbb{R}^m} = \lambda_1 \langle u_1, u_1\rangle_{\mathbb{R}^m} = \lambda_1 = \sigma_1^2.
\end{aligned} \tag{3.4}$$

We extend u_1 to an orthonormal basis $\{u_i\}_{i=1}^m$ in $\mathbb{R}^m$. Then we can write any vector as its Fourier sum utilizing the basis $\{u_1, ..., u_m\}$. Let $\tilde{u}$ be any vector in $\mathbb{R}^m$ satisfying $\|\tilde{u}\|_{\mathbb{R}^m} = 1$. From

$$\tilde{u} = \sum_{i=1}^{m} \langle \tilde{u}, u_i\rangle_{\mathbb{R}^m} u_i$$

and (3.4) we infer

$$\begin{aligned}
\sum_{j=1}^{n} |\langle y_j, \tilde{u}\rangle_{\mathbb{R}^m}|^2 &= \sum_{j=1}^{n} \left| \left\langle y_j, \sum_{i=1}^{m} \langle \tilde{u}, u_i\rangle_{\mathbb{R}^m} u_i \right\rangle_{\mathbb{R}^m} \right|^2 \\
&= \sum_{j=1}^{n} \left\langle y_j, \sum_{i=1}^{m} \langle \tilde{u}, u_i\rangle_{\mathbb{R}^m} u_i \right\rangle_{\mathbb{R}^m} \left\langle y_j, \sum_{k=1}^{m} \langle \tilde{u}, u_k\rangle_{\mathbb{R}^m} u_k \right\rangle_{\mathbb{R}^m} \\
&= \sum_{i=1}^{m}\sum_{k=1}^{m} \left\langle \sum_{j=1}^{n} \langle y_j, u_i\rangle_{\mathbb{R}^m} y_j, u_k \right\rangle_{\mathbb{R}^m} \langle \tilde{u}, u_i\rangle_{\mathbb{R}^m} \langle \tilde{u}, u_k\rangle_{\mathbb{R}^m} \\
&= \sum_{i=1}^{m}\sum_{k=1}^{m} \langle YY^T u_i, u_k\rangle_{\mathbb{R}^m} \langle \tilde{u}, u_i\rangle_{\mathbb{R}^m} \langle \tilde{u}, u_k\rangle_{\mathbb{R}^m} \\
&= \sum_{i=1}^{m}\sum_{k=1}^{m} \langle \lambda_i u_i, u_k\rangle_{\mathbb{R}^m} \langle \tilde{u}, u_i\rangle_{\mathbb{R}^m} \langle \tilde{u}, u_k\rangle_{\mathbb{R}^m} \\
&= \sum_{i=1}^{m}\sum_{k=1}^{m} \lambda_i \delta_{ik} \langle \tilde{u}, u_i\rangle_{\mathbb{R}^m} \langle \tilde{u}, u_k\rangle_{\mathbb{R}^m} = \sum_{i=1}^{m} \lambda_i |\langle \tilde{u}, u_i\rangle_{\mathbb{R}^m}|^2 \\
&\leq \lambda_1 \sum_{i=1}^{m} |\langle \tilde{u}, u_i\rangle_{\mathbb{R}^m}|^2 = \lambda_1 \|\tilde{u}\|_{\mathbb{R}^m}^2 = \lambda_1 = \sum_{j=1}^{n} |\langle y_j, u_1\rangle_{\mathbb{R}^m}|^2
\end{aligned}$$

and therefore u_1 indeed solves (3.2). ■

Analogously we can derive the optimality conditions for the ℓ-dimensional maximization problem (3.1). The result is that the maximizing vectors are the ℓ eigenvectors corresponding to the ℓ largest eigenvalues of the eigenvalue problem $YY^T u_i = \lambda_i u_i$, $i = 1, ..., \ell$. For the proof we refer the reader to [44]. Moreover, the optimal value of the cost function is $\sum_{i=1}^{\ell} \sigma_i^2$. It can be shown that this is indeed the optimal value. The proof is very similar to the proof for the optimal value of the maximization problem (3.2).
Comparing this result with our singular value analysis above we see that the vectors which are solving the problem (3.1) are the left-hand sided singular vectors u_i together with the singular values $\sigma_i = \sqrt{\lambda_i}$ of the snapshot matrix Y.

Furthermore, we use our knowledge of the SVD that the eigenvalues $\lambda_1, ..., \lambda_r$ of the matrices YY^T and Y^TY are the same and the eigenvectors u_i of the matrix YY^T can be calculated from the eigenvectors v_i of the matrix Y^TY

by $u_i = \frac{1}{\sigma_i} Y v_i$, $i = 1, ..., r$.
In the specific problem on which we will concentrate in this work (see Section 4) we are facing a matrix Y whose number of rows is much larger than its number of columns ($\text{rank}(Y) = n$). Therefore, it is easier to compute the eigenvalues of the symmetric matrix $Y^T Y \in \mathbb{R}^{n \times n}$ instead of the larger matrix $YY^T \in \mathbb{R}^{m \times m}$. This property arises from the fact that we view the matrix Y as a matrix of snapshots, this means that the columns of Y stand for only a few (discretized) solutions to a differential equation for different parameters. In our application the discretization of our domain yields more degrees of freedom than the number of snapshots. If that assumption does not hold we regard the symmetric positive definite matrix YY^T which is of a smaller size than $Y^T Y$. More on the topic of discretization follows in Section 3.3.

Let us now look at the POD basis from another point of view. We aim to find orthonormal vectors $u_1, ..., u_\ell \in \mathbb{R}^m$ such that the snapshot vectors y_j are approximated best by the truncated Fourier sum $\sum_{i=1}^{\ell} \langle y_j, u_i \rangle_{\mathbb{R}^m} u_i$. Recall that $\text{rank}(Y) = r$. Thus,

$$y_j = \sum_{i=1}^{r} \langle y_j, u_i \rangle_{\mathbb{R}^m} u_i, \;\; j = 1, ..., n.$$

However, we might not want to use all those vectors as our POD basis because r might still be a relatively large number.
When we take only $\ell (\leq r)$ basis vectors, the equality in the above equation does not hold anymore. Yet we aim to find for any $\ell \in \{1, ..., r\}$ basis vectors $u_1, ..., u_\ell$ such that the norms of the differences between y_j and its ℓ^{th} partial Fourier sum, $j = 1, ..., n$, are as small as possible. This leads to the following minimization problem:

$$\min_{u_1, ..., u_\ell} \sum_{j=1}^{n} \left\| y_j - \sum_{i=1}^{\ell} \langle y_j, u_i \rangle_{\mathbb{R}^m} u_i \right\|_{\mathbb{R}^m}^2 \text{ s.t. } \langle u_i, u_j \rangle_{\mathbb{R}^m} = \delta_{ij}. \tag{3.5}$$

It can be shown that the solution to our maximization problem (3.1), namely the left-hand sided singular vectors $u_1, ..., u_\ell$ corresponding to the ℓ largest singular values of the matrix Y, also solve the minimization problem (3.5). In order to prove this claim we make use of the following corollary.

Corollary 3.5 *Let $Y \in \mathbb{R}^{m \times n}$ be given with $\text{rank}(Y) = r \leq \min\{m, n\}$ and let $\ell \in \{1, .., r\}$. Moreover, let $Y = U \Sigma V^T$ be the singular value decompo-*

sition of Y with $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ orthogonal. Let $B^r \in \mathbb{R}^{r\times n}$ be the matrix with the entries $B^r_{ij} = \langle u_i, y_j\rangle_{\mathbb{R}^m}$. Moreover, suppose that $\hat{U}^r = [\hat{u}_1, ..., \hat{u}_r]$ is an arbitrary orthogonal matrix and $Y = \hat{U}^r C^r$ with $C^r_{ij} = \langle u_i, y_j\rangle_{\mathbb{R}^m}$. Then,

$$\|Y - \hat{U}^\ell C^\ell\|_F^2 \geq \|Y - U^\ell B^\ell\|_F^2 \tag{3.6}$$

where U^ℓ and $\hat{U}^\ell$ denote the matrices consisting of the first ℓ columns of U and $\hat{U}^r$, respectively, and B^ℓ and C^ℓ denote the matrices consisting of the first ℓ lines of B^r and C^r, respectively.

Proof. We observe that

$$\begin{aligned}\|Y - \hat{U}^\ell C^\ell\|_F^2 &= \|\hat{U}^r C^r - \hat{U}^\ell C^\ell\|_F^2 = \|\hat{U}^r (C^r - ((C^\ell)^T, 0)^T)\|_F^2 \\ &= \|(C^r - ((C^\ell)^T, 0)^T)\|_F^2 = \sum_{i=\ell+1}^{r}\sum_{j=1}^{n} |C^r_{ij}|^2.\end{aligned}$$

Analogously, we have

$$\|Y - U^\ell B^\ell\|_F^2 = \sum_{i=\ell+1}^{r}\sum_{j=1}^{n} |B^r_{ij}|^2 = \sum_{i=\ell+1}^{r}\sum_{j=1}^{n} |\langle y_j, u_i\rangle_{\mathbb{R}^m}|^2 = \sum_{i=\ell+1}^{r} \sigma_i^2.$$

From

$$\|Y\|_F^2 = \|\hat{U}^r C^r\|_F^2 = \|C^r\|_F^2 = \sum_{j=1}^{n}\sum_{i=1}^{r} |C^r_{ij}|^2$$

and

$$\|Y\|_F^2 = \|U^r B^r\|_F^2 = \|B^r\|_F^2 = \sum_{j=1}^{n}\sum_{i=1}^{r} |B^r_{ij}|^2 = \sum_{i=1}^{r}\sum_{j=1}^{n} \langle y_j, u_i\rangle_{\mathbb{R}^m}|^2 = \sum_{i=1}^{r} \sigma_i^2$$

we deduce

$$\begin{aligned}\|Y - U^\ell B^\ell\|_F^2 = \sum_{i=\ell+1}^{r} \sigma_i^2 &= \sum_{i=1}^{r} \sigma_i^2 - \sum_{i=1}^{\ell} \sigma_i^2 = \|Y\|_F^2 - \sum_{i=1}^{\ell}\sum_{j=1}^{n} |\langle y_j, u_i\rangle_{\mathbb{R}^m}|^2 \\ &\leq \|Y\|_F^2 - \sum_{i=1}^{\ell}\sum_{j=1}^{n} |\langle y_j, \hat{u}_i\rangle_{\mathbb{R}^m}|^2 = \|Y\|_F^2 - \sum_{i=1}^{\ell}\sum_{j=1}^{n} |C^r_{ij}|^2 \\ &= \sum_{i=1}^{r}\sum_{j=1}^{n} |C^r_{ij}|^2 - \sum_{i=1}^{\ell}\sum_{j=1}^{n} |C^r_{ij}|^2 = \sum_{i=\ell+1}^{r}\sum_{j=1}^{n} |C^r_{ij}|^2 \\ &= \|Y - \hat{U}^\ell C^\ell\|_F^2,\end{aligned}$$

so that the claim follows. ∎

Remark 3.6 *Because of*

$$\begin{aligned}\sum_{j=1}^{n}\left\|y_j-\sum_{i=1}^{\ell}\langle y_j,u_i\rangle_{\mathbb{R}^m}u_i\right\|_{\mathbb{R}^m}^2 &= \sum_{j=1}^{n}\sum_{k=1}^{m}\left|Y_{kj}-\sum_{i=1}^{\ell}\langle y_j,u_i\rangle_{\mathbb{R}^m}U_{ki}^{\ell}\right|^2\\ &= \sum_{j=1}^{n}\sum_{k=1}^{m}\left|Y_{kj}-\sum_{i=1}^{\ell}U_{ki}^{\ell}B_{ij}^{\ell}\right|^2\\ &= \sum_{j=1}^{n}\sum_{k=1}^{m}|Y_{kj}-(U^{\ell}B^{\ell})_{kj}|^2=\|Y-U^{\ell}B^{\ell}\|_F^2\end{aligned}$$

and similarly

$$\sum_{j=1}^{n}\left\|y_j-\sum_{i=1}^{\ell}\langle y_j,\hat{u}_i\rangle_{\mathbb{R}^m}\hat{u}_i\right\|_{\mathbb{R}^m}^2=\|Y-\hat{U}^{\ell}C^{\ell}\|_F^2,$$

the inequality (3.6) *can be written as*

$$\sum_{j=1}^{n}\left\|y_j-\sum_{i=1}^{\ell}\langle y_j,\hat{u}_i\rangle_{\mathbb{R}^m}\hat{u}_i\right\|_{\mathbb{R}^m}^2\geq\sum_{j=1}^{n}\left\|y_j-\sum_{i=1}^{\ell}\langle y_j,u_i\rangle_{\mathbb{R}^m}u_i\right\|_{\mathbb{R}^m}^2.$$

This result implies directly that the vectors $u_1,...,u_\ell$ *are a solution to the minimization problem* (3.5).

3.2 POD in general Hilbert spaces

Next we consider the POD method in a real separable Hilbert space V. In the problem on which we will lay our focus later in this work we use the space $L^2(\Omega)$ or the space $H^1(\Omega)$ for an open and bounded domain Ω as our particular Hilbert space V. Let us introduce the bounded linear operator $\mathcal{C}:L^2(\mathcal{I})\to V$ with $\mathcal{I}=[\mu_l,\mu_u]$ for $-\infty<\mu_l\leq\mu_u<\infty$ by

$$\varphi\mapsto\mathcal{C}\varphi=\int_{\mathcal{I}}\varphi(\mu)y(\mu)\mathrm{d}\mu \text{ for } \varphi\in L^2(\mathcal{I}),$$

where $y(\mu) \in V$ denotes the snapshot (a solution to (2.1)) for each parameter $\mu \in \mathcal{I}$. We define the space $L^2(\mathcal{I}, V)$ as the set of all functions y which satisfy

$$\|y\|_{L^2(\mathcal{I},V)} := \Big(\int_{\mathcal{I}} \|y(\mu)\|_V^2 \mathrm{d}\mu \Big)^{1/2} < \infty.$$

Note that

$$\|\mathcal{C}\varphi\|_V \leq \Big(\int_{\mathcal{I}} |\varphi(\mu)|^2 \mathrm{d}\mu \Big)^{1/2} \Big(\int_{\mathcal{I}} \|y(\mu)\|_V^2 \mathrm{d}\mu \Big)^{1/2} = \|\varphi\|_{L^2(\mathcal{I})} \|y\|_{L^2(\mathcal{I},V)}$$

for $\varphi \in L^2(\mathcal{I})$.

Lemma 3.7 *The adjoint operator $\mathcal{C}^* : V \to L^2(\mathcal{I})$ satisfying*

$$\langle \mathcal{C}\varphi, z \rangle_V = \langle \varphi, \mathcal{C}^* z \rangle_{L^2(\mathcal{I})} \text{ for } \varphi \in L^2(\mathcal{I}) \text{ and } z \in V$$

is given by

$$(\mathcal{C}^* z)(\mu) = \langle z, y(\mu) \rangle_V,$$

where $y(\mu)$ denotes the solution to (2.1) for the parameter $\mu \in \mathcal{I}$, again. Since $\mathcal{C}$ is bounded, $\mathcal{C}^*$ is also bounded.

Proof. We observe

$$\begin{aligned} \langle \varphi, \mathcal{C}^* z \rangle_{L^2(\mathcal{I})} &= \int_{\mathcal{I}} \varphi(\mu)(\mathcal{C}^* z)(\mu) \mathrm{d}\mu = \int_{\mathcal{I}} \varphi(\mu) \langle z, y(\mu) \rangle_V \mathrm{d}\mu \\ &= \int_{\mathcal{I}} \langle z, \varphi(\mu) y(\mu) \rangle_V \mathrm{d}\mu = \Big\langle z, \int_{\mathcal{I}} \varphi(\mu) y(\mu) \mathrm{d}\mu \Big\rangle_V \\ &= \Big\langle \int_{\mathcal{I}} \varphi(\mu) y(\mu) \mathrm{d}\mu, z \Big\rangle_V = \langle \mathcal{C}\varphi, z \rangle_V \end{aligned}$$

for all $\varphi \in L^2(\mathcal{I}), z \in V$. This gives the claim. ∎

We now define the linear operator $\mathcal{R} = \mathcal{C}\mathcal{C}^* : V \to V$ which is of the form

$$\mathcal{R}z = \int_{\mathcal{I}} \langle z, y(\mu) \rangle_V y(\mu) \mathrm{d}\mu \text{ for } z \in V \tag{3.7}$$

and the linear operator $\mathcal{K} = \mathcal{C}^*\mathcal{C} : L^2(\mathcal{I}) \to L^2(\mathcal{I})$ which can be written as

$$(\mathcal{K}\varphi)(\bar{\mu}) = \int_{\mathcal{I}} \langle y(\mu), y(\bar{\mu}) \rangle_V \varphi(\mu) \mathrm{d}\mu \text{ for } \varphi \in L^2(\mathcal{I}).$$

Since $\mathcal{C}$ is bounded, $\mathcal{R}$ and $\mathcal{K}$ are bounded.
From [25] it follows that the mapping $y : \mathcal{I} \to V, \mu \mapsto y(\mu)$ is continuous. Thus,

$$\int_{\mathcal{I}} \int_{\mathcal{I}} |\langle y(\mu), y(\bar{\mu})\rangle_V|^2 \, \mathrm{d}\bar{\mu}\mathrm{d}\mu < \infty.$$

Hence, the operator $\mathcal{K} = \mathcal{C}\mathcal{C}^*$ is compact, and therefore, $\mathcal{R} = \mathcal{C}^*\mathcal{C}$ is compact as well.
From the Hilbert-Schmidt theorem (see, e.g., [39]) it follows that there exists a complete orthonormal basis $\{\psi_i\}_{i\in\mathbb{N}}$ for V and a sequence $\{\lambda_i\}_{i\in\mathbb{N}}$ of non-negative real numbers so that the following eigenvalue equation holds:

$$\mathcal{R}\psi_i = \lambda_i\psi_i, \quad \lambda_1 \geq \lambda_2 \geq \ldots, \quad \text{and } \lambda_i \to 0 \text{ as } i \to \infty.$$

Each non-zero eigenvalue of $\mathcal{R}$ has finite multiplicity and the only possible accumulation point of the spectrum of $\mathcal{R}$ is 0 (see [24]). Let us note that

$$\int_{\mathcal{I}} \|y(\mu)\|_V^2 \, \mathrm{d}\mu = \sum_{i=1}^{\infty} \lambda_i.$$

Similar to Section 3.1, we introduce the cost functional $J : V \to \mathbb{R}$ by

$$J(\psi) = \int_{\mathcal{I}} \|y(\mu) - \langle y(\mu), \psi\rangle_V \ \psi\|_V^2 \mathrm{d}\mu.$$

It can be proved – see [21], for instance – that the first-order optimality condition to the optimization problem

$$\min_{\psi\in V} J(\psi) \text{ s.t. } \|\psi\|_V^2 = 1 \tag{3.8}$$

is equivalent to the infinite-dimensional eigenvalue problem

$$\mathcal{R}\psi = \lambda\psi.$$

Analogously, the first-order conditions for the ℓ-dimensional minimization problem are given by

$$\mathcal{R}\psi_i = \lambda_i\psi_i \text{ for } i = 1, ..., \ell. \tag{3.9}$$

Lemma 3.8 *The eigenvalues of $\mathcal{K}$ are the same as the eigenvalues of $\mathcal{R}$ and the normalized eigenfunctions of $\mathcal{K}$ are given by*

$$v_i(\mu) = \frac{1}{\sqrt{\lambda_i}}(\mathcal{C}^*\psi_i)(\mu) = \frac{1}{\sqrt{\lambda_i}}\langle \psi_i, y(\mu)\rangle_V, \quad 1 \leq i \leq \ell.$$

Proof. We choose $\lambda \in \{\lambda_1, ..., \lambda_\ell\}$ and denote its associated eigenfunction by ψ. Then we have

$$\lambda\psi = \mathcal{R}\psi = \mathcal{C}\mathcal{C}^*\psi.$$

By applying the operator $\mathcal{C}^*$ on the left-hand side and on the right-hand side of the equation and multiplying the resulting equation with $\frac{1}{\sqrt{\lambda}}$ in order to obtain the normalized eigenfunction we obtain

$$\frac{1}{\sqrt{\lambda}}\mathcal{C}^*\lambda\psi = \frac{1}{\sqrt{\lambda}}\mathcal{C}^*\mathcal{C}\mathcal{C}^*\psi,$$

which can be written as

$$\lambda\left(\frac{1}{\sqrt{\lambda}}\mathcal{C}^*\psi\right) = \mathcal{C}^*\mathcal{C}\left(\frac{1}{\sqrt{\lambda}}\mathcal{C}^*\psi\right) = \mathcal{K}\left(\frac{1}{\sqrt{\lambda}}\mathcal{C}^*\psi\right).$$

This gives us the eigenvalue problem for the operator $\mathcal{K}$ with the eigenfunction $v = \frac{1}{\sqrt{\lambda}}\mathcal{C}^*\psi$. The function v is normalized because

$$\begin{aligned}\|v\|^2_{L^2(\mathcal{I})} = \langle v, v\rangle_{L^2(\mathcal{I})} &= \left\langle \frac{1}{\sqrt{\lambda}}\mathcal{C}^*\psi, \frac{1}{\sqrt{\lambda}}\mathcal{C}^*\psi \right\rangle_{L^2(\mathcal{I})} = \frac{1}{\lambda}\langle \mathcal{C}^*\psi, \mathcal{C}^*\psi\rangle_{L^2(\mathcal{I})} \\ &= \frac{1}{\lambda}\langle \mathcal{C}\mathcal{C}^*\psi, \psi\rangle_V = \frac{1}{\lambda}\langle \lambda\psi, \psi\rangle_V = \langle \psi, \psi\rangle_V = 1\end{aligned}$$

and therefore $\|v\|_{L^2(\mathcal{I})} = 1$. ∎

Remark 3.9 (Method of snapshots) *To obtain the eigenfunctions of* (3.9) *we use a technique as presented in [41]. We compute the eigenvalues and eigenfunctions from the problem*

$$\mathcal{K}v_i = \lambda_i v_i \text{ for } i = 1, ..., \ell$$

and calculate the POD basis functions ψ_i afterwards by

$$\psi_i = \frac{1}{\sqrt{\lambda_i}}\mathcal{C}v_i \text{ for } i = 1, ..., \ell. \tag{3.10}$$

Of course, (3.10) *is based on SVD.*

3.3 POD in discretized functional spaces

Suppose that the snapshots on a discrete grid of the interval $\mathcal{I} = [q_l, q_u]$ are given by the FE solution to (2.1) in the form

$$y_j^h(x) = \sum_{l=1}^{n_{FE}} Y_{lj}\varphi_l(x), j = 1, ..., n, \tag{3.11}$$

where $Y \in \mathbb{R}^{n_{FE}\times n}$ is the matrix containing the coefficients in the Galerkin ansatz for each snapshot in each column $Y_{\cdot,j} = \mathrm{y}_j \in \mathbb{R}^{n_{FE}}$, $j = 1, ..., n$. The functions $\{\varphi_l\}_{l=1}^{n_{FE}}$ denote the FE ansatz functions as introduced in Section 2.3.

The POD basis functions (which we are looking for) can be written as linear combinations of the finite elements:

$$\psi_i^h(x) = \sum_{k=1}^{n_{FE}} \mathrm{U}_{ki}\varphi_k(x), i = 1..., \ell. \tag{3.12}$$

Then the continuous minimization problem (3.8) can be approximated by its trapezoidal sum (with α_j representing the trapezoidal weights, $j = 1, ..., n$):

$$\min_{\psi_1^h,...,\psi_\ell^h} \sum_{j=1}^{n} \alpha_j \Big\| y_j^h - \sum_{i=1}^{\ell} \langle y_j^h, \psi_i^h\rangle_{L^2(\Omega)}\psi_i^h \Big\|_{L^2(\Omega)}^2. \tag{3.13}$$

The associated maximization problem (compare Section 3.1) can be written as

$$\max_{\psi_1^h,...,\psi_\ell^h} \sum_{i=1}^{\ell}\sum_{j=1}^{n} \alpha_j |\langle y_j^h, \psi_i^h\rangle_{L^2(\Omega)}|^2 \text{ s.t. } \langle \psi_i^h, \psi_j^h\rangle_{L^2(\Omega)} = \delta_{ij}. \tag{3.14}$$

By M_{ij} we denote the L^2-inner product of the FE-basis functions φ_i and φ_j, respectively. That means

$$M_{ij} = \int_\Omega \varphi_i(x)\varphi_j(x)\,\mathrm{d}x. \tag{3.15}$$

The matrix $M = ((M_{ij})) \in \mathbb{R}^{n_{FE}\times n_{FE}}$ is called mass matrix. If we choose the H^1-inner product instead of the L^2-inner product in (3.14), we will come across the so called stiffness matrix $S = ((S_{ij})) \in \mathbb{R}^{n_{FE}\times n_{FE}}$ with

$$S_{ij} = \int_\Omega \Big(\varphi_i(x)\varphi_j(x) + \nabla\varphi_i(x)\cdot\nabla\varphi_j(x)\Big)\,\mathrm{d}x. \tag{3.16}$$

Thus, we deduce

$$\langle y_j^h, \psi_i^h \rangle_{L^2(\Omega)} = \int_\Omega y_j^h(x)\psi_i^h(x)\,\mathrm{d}x = \sum_{l=1}^{n_{FE}}\sum_{k=1}^{n_{FE}} Y_{lj} \int_\Omega \varphi_l(x)\varphi_k(x)\,\mathrm{d}x\ \mathrm{U}_{ki}$$
$$= \sum_{l=1}^{n_{FE}}\sum_{k=1}^{n_{FE}} Y_{jl}^T W_{lk}\mathrm{U}_{ki} = (Y_{\cdot,j})^T W \mathrm{U}_{\cdot,i}$$

with $W = M$. If we choose the H^1-inner product, we have again

$$\langle y_j^h, \psi_i^h \rangle_{H^1(\Omega)} = (Y_{\cdot,j})^T W \mathrm{U}_{\cdot,i},$$

but now with $W = S$.
As we can see, the L^2-inner product of the two FE-based functions y_j^h and ψ_i^h can be computed by the vector-matrix-vector product $(Y_{\cdot,j})^T M \mathrm{U}_{\cdot,i}$, where $Y_{\cdot,j}$ is the column vector whose entries are the coefficients to the FE-basis functions of the snapshot y_j^h, and $\mathrm{U}_{\cdot,i}$ is the vector of the coefficients to the FE-basis functions of the POD-basis function ψ_i^h, which we aim to find as a solution to the maximization problem (3.14).

Let us first consider the maximization problem in one variable, that means

$$\max_{\psi^h \in \mathrm{span}\{\varphi_1,\dots,\varphi_{n_{FE}}\}} \sum_{j=1}^{n} \alpha_j |\langle y_j^h, \psi^h \rangle_{L^2(\Omega)}|^2 \text{ s.t. } \|\psi^h\|_{L^2(\Omega)} = 1. \tag{3.17}$$

Because the finite elements $\{\varphi_1, ..., \varphi_{n_{FE}}\}$ are fixed, (3.17) can be replaced by

$$\max_{\mathrm{u}_1,\dots,\mathrm{u}_{n_{FE}}} \sum_{j=1}^{n} \alpha_j \left| \left\langle y_j^h, \sum_{i=1}^{n_{FE}} \mathrm{u}_i\varphi_i \right\rangle_{L^2(\Omega)} \right|^2 \text{ s.t. } \left\| \sum_{i=1}^{n_{FE}} \mathrm{u}_i\varphi_i \right\|_{L^2(\Omega)} = 1 \tag{3.18}$$

with $\mathrm{u}_i = \mathrm{U}_{i1}$ in (3.12).
Setting $\mathrm{u} = (\mathrm{u}_1, ..., \mathrm{u}_{n_{FE}})^T$, (3.18) leads to the Lagrange function $\mathcal{L} : \mathbb{R}^{n_{FE}} \times \mathbb{R} \to \mathbb{R}$ given by

$$\mathcal{L}(\mathrm{u}, \lambda) = \sum_{j=1}^{n} \alpha_j \left| \left\langle y_j^h, \sum_{i=1}^{n_{FE}} \mathrm{u}_i\varphi_i \right\rangle_{L^2(\Omega)} \right|^2 + \lambda\left(1 - \left\| \sum_{i=1}^{n_{FE}} \mathrm{u}_i\varphi_i \right\|_{L^2(\Omega)}^2\right).$$

It can be proved by analogous arguments as in Section 3.1 that there exists a Lagrange multiplier λ together with a solution u to the problem (3.17)

satisfying $\nabla\mathcal{L}(\mathrm{u},\lambda)=0$.

In order to obtain an eigenvalue problem similar to (3.3) we calculate the derivatives of $\mathcal{L}$ with respect to u_i, for $i=1,...,n_{FE}$.

$$\begin{aligned}
\nabla_{\mathrm{u}_i}\mathcal{L}(\mathrm{u},\lambda) =& \frac{\partial}{\partial \mathrm{u}_i}\Bigg(\sum_{j=1}^{n}\alpha_j\Big|\sum_{k=1}^{n_{FE}}\sum_{l=1}^{n_{FE}}(Y_{jl})^T W_{lk}\mathrm{u}_k\Big|^2 \\
&+\lambda\Bigg(1-\sum_{k=1}^{n_{FE}}\sum_{l=1}^{n_{FE}}\mathrm{u}_l W_{lk}\mathrm{u}_k\Bigg)\Bigg) \\
=& 2\sum_{j=1}^{n}\alpha_j\Bigg(\sum_{k=1}^{n_{FE}}\sum_{l=1}^{n_{FE}}(Y_{jl})^T W_{lk}\mathrm{u}_k\Bigg)\sum_{s=1}^{n_{FE}}(Y_{js})^T W_{si} \\
&+\lambda\Bigg(-\sum_{k=1}^{n_{FE}}W_{ik}\mathrm{u}_k-\sum_{l=1}^{n_{FE}}\mathrm{u}_l W_{li}\Bigg) \\
=& 2\sum_{k=1}^{n_{FE}}\sum_{l=1}^{n_{FE}}\sum_{s=1}^{n_{FE}}\Bigg(W_{is}\sum_{j=1}^{n}Y_{sj}\alpha_j(Y^T)_{jl}W_{lk}\mathrm{u}_k\Bigg) \\
&+\lambda\Bigg(-\sum_{k=1}^{n_{FE}}W_{ik}\mathrm{u}_k-\sum_{l=1}^{n_{FE}}W_{il}\mathrm{u}_l\Bigg) \\
=& 2(WYDY^TW\mathrm{u})_i-2\lambda(W\mathrm{u})_i
\end{aligned}$$

with $D=\mathrm{diag}(\alpha_1,...\alpha_n)$ and $W=M$. Of course, if we apply the H^1-inner product in (3.17) instead of the L^2-inner product, we have $W=S$.
Hence, the condition $\nabla_{\mathrm{u}}\mathcal{L}(\mathrm{u},\lambda)=0$ leads to the generalized eigenvalue problem

$$2(WYDY^TW\mathrm{u}-\lambda W\mathrm{u})=0, \text{ or equivalently, } WYDY^TW\mathrm{u}=\lambda W\mathrm{u}.$$

We set $\bar{Y}=W^{1/2}YD^{1/2}$ and $\bar{u}=W^{1/2}\mathrm{u}$, where the matrices $W^{1/2}$ and $D^{1/2}$ are defined as in Definition A.7 in the appendix, and obtain

$$WYDY^TW\mathrm{u}=W^{1/2}W^{1/2}YD^{1/2}D^{1/2}Y^TW^{1/2}W^{1/2}\mathrm{u}=W^{1/2}\bar{Y}\bar{Y}^T\bar{u}$$

and

$$\lambda W\mathrm{u}=\lambda W^{1/2}W^{1/2}\mathrm{u}=\lambda W^{1/2}\bar{u}.$$

Consequently, we have

$$W^{1/2}\bar{Y}\bar{Y}^T\bar{u}=\lambda W^{1/2}\bar{u}.$$

By multiplying both sides of the above equation from the left-hand side with $W^{-1/2}$ we obtain the symmetric eigenvalue problem

$$\bar{Y}\bar{Y}^T\bar{u} = \lambda\bar{u}.$$

Similarly as in Section 3.1 we can proceed for the problem in ℓ variables $\psi_1^h, ..., \psi_\ell^h$. This leads us to the following optimality conditions:

$$\bar{Y}\bar{Y}^T\bar{u}_i = \lambda_i\bar{u}_i \text{ for } i = 1, ..., \ell$$

where $\bar{Y} = W^{1/2}YD^{1/2}$ and $\bar{u}_i = W^{1/2}\mathrm{U}_{\cdot,i}$.

Equivalently, we can calculate the eigenvectors from the eigenvalue problem (method of snapshots – compare Remark 3.9)

$$\bar{Y}^T\bar{Y}\bar{v}_i = \lambda_i\bar{v}_i \text{ for } i = 1, ..., \ell \tag{3.19}$$

and set

$$\begin{aligned}\mathrm{U}_{\cdot,i} = W^{-1/2}\bar{u}_i = W^{-1/2}\tfrac{1}{\sqrt{\lambda_i}}\bar{Y}\bar{v}_i = W^{-1/2}\tfrac{1}{\sqrt{\lambda_i}}W^{1/2}YD^{1/2}\bar{v}_i \\ = \tfrac{1}{\sqrt{\lambda_i}}YD^{1/2}\bar{v}_i\end{aligned} \tag{3.20}$$

for $i = 1, ..., \ell$.
Because of the symmetry of W we find that $\bar{Y}^T\bar{Y}$ can be written as

$$\bar{Y}^T\bar{Y} = D^{1/2}Y^TW^{1/2}W^{1/2}YD^{1/2} = D^{1/2}Y^TWYD^{1/2}.$$

Therefore, to solve (3.19) and compute $\mathrm{U}_{\cdot,i}$ by (3.20) we do not require to evaluate $W^{1/2}$.

When we want to compute the POD basis of rank ℓ numerically, we first calculate the snapshots by the FE method. Alternatively, we could compute the snapshots by the method of finite differences or by the method of finite volumes. Throughout this work, the coefficients to the FE basis functions for each snapshot form the vectors y_j which are the columns of the snapshot matrix Y.
After having built the matrix product $\bar{Y}^T\bar{Y}$, we apply an (iterative) eigenvalue solver to that matrix that gives us only the first ℓ eigenvalues $\{\lambda_i\}_{i=1}^{\ell}$ (that means the ℓ largest ones) together with their matching eigenvectors

$\bar{v}_i$. Corresponding to (3.20) we calculate the vectors $\mathrm{U}_{\cdot,i}$, $i = 1, ..., \ell$.
The first ℓ POD basis functions $\psi_1^h, ...\psi_\ell^h$ are then computed by (3.12).
From Section 3.1 we know that the vectors $\mathrm{U}_{\cdot,1}, ..., \mathrm{U}_{\cdot,\ell}$ (i.e., the functions $\psi_1^h, ..., \psi_\ell^h$ with $\psi_i^h = \sum_{k=1}^{n_{FE}} \mathrm{U}_{ki}\varphi_k$ for $i = 1, ..., \ell$) solve the maximization problem (3.14) as they are the eigenvectors corresponding to the largest eigenvalues of the eigenvalue problem $\bar{Y}\bar{Y}^T\mathrm{u} = \lambda\mathrm{u}$ and therefore yield the maximal value for

$$\sum_{i=1}^{\ell}\sum_{j=1}^{n}\left|\alpha_j\left\langle y_j, \sum_{k=1}^{n_{FE}}\mathrm{U}_{ki}\varphi_k\right\rangle_{L^2(\Omega)}\right|^2$$

among all other ℓ orthonormal functions $\{\tilde{\psi}_i^h\}_{i=1}^{\ell}$ represented by a matrix $\tilde{\mathrm{U}} \in \mathbb{R}^{n_{FE}\times\ell}$.

3.4 POD error estimates

Let $q \in \mathcal{I}$, $B(\cdot,\cdot;q)$ be given by (2.5) and F by (2.2). Moreover, $\{\psi_i\}_{i=1}^{\ell}$ are the POD basis functions computed as in Section 3.2. In this section we give error estimates between the solution $u = u(q)$ to

$$B(u,\phi;q) = \langle F,\phi\rangle_{H^1(\Omega)',H^1(\Omega)} \text{ for all } \phi \in H^1(\Omega), \tag{3.21}$$

and the corresponding POD solution $u^\ell = u^\ell(q) \in V^\ell = \mathrm{span}\{\psi_i\}_{i=1}^{\ell}$ to the variational equation

$$B(u^\ell,\phi;q) = \langle F,\phi\rangle_{H^1(\Omega)',H^1(\Omega)} \text{ for all } \phi \in V^\ell. \tag{3.22}$$

First of all, let us introduce the following inner product on $H^1(\Omega)$:

Definition 3.10 *For u, $v \in H^1(\Omega)$ we define the inner product $\langle\cdot,\cdot\rangle_{H^1_\Gamma(\Omega)}$ for $c, \sigma > 0$ by*

$$\langle u, v\rangle_{H^1_\Gamma(\Omega)} = c\int_\Omega \nabla u\cdot\nabla v\,\mathrm{d}x + \sigma\int_\Gamma uv\,\mathrm{d}s$$

and its induced norm $\|\cdot\|_{H^1_\Gamma(\Omega)}$ by

$$\|u\|_{H^1_\Gamma(\Omega)} = \sqrt{\langle u,u\rangle_{H^1_\Gamma(\Omega)}}$$

for all $u \in H^1(\Omega)$.

It follows from [12] that $\|\cdot\|_{H^1_\Gamma(\Omega)}$ is an equivalent norm to $\|\cdot\|_{H^1(\Omega)}$ on $H^1(\Omega)$, thus there exists a constant $c_{H^1} > 0$ such that

$$\|u\|_{H^1(\Omega)} \leq c_{H^1}\|u\|_{H^1_\Gamma(\Omega)} \text{ for all } u \in H^1(\Omega). \tag{3.23}$$

Since $H^1(\Omega)$ is continuously (even compactly) injected in $L^2(\Omega)$, there exists also a constant $c_{L^2} > 0$ such that

$$\|u\|_{L^2(\Omega)} \leq c_{L^2}\|u\|_{H^1(\Omega)} \text{ for all } u \in H^1(\Omega). \tag{3.24}$$

The inequalities (3.23) and (3.24) imply a Poincaré inequality

$$\|u\|_{L^2(\Omega)} \leq c_V\|u\|_{H^1_\Gamma(\Omega)} \text{ for all } u \in H^1(\Omega)$$

with $c_V = c_{L^2}c_{H^1} > 0$.
For simplicity, let us assume that in our elliptic system (2.1) the parameter β is zero, i.e., there is no convection. Then the parametrized bilinear form (2.5) can be written as

$$B(u, \phi; q) = \langle u, \phi\rangle_{H^1_\Gamma(\Omega)} + q\langle u, \phi\rangle_{L^2(\Omega)} \tag{3.25}$$

for $u, \phi \in H^1(\Omega)$ and $q \in \mathcal{I}$. It follows that the mapping $q \mapsto u(q)$ is Lipschitz-continuous (compare Section 3.2). Thus, the operator $\mathcal{R}$ introduced in (3.7) is compact, where $V = H^1(\Omega)$.
Now let us refer to [25], where it is shown that for bilinear forms like the one in (3.25) we are able to formulate error estimates between the solution $u = u(q)$ to (3.21) and the POD solution $u^\ell(q)$ to (3.22) for $q \in \mathcal{I}$.

First we utilize a result from [25] for the case that the solution to (3.21) is given for all $q \in \mathcal{I}$ (the continuous set of snapshots).

Proposition 3.11 *The error between the solution $u = u(q)$ to (3.21) and the POD solution $u^\ell(q)$ to (3.22) for $q \in \mathcal{I}$ can be estimated by*

$$\int_{\mathcal{I}} \|u(q) - u^\ell(q)\|^2_{H^1(\Omega)}\,\mathrm{d}q \leq C_{\mathrm{cont}} \sum_{i=\ell+1}^{\infty} \lambda_i,$$

where $C_{\mathrm{cont}} > 0$ depends on the constants q_l, q_u, $\alpha_2(q_l)$, and c_V, only.

As we can see, the error between the solution $u = u(q)$ to (3.21) and the POD solution $u^\ell(q)$ to (3.22) for $q \in \mathcal{I}$ can be estimated in terms of the eigenvalues corresponding to the not modeled eigenfunctions. If the eigenvalues $\{\lambda_i\}_{i=1}^\infty$ decay rapidly, we deduce that u^ℓ is close to u for a small value of ℓ. This motivates the notion that (3.22) is a low-dimensional POD model for (3.21).

Now we look at the (numerically more interesting) case, where the snapshots are given only on a discrete grid $\{q_i\}_{i=1}^n \subset \mathcal{I}$. This case corresponds to Section 3.3, where the number of columns of the snapshot matrix Y is n. Suppose that for given $0 < q_l < q_u$ we define an equidistant grid on the interval $[q_l, q_u]$ with $n > 1$ grid points to the grid size δq by

$$q_i = q_l + (i-1)\delta q \text{ for } i \in \{1, ..., n\}, \text{ and } \delta q := \frac{q_u - q_l}{n-1}. \tag{3.26}$$

The eigenvalues to the problem (3.19) depend on the grid (3.26) and are now denoted by λ_i^n for $i = 1, ..., d^n$, where d^n is the rank of the matrix $\bar{Y}^T\bar{Y}$ in (3.19). The POD basis functions $\psi_1^h, ..., \psi_\ell^h$ are computed by (3.12), where the vectors $\mathrm{U}_{\cdot,i}$ are calculated for $i = 1, ..., \ell$ as in (3.19) and (3.20) (method of snapshots).

Proposition 3.12 *The error between the solution $u = u(q_i)$ to* (3.21) *and the POD solution $u^\ell(q_i)$ to* (3.22) *for q_i taken from* (3.26) *can be estimated by*

$$\sum_{j=1}^{n} \alpha_j \|u(q_j) - u^\ell(q_j)\|^2_{H^1(\Omega)} \leq C_{\text{disc}} \sum_{i=\ell+1}^{d^n} \lambda_i^n,$$

where $C_{\text{disc}} > 0$ depends on the constants q_l, q_u, $\alpha_2(q_l)$, and c_V, only. In contrast to the case of a continuous snapshot set, the rank d^n of the matrix $\bar{Y}^T\bar{Y}$ and its eigenvalues $\lambda_i^n, i = 1, ..., d^n$ depend on the grid chosen in (3.26) *for the computation of the POD basis.*

Again, the error between the solution $u = u(q_i)$ to (3.21) and the POD solution $u^\ell(q_i)$ to (3.22) for q_i taken from (3.26) can be estimated in terms of the eigenvalues corresponding to the not-modelled eigenfunctions.

Note that in Section 3.5.3 we will find that the eigenvalues λ_i^n of the matrix $\bar{Y}^T\bar{Y}$ converge towards the eigenvalues λ_i for $1 \leq i \leq \ell$ (with ℓ fixed). Hence, both error estimates (the continuous and the discrete one) are closely connected.

Remark 3.13 *In [25] a further result corresponding to Proposition 3.12 is presented, where the error between the POD state and the exact state is estimated as above, except that the grid* $q_l \leq q_1 < ... < q_n \leq q_u$ *for the snapshot set is not equal to the grid* $q_l \leq \bar{q}_1 < ... < \bar{q}_m \leq q_u$ *on which we calculate the error. Defining*

$$\Delta q = \max\left\{ \max_{2\leq j\leq n}(q_j - q_{j-1}), \max_{2\leq j\leq m}(\bar{q}_j - \bar{q}_{j-1}) \right\}$$

we can estimate the error by

$$\sum_{k=1}^{m} \beta_k \|u(\bar{q}_k) - u^\ell(\bar{q}_k)\|^2_{H^1(\Omega)} \leq C\left(C_{\text{grid}} \sum_{i=\ell+1}^{d^n} \lambda_i^n + \Delta q^2 \right),$$

with a constant $C > 0$ *depending on* q_l, q_u, $\alpha_2(q_l)$, *and* c_V, *but independent of the grids, and a constant* $C_{\text{grid}} > 0$ *depending on the grids.*

3.5 Numerical results concerning POD calculation

This section is devoted to present numerical test examples which shall confirm our theoretical results presented in Section 3 so far. All coding is done in MATLAB 6.1 using routines from the FEMLAB 2.2 package concerning finite element implementation. The programs are executed on a standard 1.7 Ghz desktop PC.

3.5.1 POD solution for given differential equation

Our aim is to compare the POD solution to the elliptic equation (2.1) for given parameters to the FE solution to (2.1). The FE solution is computed, for instance, by using the FEMLAB routine `fesolve`.
In order to obtain the POD solution u^ℓ to (2.1) for fixed ℓ, we need to compute the first ℓ POD ansatz functions. Therefore we proceed as described in Section 3.3. For a given domain $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, and a given elliptic differential equation we compute the snapshots by the method of finite elements for a discrete grid of parameters $\{q_i\}_{i=1}^n$ as in (3.26). In our numerical tests we choose the trapezoidal weights

$$\alpha_1 = \alpha_n = \frac{\delta q}{2} \text{ and } \alpha_i = \delta q \text{ for } i = 2, ..., n-1$$

with δq taken as in (3.26). When we apply the method of snapshots as introduced in Remark 3.9, we obtain the eigenvalue problem

$$D^{1/2}Y^TWYD^{1/2}\bar{v}_i = \lambda_i\bar{v}_i. \tag{3.27}$$

We compute the ℓ largest eigenvalues $\lambda_1, ..., \lambda_\ell$ and their corresponding eigenvectors $\bar{v}_1, ..., \bar{v}_\ell$, then compute the vector of coefficients $\mathrm{U}_{\cdot,i}$ of each POD basis function ψ_i, $i = 1, ...\ell$, to the FE basis functions by

$$\mathrm{U}_{\cdot,i} = \frac{1}{\sqrt{\lambda_i}}Y\bar{v}_i. \tag{3.28}$$

The first ℓ POD ansatz functions are then given by

$$\psi_i(x) = \sum_{j=1}^{n_{FE}} \mathrm{U}_{ji}\varphi_j(x) \text{ for } i = 1, ..., \ell. \tag{3.29}$$

The POD solution $u^\ell \in V^\ell = \mathrm{span}\{\psi_i\}_{i=1}^\ell \subset H^1(\Omega)$ given by

$$u^\ell(x) = \sum_{i=1}^{\ell} \mathrm{u}_i^\ell\psi_i(x) \tag{3.30}$$

satisfies the variational equation

$$B(u^\ell, v^\ell; q) = \langle F, v^\ell\rangle_{H^1(\Omega)',H^1(\Omega)} \text{ for all } v^\ell \in V^\ell \text{ and } q \in \mathcal{I}.$$

Especially,

$$B(u^\ell, \psi_j; q) = \langle F, \psi_j\rangle_{H^1(\Omega)',H^1(\Omega)} \tag{3.31}$$

must hold for $j = 1, ..., \ell$ and any $q \in \mathcal{I}$.
Analogously to the method of finite elements (see Section 2.3), we substitute (3.30) into (3.31) and obtain a result similar to (2.9), namely

$$\sum_{i=1}^{\ell} B(\psi_i, \psi_j; q)\mathrm{u}_i^\ell = \langle F, \psi_j\rangle_{H^1(\Omega)',H^1(\Omega)}, 1 \le j \le \ell$$

which leads to the linear equation system

$$(\mathrm{B}^\ell)^T\mathrm{u}^\ell = \mathrm{F}^\ell \tag{3.32}$$

with $\mathrm{B}_{ij}^{\ell} = B(\psi_i, \psi_j; q)$, $\mathrm{B}^{\ell} = ((\mathrm{B}_{ij}^{\ell})) \in \mathbb{R}^{\ell \times \ell}$, $\mathrm{F}_j^{\ell} = \langle F, \psi_j \rangle_{H^1(\Omega)', H^1(\Omega)}$, $\mathrm{F}^{\ell} = (\mathrm{F}_j^{\ell}) \in \mathbb{R}^{\ell}$, and $\mathrm{u}^{\ell} = (\mathrm{u}_j^{\ell}) \in \mathbb{R}^{\ell}$. This linear equation system can be solved uniquely, provided B^{ℓ} is regular, and the POD solution u^{ℓ} is then given by (3.30).

Remark 3.14 *Note that*

$$\begin{aligned}\mathrm{B}_{ik}^{\ell} = B(\psi_i, \psi_k; q) &= B\Big(\sum_{j=1}^{n_{FE}} \mathrm{U}_{ji}\varphi_j, \sum_{l=1}^{n_{FE}} \mathrm{U}_{lk}\varphi_l; q\Big) \\ &= \sum_{j=1}^{n_{FE}}\sum_{l=1}^{n_{FE}} (\mathrm{U}^T)_{ij} B(\varphi_j, \varphi_l; q)\mathrm{U}_{lk} = \sum_{j=1}^{n_{FE}}\sum_{l=1}^{n_{FE}} (\mathrm{U}^T)_{ij}\mathrm{B}_{jl}\mathrm{U}_{lk} = (\mathrm{U}^T\mathrm{B}\mathrm{U})_{ik}\end{aligned}$$

and

$$\begin{aligned}\mathrm{F}_k^{\ell} = \langle F, \psi_k \rangle_{H^1(\Omega)', H^1(\Omega)} &= \langle F, \sum_{j=1}^{n_{FE}} \mathrm{U}_{jk}\varphi_j \rangle_{H^1(\Omega)', H^1(\Omega)} \\ &= \sum_{j=1}^{n_{FE}} \mathrm{U}_{jk}\langle F, \varphi_j \rangle_{H^1(\Omega)', H^1(\Omega)} = \sum_{j=1}^{n_{FE}} (\mathrm{U}^T)_{kj}\mathrm{F}_j = (\mathrm{U}^T\mathrm{F})_k\end{aligned}$$

hold, thus $\mathrm{B}^{\ell} = \mathrm{U}^T\mathrm{B}\mathrm{U}$ *and* $\mathrm{F}^{\ell} = \mathrm{U}^T\mathrm{F}$*, where* $\mathrm{B} = ((\mathrm{B}_{ij})) \in \mathbb{R}^{n_{FE} \times n_{FE}}$ *and* $\mathrm{F} = (\mathrm{F}_j) \in \mathbb{R}^{n_{FE}}$ *are given in* (2.10) *and* (2.11)*.*

Run 3.1 *Let the domain* Ω *be the unit square* $(0,1) \times (0,1) \subset \mathbb{R}^2$*. We choose* $c = 0.75$*,* $\beta = (1,1)^T$*,* $f(x) = x_1$*,* $\sigma = 1.5$*, and* $g(x) = -1$ *in* (2.1)*. To compute the snapshots* $\{\mathrm{y}_j\}_{j=1}^{n}$ *for the POD model we apply a finite element discretization with a rectangular, equidistant mesh with mesh size* $h = 1/50$*. Hence, our FE-discretization yields* 2601 *degrees of freedom. As ansatz functions we utilize piecewise linear finite elements* $\{\varphi_i\}_{i=1}^{n_{FE}}$*, with* $n_{FE} = 2601$*. Let* $q_l = 0.5$*,* $q_u = 50.5$*, and* $\delta q = 1$ *in* (3.26)*, thus we obtain 51 snapshots* $\mathrm{y}_j = Y_{\cdot,j} \in \mathbb{R}^{n_{FE}}$*,* $j = 1, ..., 51$*, where* $Y \in \mathbb{R}^{n_{FE} \times 51}$ *is the coefficient matrix satisfying*

$$u^h(q_j) = \sum_{i=1}^{n_{FE}} Y_{ij}\varphi_i$$

for the parameters $\{q_j\}_{j=1}^{51}$*. Then we determine* $\ell = 7$ *POD basis functions by applying* (3.27) *with the mass matrix* M *as our weight matrix* W*,* (3.28)*, and*

(3.29). *The POD state is then given by* (3.30)*, where the coefficient vector* u^ℓ *is given by the unique solution to* (3.32)*. The computation of the optimal POD state takes 0.67 seconds, plus 39.66 seconds for the computation of the 51 snapshots.*
The POD solutions taking $\ell = 7$ *POD ansatz functions for* $q = 25$ *(left plot) and for* $q = 5$ *(right plot) are presented in Figure 3.1. In Figure 3.2 we have the FE solution for* $q = 25$ *(left plot) and the difference between the POD solution and the FE solution for* $q = 25$ *(right plot). It appears that these two solutions almost coincide. In Figure 3.3 and Figure 3.4 the first four POD ansatz functions which are denoted by* ψ_1*,* ψ_2*,* ψ_3*, and* ψ_4*, are plotted. Now we compute POD solutions for different values for* ℓ*. Our goal is to show that the relative error between the POD solution and the FE solution to a given elliptic equation decreases as the number of POD basis functions increases.*

In Figure 3.5 (right plot) the relative error for the difference of the POD and FE solution is plotted for different values of $q \in [-15.5, 65.5]$ *and for different numbers of POD basis functions* $1 \leq \ell \leq 5$*. Notice that the relative* $L^2(\Omega)$ *error decreases with increasing* ℓ*. Moreover, we observe that the POD basis is also suitable for parameter values (*$q \in (50.5, 65.5]$*, i.e.,* $q \notin \mathcal{I}$*) that are not contained in the snapshot computation in this example. For negative* q *we observe that the quality of the POD approximation becomes worse. In the left plot of Figure 3.5 we see the real and estimated decay (see Section 3.5.2) of the first 5 eigenvalues of* $\bar{Y}^T\bar{Y}$*.*

An alternative in the computation of the POD basis is the subtraction of the mean vector $\mathrm{y}_{mean} = \frac{1}{n}\sum_{j=1}^{n} Y_{\cdot,j}$ for each snapshot $\mathrm{y}_j = Y_{\cdot,j} \in \mathbb{R}^{n_{FE}}$, $j = 1, ..., n$. The resulting snapshot matrix is denoted by $\hat{Y} \in \mathbb{R}^{n_{FE} \times n}$, consisting of the column vectors $\hat{\mathrm{y}}_j$, $j = 1, ..., n$.
The subtraction of the mean vector need not be done in fact, yet we might need more POD basis functions if we renounce the substraction of the mean vector. In applications where we have turbulence or chaotic behavior, it is essential to subtract the mean vector from the snapshots – see [21], for instance. Let us refer to [36] for an application where this procedure appears to be useful. We have tested this procedure in a numerical experiment in the setting of Run 3.1.

Run 3.2 *We proceed as in Run 3.1 except that for the snapshots we choose*

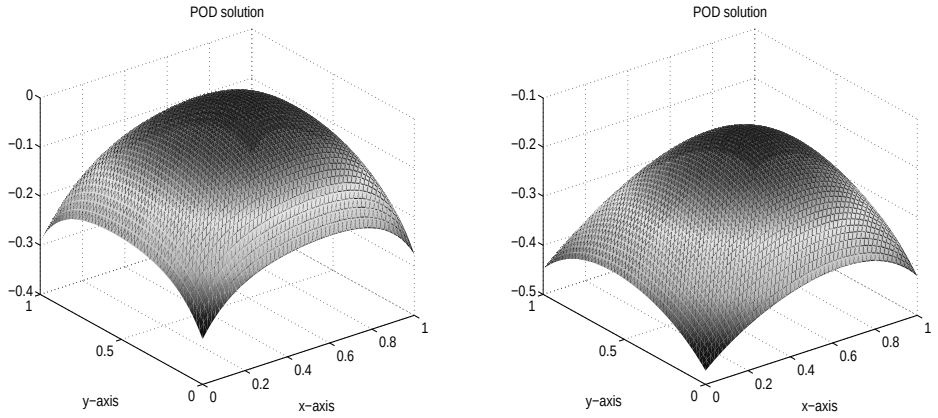

Figure 3.1: Run 3.1: POD solution with $\ell = 7$ ansatz functions for $q = 25$ (left plot) and POD solution for $q = 5$ (right plot).

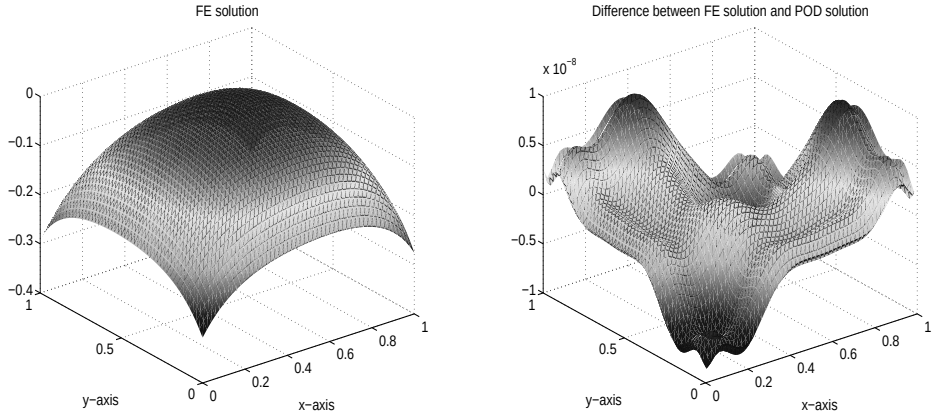

Figure 3.2: Run 3.1: FE solution for $q = 25$ (left plot) and difference between FE and POD solution taking $\ell = 7$ ansatz functions for $q = 25$ (right plot).

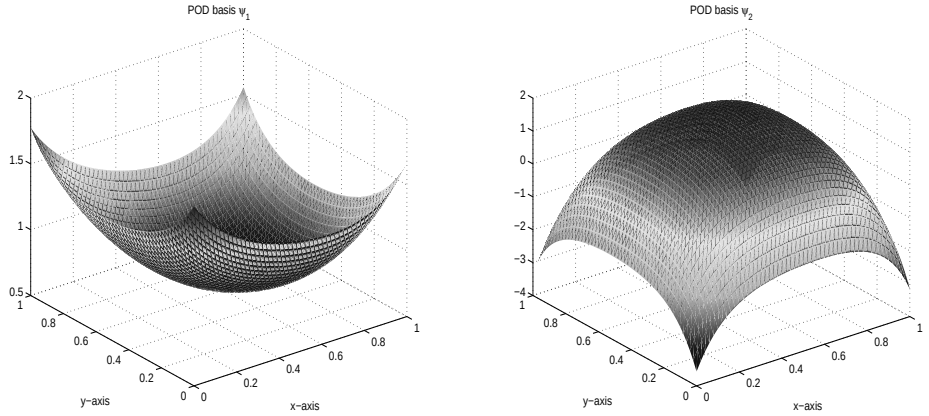

Figure 3.3: Run 3.1: POD basis functions ψ_1 (left plot) and ψ_2 (right plot).

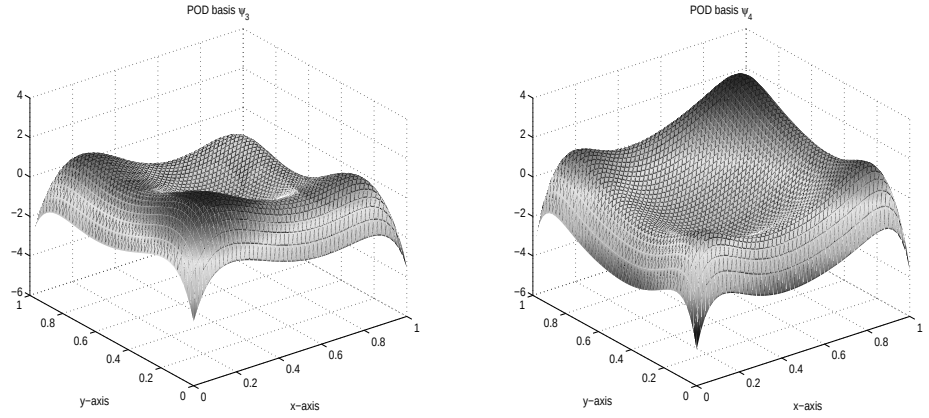

Figure 3.4: Run 3.1: POD basis functions ψ_3 (left plot) and ψ_4 (right plot).

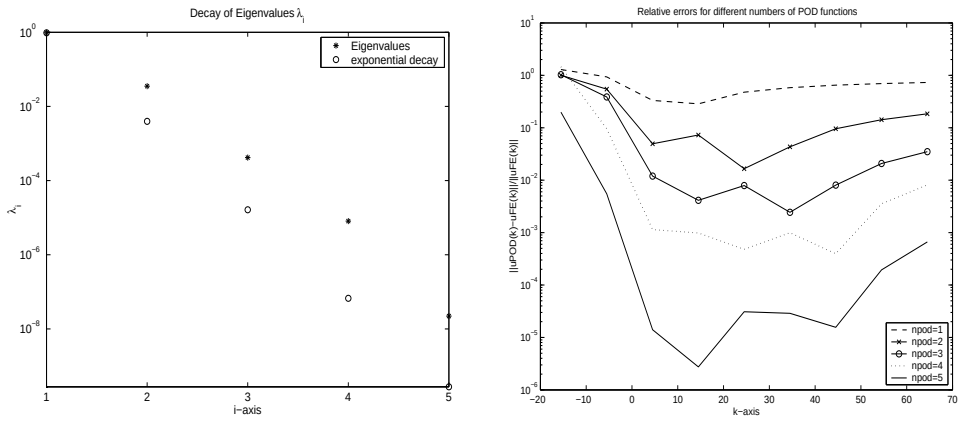

Figure 3.5: Run 3.1: Decay of the first eigenvalues (left plot) and relative errors (L^2-norm) for several values for ell (right plot).

the vectors $\hat{\mathrm{y}}_j$ *instead of* y_j*, which suffice*

$$\hat{\mathrm{y}}_j = \mathrm{y}_j - \mathrm{y}_{mean}.$$

Using the resulting matrix $\hat{Y}$ *instead of* Y *in the computation of the POD basis functions, we have to make the ansatz*

$$u^\ell(x) = \sum_{j=1}^{n_{FE}} (\mathrm{y}_{mean})_j \varphi_j(x) + \sum_{i=1}^{\ell} \mathrm{u}_i^\ell \psi_i^h(x).$$

We define by

$$\mathcal{E}_{L^2} = \frac{\|u^h - u^\ell\|_{L^2(\Omega)}}{\|u^h\|_{L^2(\Omega)}}$$

the relative error in the L^2*-norm, and by*

$$\mathcal{E}_{H^1} = \frac{\|u^h - u^\ell\|_{H^1(\Omega)}}{\|u^h\|_{H^1(\Omega)}}$$

the relative error in the H^1*-norm.*
The relative errors between u^h *and* u^ℓ *for* $\ell = 5$ *and* $\ell = 6$ *are compared in Table 3.2 for the two different strategies – once with subtracting the mean values, and once without subtracting them. As we can see, the relative error is significantly smaller in the first case. The error estimates for the snapshot*

$\ell = 5$:

Snapshot set	$\{\hat{y}_i\}_{i=1}^{51}$	$\{y_i\}_{i=1}^{51}$
$\mathcal{E}_{L^2}$	$1.2 \cdot 10^{-6}$	$1.1 \cdot 10^{-5}$
$\mathcal{E}_{H^1}$	$1.9 \cdot 10^{-6}$	$2.9 \cdot 10^{-5}$

$\ell = 6$:

Snapshot set	$\{\hat{y}_i\}_{i=1}^{51}$	$\{y_i\}_{i=1}^{51}$
$\mathcal{E}_{L^2}$	$4.1 \cdot 10^{-8}$	$1.1 \cdot 10^{-6}$
$\mathcal{E}_{H^1}$	$7.6 \cdot 10^{-8}$	$1.8 \cdot 10^{-6}$

Table 3.1: Run 3.2: Relative errors in the L^2-norm and in the H^1-norm for $\ell = 5$ and $\ell = 6$ for different strategies in the computation of the POD basis.

set $\{\hat{y}_i\}_{i=1}^{5}1$ *and* $\ell = 5$ *are almost as small as for the snapshot set* $\{y_i\}_{i=1}^{5}1$ *and* $\ell = 6$.

Run 3.3 *Let our domain* Ω *be* $B(0;1) \setminus B(0;0.25) \subset \mathbb{R}^2$*, the circle with radius* 1 *minus its concentric circle with radius* 0.25*. The FE-discretization yields* 3256 *degrees of freedom. The snapshots are computed for the same parameters* $\{q_i\}_{i=1}^{51}$ *as in Run 3.1. In the computation of the POD basis functions, the stiffness matrix* S *is used as the weight matrix* W*. For the parameters* $c = 1.5$*,* $\beta = (2,1)^T$*,* $f(x) = -x_1$*,* $\sigma = 1$*, and* $g(x) = x_1$ *in* (2.1) *the POD solution taking* $\ell = 5$ *POD basis functions is shown for* $q = 2$ *in Figure 3.3 (left plot). In the right plot of Figure 3.3 we see the difference between the POD solution and the corresponding FE solution for* $q = 2$*. Again the difference is very small. We obtain the relative errors*

$$\frac{\|u^h - u^\ell\|_{L^2(\Omega)}}{\|u^h\|_{L^2(\Omega)}} \approx 9.1 \cdot 10^{-5} \text{ and } \frac{\|u^h - u^\ell\|_{H^1(\Omega)}}{\|u^h\|_{H^1(\Omega)}} \approx 3.9 \cdot 10^{-4}.$$

on the overall domain Ω*.*

Now we apply our algorithm on a 3-dimensional problem.

Run 3.4 *Let our domain* Ω *be* $(0,1) \times (0,1) \times (0,1) \subset \mathbb{R}^3$*, the cube with side length* 1*. We apply a grid with* 2424 *degrees of freedom. The FE-discretization of the domain is shown in Figure 3.4. The parameters in* (2.1) *are* $c = 2$*,* $\beta = (1,-1,2)^T$*,* $f(x) = 1$*,* $\sigma = -0.5$*, and* $g(x) = 1$*. The snapshots are computed for the same parameters* $\{q_i\}_{i=1}^{51}$ *as in Run 3.1. The mass matrix* M *is used as the weight matrix* W *in the computation of the*

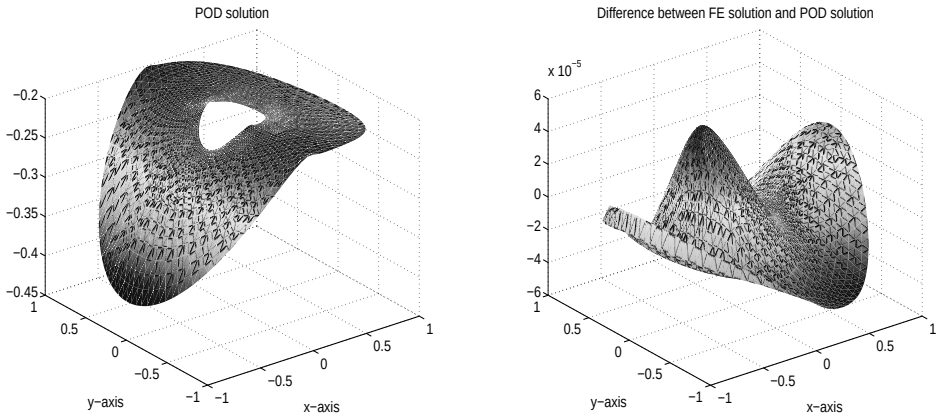

Figure 3.6: Run 3.3: POD solution (left plot) and difference between FE and POD solution taking $\ell = 5$ POD ansatz functions (right plot).

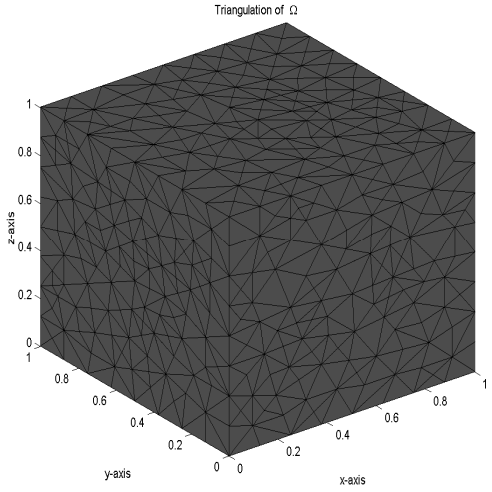

Figure 3.7: Run 3.4: FE-discretization of the 3-dimensional domain Ω.

POD basis functions. The relative errors between the POD solution taking $\ell = 6$ *POD ansatz functions and the respective FE solution are*

$$\frac{\|u^h - u^\ell\|_{L^2(\Omega)}}{\|u^h\|_{L^2(\Omega)}} \approx 9.5 \cdot 10^{-9} \text{ and } \frac{\|u^h - u^\ell\|_{H^1(\Omega)}}{\|u^h\|_{H^1(\Omega)}} \approx 3.4 \cdot 10^{-8}.$$

3.5.2 Experimental Order of Decay

In order to estimate the decay of the eigenvalues of the weighted snapshot matrix $\bar{Y}^T\bar{Y} = D^{1/2}Y^TWYD^{1/2}$ let us introduce a technique known as the *Experimental Order of Decay*. Let us refer to [20] on error estimates for general linear-quadratic optimal control problems and [25] where this strategy is presented in the context of an elliptic problem. Let us assume that the first L eigenvalues decay approximately in an exponential manner. Thus, we write the eigenvalues in the following way:

$$\lambda_i = \lambda_1 e^{-\alpha(i-1)} \text{ , } i = 1, ..., L, \tag{3.33}$$

where α is a parameter which is still to determine. This value can be determined numerically. We will show some numerical examples regarding our elliptic system later in this section. We choose a set of parameters $\mathcal{I} = [q_l, q_u]$, where q_l and q_u are the lower and the upper bound of the set, respectively. We define a grid in $\mathcal{I}$ as in (3.26). From Proposition 3.11 we infer that the integral over the errors between the POD solution u^ℓ and the exact solution u can be approximated by the sum of the eigenvalues corresponding to the eigenfunctions which have not been used for the POD model, that means

$$T(\ell) := \int_{\mathcal{I}} \|u(q) - u^\ell(q)\|^2_{H^1(\Omega)} \, \mathrm{d}q \sim \sum_{i=\ell+1}^{\infty} \lambda_i. \tag{3.34}$$

We approximate this integral by its trapezoidal approximation:

$$\begin{aligned} T^{\delta q}(\ell) :=& \frac{\delta q}{2}\Big(\|u^h(q_l) - u^\ell(q_l)\|^2_{H^1(\Omega)} + \|u^h(q_u) - u^\ell(q_u)\|^2_{H^1(\Omega)}\Big) \\ &+ \delta q \sum_{i=2}^{n-1} \|u^h(q_i) - u^\ell(q_i)\|^2_{H^1(\Omega)}. \end{aligned}$$

In this formula $u^h(q)$ denotes the solution to the differential equation (2.8) with parameter q by using the FE method. Instead of the norm $\|\cdot\|_{H^1(\Omega)}$

we could also use $\|\cdot\|_{L^2(\Omega)}$ in the above equation and in (3.34). Now, when we look at the fraction

$$\frac{T^{\delta q}(\ell)}{T^{\delta q}(\ell+1)}$$

and we use our knowledge that $T^{\delta q}(\ell)$ approximates $T(\ell)$ for all ℓ, we deduce

$$\begin{aligned}
\frac{T^{\delta q}(\ell)}{T^{\delta q}(\ell+1)} \sim \frac{T(\ell)}{T(\ell+1)} \sim \frac{\sum_{i=\ell+1}^{\infty}\lambda_i}{\sum_{i=\ell+2}^{\infty}\lambda_i} &= \frac{\sum_{i=\ell+1}^{\infty}\lambda_1\cdot e^{-\alpha(i-1)}}{\sum_{i=\ell+2}^{\infty}\lambda_1\cdot e^{-\alpha(i-1)}} \\
&= \frac{\sum_{i=\ell+1}^{\infty}e^{-\alpha(i-1)}}{\sum_{i=\ell+2}^{\infty}e^{-\alpha(i-1)}} = \frac{e^{-\alpha\ell}+e^{-\alpha(\ell+1)}+...}{e^{-\alpha(\ell+1)}+e^{-\alpha(\ell+2)}+...} \\
&= \frac{e^{\alpha\ell}(e^{-\alpha\ell}+e^{-\alpha(\ell+1)}+...)}{e^{\alpha\ell}(e^{-\alpha(\ell+1)}+e^{-\alpha(\ell+2)}+...)} = \frac{1+e^{-\alpha}+(e^{-\alpha})^2+...}{e^{-\alpha}+(e^{-\alpha})^2+...} \\
&= \frac{\sum_{i=0}^{\infty}(e^{-\alpha})^i}{\sum_{i=1}^{\infty}(e^{-\alpha})^i} = \frac{\sum_{i=0}^{\infty}(e^{-\alpha})^i}{\sum_{i=0}^{\infty}(e^{-\alpha})^i-1} = \frac{\frac{1}{1-e^{-\alpha}}}{\frac{1}{1-e^{-\alpha}}-1} \\
&= \frac{\frac{1}{1-e^{-\alpha}}}{\frac{1}{1-e^{-\alpha}}-\frac{1-e^{-\alpha}}{1-e^{-\alpha}}} = \frac{1}{e^{-\alpha}} = e^{\alpha}.
\end{aligned}$$

By taking the logarithm we obtain

$$Q(\ell) := \ln\frac{T^{\delta q}(\ell)}{T^{\delta q}(\ell+1)} \sim \alpha.$$

If we compute $Q(\ell)$ for $\ell = 1,...,L$ and average over all these values, we receive an estimation for α (the exponent by which the eigenvalues exponentially decay) according to our assumption:

$$\alpha_{EOD} = \frac{1}{L}\sum_{\ell=1}^{L}Q(\ell).$$

In the following numerical experiment we want to compare the real decay of the first eigenvalues with the decay of the 'approximation of the first eigenvalues' which we obtain by applying (3.33).

Run 3.5 *We choose the parameters $c = 2$, $\beta = (0,1)^T$, $f(x) = x_1$, $\sigma = -1$, and $g(x) = 2$ in (2.1). As the weight matrix W in the computation of the POD ansatz functions we choose the stiffness matrix S. In Figure 3.8 (left plot) we show the real and estimated decay of the first 7 eigenvalues, where*

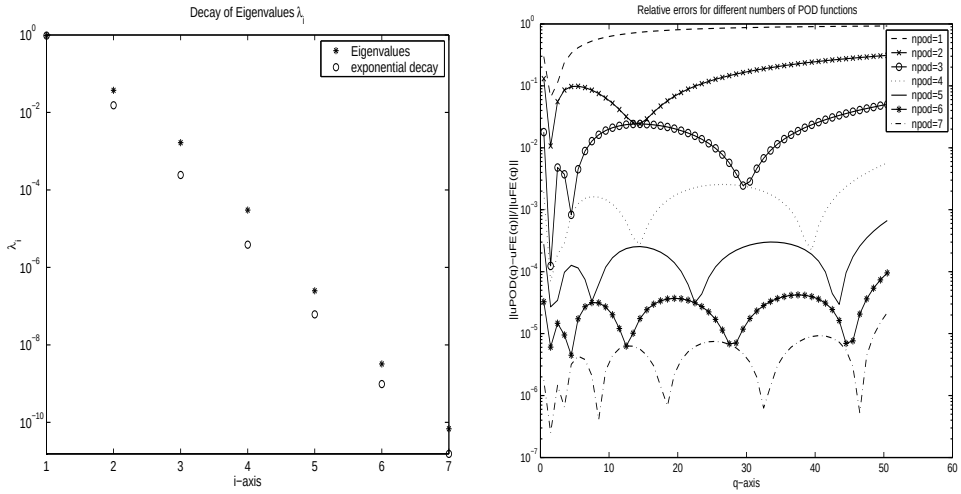

Figure 3.8: Run 3.5: Real and estimated decay of the first 7 eigenvalues (left plot) and relative error in the L^2-norm (right plot).

the estimated decay is computed as described above, using the H^1*-norm in* (3.34)*. Notice that we have normalized all eigenvalues so that* $\sum_{i=1}^{d^n} \lambda_i^n = 1$ *holds. Apparently the experimental order of decay leads to a good estimate for the decay of the first 7 eigenvalues. In the right plot of Figure 3.8 we observe that the relative error for the difference between the POD and the respective FE solution is already smaller than* 10^{-3} *for* $\ell = 5$*.*

3.5.3 Observed convergence of eigenvalues as the number of snapshots increases

Let us remind the reader of the operator $\mathcal{R} : V \to V$ (compare (3.7)) which is defined as

$$\mathcal{R}\psi = \int_{\mathcal{I}} \langle y(\mu), \psi \rangle_V \, y(\mu) \mathrm{d}\mu \text{ for } \psi \in V.$$

Now we define the discrete operator $\mathcal{R}^n : V^h \to V^h$ by

$$\mathcal{R}^n \psi^h = \sum_{j=1}^{n} \alpha_j \langle y^h(\mu_i), \psi^h \rangle_V y^h(\mu_i) \text{ for } \psi^h \in V^h,$$

where the functions $y^h(\mu_i) \in V^h$, $i = 1, ..., n$, are the snapshots (in FE form) on a grid on the interval $\mathcal{I}$ and α_i, $i = 1, ..., n$, are the trapezoidal weights. The operator $\mathcal{R}^n$ is the trapezoidal approximation for $\mathcal{R}$, provided the mapping $\mu \mapsto y(\mu) \in V$ is continuous. Thus,

$$\lim_{n\to\infty} \|\mathcal{R}^n - \mathcal{R}\|_{L(V)} := \lim_{n\to\infty} \sup_{\|z\|_V=1} \|(\mathcal{R}^n - \mathcal{R})z\|_V = 0.$$

We have already shown in Section 3.2 and Section 3.3 that $\mathcal{R}\psi_i = \lambda_i\psi_i$ and $\mathcal{R}^n\psi_i^n = \lambda_i^n\psi_i^n$ are the first-order optimality conditions to the continuous (3.8) and the discretized optimization problem (3.13), respectively.
It is proved in [24] that

$$\sum_{i=1}^{\infty} \lambda_i^n \to \sum_{i=1}^{\infty} \lambda_i \text{ as } n \to \infty. \tag{3.35}$$

Now we choose and fix ℓ such that $\lambda_\ell \neq \lambda_{\ell+1}$. Then, by the arguments stated in [24], it follows that

$$\lambda_i^n \to \lambda_i \text{ for } 1 \leq i \leq \ell \text{ as } n \to \infty.$$

This hypothesis is tested in a numerical experiment. Of course, we cannot determine the eigenvalues of the continuous snapshot set numerically, yet we can calculate the eigenvalues for sufficiently fine grids in the interval $\mathcal{I} = [q_l, q_u]$. Our goal is to confirm the theoretical result in (3.35) by computing the eigenvalues for several discrete grids – from a very coarse one to a very fine one – and comparing the respective eigenvalues.

Run 3.6 *In the setting of Run 3.4 let the weight matrix W be the stiffness matrix S, that means that we consider the maximization problem*

$$\max_{\psi_1^h,\ldots,\psi_\ell^h} \sum_{i=1}^{\ell} \sum_{j=1}^{n} \alpha_j |\langle y_j^h, \psi_i^h \rangle_{H^1(\Omega)}|^2 \text{ s.t. } \langle \psi_i^h, \psi_j^h \rangle_{H^1(\Omega)} = \delta_{ij}.$$

Moreover, we set $q_l = 0.5$ and $q_u = 24.5$ and fix the number of POD basis functions by $\ell = 6$. For different grid sizes δq we obtain the normalized eigenvalues (that means, divided by $\sum_{i=1}^{d^n} \lambda_i^n$) of the matrix $\bar{Y}^T\bar{Y}$ (which corresponds to the discretized operator $\mathcal{R}^n$) stated in Table 3.2. Apparently the eigenvalues indeed converge as the grid size δq decreases, i.e., as the number of snapshots n increases.

	$n = 7$	$n = 13$	$n = 25$	$n = 49$	$n = 97$	$n = 193$
λ_1^n	0.8997	0.9099	0.9127	0.9127	0.9123	0.9121
λ_2^n	0.0997	0.0896	0.0869	0.0869	0.0872	0.0874
λ_3^n	$5.8 \cdot 10^{-4}$	$5.1 \cdot 10^{-4}$	$4.6 \cdot 10^{-4}$	$4.5 \cdot 10^{-4}$	$4.5 \cdot 10^{-4}$	$4.6 \cdot 10^{-4}$
λ_4^n	$1.1 \cdot 10^{-6}$	$1.3 \cdot 10^{-6}$	$1.1 \cdot 10^{-6}$	$1.0 \cdot 10^{-6}$	$1.0 \cdot 10^{-6}$	$1.1 \cdot 10^{-6}$
λ_5^n	$1.8 \cdot 10^{-10}$	$2.5 \cdot 10^{-10}$	$2.2 \cdot 10^{-10}$	$2.1 \cdot 10^{-10}$	$2.0 \cdot 10^{-10}$	$2.1 \cdot 10^{-10}$
λ_6^n	$5.8 \cdot 10^{-14}$	$1.4 \cdot 10^{-13}$	$1.4 \cdot 10^{-13}$	$1.2 \cdot 10^{-13}$	$1.2 \cdot 10^{-13}$	$1.2 \cdot 10^{-13}$

Table 3.2: Run 3.6: Eigenvalues $\{\lambda_i^n\}_{i=1}^6$ for different numbers of snapshots n (i.e., different grid sizes δq) in the set $\mathcal{I} = [0.5, 24.5]$.

4 Parameter estimation

In this section we apply a POD Galerkin discretization to a parameter estimation problem for an elliptic differential equation. Our goal is to identify a scalar parameter in the elliptic equation from boundary measurements of the function u.
Recall the elliptic differential equation which we introduced in Section 2.2:

$$-c\Delta u + \beta \cdot \nabla u + qu = f \quad \text{in } \Omega, \tag{4.1a}$$

$$c\frac{\partial u}{\partial n} + \sigma u = g \quad \text{on } \Gamma = \partial\Omega. \tag{4.1b}$$

From now on we presume that we are considering the differential equation (4.1) in which all parameters except for the scalar q are given, and that we have measurements for u on the boundary of the domain Ω. We will observe in our numerical tests (see Section 4.5) that we obtain good results for the estimated parameter when we use the data given on the boundary $\Gamma = \partial\Omega$. Furthermore, we must take into consideration that the given data for the solution of the differential equation might be corrupted due to measurement noise. In Section 4.5.2 we will see that we find satisfactory results for little perturbed data, too.

4.1 Formulation of the parameter estimation problem as an optimal control problem

Writing (4.1) in abstract form we introduce the operator $e : \mathbb{R} \times H^1(\Omega) \rightarrow$

$H^1(\Omega)'$ by

$$\langle e(q,u),\varphi\rangle_{H^1(\Omega)',H^1(\Omega)} = \int_\Omega c\nabla u\cdot\nabla\varphi\,\mathrm{d}x + \int_\Omega \beta\cdot\nabla u\varphi\,\mathrm{d}x + \int_\Omega qu\varphi\,\mathrm{d}x - \int_\Omega f\varphi\,\mathrm{d}x + \int_\Gamma \sigma u\varphi\,\mathrm{d}s - \int_\Gamma g\varphi\,\mathrm{d}s$$

for $(q,u)\in\mathbb{R}\times H^1(\Omega)$, $\varphi\in H^1(\Omega)$. Notice that

$$e(q,u) = B(u,\cdot;q) - F \in H^1(\Omega)'$$

with B as introduced in (2.5) and F from (2.2).
Moreover, let us define the quadratic cost functional $J:\mathbb{R}\times H^1(\Omega)\to\mathbb{R}$ by

$$J(q,u) = \frac{\alpha}{2}\int_\Gamma |u-u_\Gamma|^2\mathrm{d}s + \frac{\kappa}{2}|q-q_d|^2. \tag{4.2}$$

The values $\alpha\geq 0$ and $\kappa>0$ can be viewed as weights or regulization parameters that indicate how much weight we give to the respective quadratic terms in the optimization problem.
The scalar q_d is a possible a-priori estimate for the sought-after parameter q. If we do not have an estimate at all, q_d is set equal to zero.
The function $u_\Gamma\in L^2(\Gamma)$ represents the given data for u on the boundary Γ. These measurements, of course, can only be taken on a grid of Γ, in practice.

Remark 4.1 *Of course it is also possible that one has measurements for the function u in the whole domain Ω (not only on the boundary Γ) or even measurements for the gradient of u in the domain Ω. More generally, our cost function can be expanded to*

$$\hat{J}(q,u) = \frac{\alpha_1}{2}\int_\Omega |u-u_\Omega|^2\mathrm{d}x + \frac{\alpha_2}{2}\int_\Omega |\nabla u - v_\Omega|^2\mathrm{d}x + \frac{\alpha_3}{2}\int_\Gamma |u-u_\Gamma|^2\mathrm{d}s + \frac{\kappa}{2}|q-q_d|^2$$

with $u_\Omega\in L^2(\Omega)$, $v_\Omega\in L^2(\Omega)^d$, $\alpha_1\geq 0$, $\alpha_2\geq 0$, $\alpha_3\geq 0$, and $\kappa>0$. Here, we focus on the case $\alpha_1=\alpha_2=0$ and $\alpha_3=:\alpha>0$.

Using the notation above we obtain the infinite-dimensional optimization problem

$$\textbf{(P)}\qquad \min J(q,u) \text{ s.t. } e(q,u)=0 \text{ in } H^1(\Omega)' \text{ and } q_a\leq q.$$

The value q_a stands for a lower bound of the parameter q, which we aim to find as a solution to our minimization problem. The inequality constraint in (**P**) is optional, yet in many practical problems it is necessary to demand that $q \geq 0$ holds. Moreover, an upper bound to q can also be added to the optimization problem. This additional inequality constraint does not make a serious difference in the optimization analysis, though, so we neglect it in our study.

Theorem 4.2 *If*

$$\alpha_2(q_a) = \min\left\{\frac{c}{2}, q_a - \frac{\|\beta\|^2_{\mathbb{R}^d}}{2c}\right\} + \min\{0, \sigma C_\Gamma^2\} > 0,$$

(**P**) *admits a local solution* $x^* = (q^*, u^*)$.

Proof. First we need to check wether the set of feasible points

$$\mathcal{F}(\mathbf{P}) = \{(q,u) : e(q,u) = 0 \text{ and } q_a \leq q\}$$

is non-empty. Assume that the conditions from Proposition 2.10 are fulfilled for all $q \in Q_{ad} = \{q \in \mathbb{R} : q \geq q_a\}$. Then we choose $q = q_a$ and from Proposition 2.10 we know that there exists a unique solution $u = u(q)$ associated with q, satisfying $e(q,u) = 0$.
It follows that there exists a minimizing sequence $\{q^n, u^n\}_{n\in\mathbb{N}}$ in $\mathcal{F}(\mathbf{P})$ and

$$0 \leq \inf_{(q,u)\in\mathcal{F}(\mathbf{P})} J(q,u) = \lim_{n\to\infty} J(q^n, u^n) < \infty. \tag{4.3}$$

Now we show that this minimizing sequence is bounded. This is done by a proof by contradiction.
Let us assume that $\lim_{n\to\infty} q^n = \infty$. Then J is radially unbounded in q and $\lim_{n\to\infty} J(q^n, u^n) = \infty$. Since $\kappa > 0$, this is in the contradiction to (4.3). Therefore we deduce that there exists a constant C_q independent from n such that $|q^n| \leq C_q$ for all $n \in \mathbb{N}$. We use the theorem of Bolzano-Weierstrass and conclude that there exists a subsequence $\{q^{n_k}\}_{k\in\mathbb{N}}$ with $q^{n_k} \to q^*$ as $k \to \infty$ for a $q^* \in \mathbb{R}$. Since Q_{ad} is closed, $q^* \geq q_a$ holds, too.
Next we assume that $\lim_{n\to\infty} \|u^n\|_{H^1(\Omega)} = \infty$. Because $(q^n, u^n) \in \mathcal{F}(\mathbf{P})$ for every $n \in \mathbb{N}$, the equality constraint $e(q^n, u^n) = 0$ holds, i.e.,

$$-c\Delta u^n + \beta \cdot \nabla u^n + q^n u^n = f \quad \text{in } \Omega, \tag{4.4a}$$

$$c\frac{\partial u^n}{\partial n} + \sigma u^n = g \quad \text{on } \Gamma = \partial\Omega. \tag{4.4b}$$

By multiplying (4.4a) with u^n and building the integral over Ω we infer

$$\begin{aligned}\int_\Omega c\|\nabla u^n\|_{\mathbb{R}^d}^2 \,\mathrm{d}x + \int_\Omega \beta\cdot\nabla u^n\, u^n \,\mathrm{d}x + \int_\Omega q^n|u^n|^2 \,\mathrm{d}x + \int_\Gamma \sigma|u^n|^2 \,\mathrm{d}s \\ = \int_\Omega f\, u^n \,\mathrm{d}x + \int_\Gamma g\, u^n \,\mathrm{d}s.\end{aligned} \tag{4.5}$$

Let us remind of the parametrized bilinear form $B(\cdot,\cdot;q) : H^1(\Omega)\times H^1(\Omega) \to \mathbb{R}$ for fixed $q \geq q_a$ which we introduced in Section 2.2:

$$B(u,\phi;q) = c\int_\Omega \nabla u\cdot\nabla\phi \mathrm{d}x + \sigma\int_\Gamma u\phi \mathrm{d}s + q\int_\Omega u\phi \mathrm{d}x + \int_\Omega \beta\cdot\nabla u\phi \mathrm{d}x$$

for $u,\phi \in H^1(\Omega), q \in \mathcal{I}$. Obviously equation (4.5) can be written as

$$B(u^n,u^n;q^n) = \int_\Omega f\, u^n \,\mathrm{d}x + \int_\Gamma g\, u^n \,\mathrm{d}s.$$

Using Remark 2.11 we infer that

$$\int_\Omega f\, u^n \,\mathrm{d}x + \int_\Gamma g\, u^n \,\mathrm{d}s = B(u^n,u^n;q^n) \geq \alpha_2(q^n)\|u^n\|_{H^1(\Omega)}^2$$

holds, where

$$\alpha_2(q^n) = \min\left\{\frac{c}{2}, q^n - \frac{\|\beta\|_{\mathbb{R}^d}^2}{2c}\right\} + \min\{0,\sigma C_\Gamma^2\}.$$

By applying Young's inequality (see Lemma A.2) with $\epsilon = \frac{1}{\alpha_2(q^n)}$ twice, and the trace theorem (see Lemma A.14) with C_Γ denoting the trace constant, we deduce

$$\begin{aligned}\alpha_2(q^n)\|u^n\|_{H^1(\Omega)}^2 \leq& \|f\|_{L^2(\Omega)}\|u^n\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma)}\|u^n\|_{L^2(\Gamma)} \\ \leq& \frac{1}{\alpha_2(q^n)}\|f\|_{L^2(\Omega)}^2 + \frac{\alpha_2(q^n)}{4}\|u^n\|_{L^2(\Omega)}^2 + C_\Gamma\|g\|_{L^2(\Gamma)}\|u^n\|_{H^1(\Omega)} \\ \leq& \frac{1}{\alpha_2(q^n)}\|f\|_{L^2(\Omega)}^2 + \frac{\alpha_2(q^n)}{4}\|u^n\|_{H^1(\Omega)}^2 + \frac{C_\Gamma^2}{\alpha_2(q^n)}\|g\|_{L^2(\Gamma)}^2 \\ &+ \frac{\alpha_2(q^n)}{4}\|u^n\|_{H^1(\Omega)}^2.\end{aligned}$$

Thus, we observe

$$\frac{\alpha_2(q^n)}{2}\|u^n\|_{H^1(\Omega)}^2 \leq \frac{1}{\alpha_2(q^n)}\|f\|_{L^2(\Omega)}^2 + \frac{C_\Gamma^2}{\alpha_2(q^n)}\|g\|_{L^2(\Gamma)}^2$$

and therefore

$$\begin{aligned}\|u^n\|^2_{H^1(\Omega)} &\leq \frac{2}{\alpha_2(q^n)}\left(\frac{1}{\alpha_2(q^n)}\|f\|^2_{L^2(\Omega)} + \frac{C_\Gamma^2}{\alpha_2(q^n)}\|g\|^2_{L^2(\Gamma)}\right)\\ &= \frac{2}{\alpha_2(q^n)^2}\left(\|f\|^2_{L^2(\Omega)} + C_\Gamma^2\|g\|^2_{L^2(\Gamma)}\right) \leq C\end{aligned}$$

for a constant $C > 0$, provided $\alpha_2(q^n) > 0$. Since $q^n \geq q_a$ for all $n \in \mathbb{N}$, it follows by assumption that $\alpha_2(q^n) \geq \alpha_2(q_a) > 0$ holds for all $n \in \mathbb{N}$. Then, $\{u^n\}_{n\in\mathbb{N}}$ is uniformly bounded in the Hilbert space $H^1(\Omega)$, and thus a subsequence converges weakly towards an element $u^* \in H^1(\Omega)$ (we write $u^{n_k} \rightharpoonup u^*$), that means there exists a subsequence $\{u^{n_k}\}_{k\in\mathbb{N}}$ with

$$\lim_{k\to\infty} \langle u^{n_k} - u^*, \varphi\rangle_{H^1(\Omega)} = 0 \text{ for all } \varphi \in H^1(\Omega).$$

In particular,

$$\lim_{k\to\infty} \int_\Omega (u^{n_k} - u^*)\varphi \,\mathrm{d}x = 0 \text{ for all } \varphi \in L^2(\Omega) \tag{4.6}$$

and

$$\lim_{k\to\infty} \int_\Omega \nabla(u^{n_k} - u^*) \cdot \nabla\varphi \,\mathrm{d}x = 0 \text{ for all } \varphi \in H^1(\Omega) \tag{4.7}$$

hold. Equation (4.7) is equivalent to

$$\int_\Omega \nabla(u^{n_k} - u^*) \cdot \psi \,\mathrm{d}x \to 0 \text{ for all } \psi \in L^2(\Omega)^d. \tag{4.8}$$

Because the trace operator $\tau_\Gamma : H^1(\Omega) \to L^2(\Gamma)$ is linear and bounded, it follows that

$$\begin{aligned}\|\tau_\Gamma u^{n_k}\|_{L^2(\Gamma)} &\leq \sup_{\|\varphi\|_{H^1(\Omega)}=1} \|\tau_\Gamma\varphi\|_{L^2(\Gamma)}\|u^{n_k}\|_{H^1(\Omega)}\\ &\leq \max_{\|\varphi\|_{H^1(\Omega)}=1} C_\Gamma\|\varphi\|_{H^1(\Omega)}\|u^{n_k}\|_{H^1(\Omega)} = C_\Gamma\|u^{n_k}\|_{H^1(\Omega)} \leq C.\end{aligned}$$

Since u^{n_k} is bounded in $H^1(\Omega)$, $\tau_\Gamma u^{n_k}$ is bounded in $L^2(\Gamma)$, and thus $\tau_\Gamma u^{n_k} \rightharpoonup \tau_\Gamma u^*$ follows. Hence,

$$\lim_{k\to\infty} \int_\Gamma (u^{n_k} - u^*)\varphi \,\mathrm{d}x = 0 \text{ for all } \varphi \in L^2(\Gamma) \tag{4.9}$$

is also a characterization of weak convergence.
Because $(q^{n_k}, u^{n_k}) \in \mathcal{F}(\mathbf{P})$ for all $k \in \mathbb{N}$,

$$\int_\Omega c\nabla u^{n_k} \cdot \nabla\varphi \,\mathrm{d}x + \int_\Omega \beta \cdot \nabla u^{n_k}\, \varphi \,\mathrm{d}x + \int_\Omega q^{n_k} u^{n_k} \varphi \,\mathrm{d}x - \int_\Omega f\varphi \,\mathrm{d}x$$
$$+ \int_\Gamma \sigma u^{n_k} \varphi \mathrm{d}s - \int_\Gamma g\varphi \,\mathrm{d}s = 0$$

holds for all $k \in \mathbb{N}$ and for all $\varphi \in H^1(\Omega)$. We view every integral term in the sum containing u^{n_k} or q^{n_k} separately and check whether it converges (weakly) towards the respective term with q^* and u^*.
It is clear that $\int_\Omega c\nabla u^{n_k} \cdot \nabla\varphi \,\mathrm{d}x$ converges towards $\int_\Omega c\nabla u^* \cdot \nabla\varphi \,\mathrm{d}x$ because

$$\int_\Omega c\nabla u^{n_k} \cdot \nabla\varphi \,\mathrm{d}x - \int_\Omega c\nabla u^* \cdot \nabla\varphi \,\mathrm{d}x = \int_\Omega \nabla(u^{n_k} - u^*) \cdot \nabla(c\varphi) \,\mathrm{d}x,$$

$c\varphi \in H^1(\Omega)$, and (4.7). Because of the weak convergence of $\{u^{n_k}\}_{k\in\mathbb{N}}$ the claim follows.
The term $\int_\Omega \beta \cdot \nabla u^{n_k}\, \varphi \,\mathrm{d}x$ converges towards $\int_\Omega \beta \cdot \nabla u^*\, \varphi \,\mathrm{d}x$ because

$$\int_\Omega \beta \cdot \nabla u^{n_k}\, \varphi \,\mathrm{d}x - \int_\Omega \beta \cdot \nabla u^*\, \varphi \,\mathrm{d}x = \int_\Omega \nabla(u^{n_k} - u^*) \cdot (\varphi\beta) \,\mathrm{d}x,$$

$\varphi\beta \in L^2(\Omega)^d$, and (4.8).
Next we view the term $\int_\Omega q^{n_k} u^{n_k} n\varphi \,\mathrm{d}x$. We observe

$$\begin{aligned}
\left| \int_\Omega q^{n_k} u^{n_k} \varphi \,\mathrm{d}x - \int_\Omega q^* u^* \varphi \,\mathrm{d}x \right| &= \left| \int_\Omega (q^{n_k} u^{n_k} - q^* u^*) \varphi \,\mathrm{d}x \right| \\
&= \left| \int_\Omega (q^{n_k} - q^*) u^{n_k} \varphi \,\mathrm{d}x + \int_\Omega q^* (u^{n_k} - u^*) \varphi \,\mathrm{d}x \right| \\
&\le |q^{n_k} - q^*| \int_\Omega |u^{n_k} \varphi| \,\mathrm{d}x \\
&\quad + \left| \int_\Omega (u^{n_k} - u^*)(q^* \varphi) \,\mathrm{d}x \right| \\
&\le |q^{n_k} - q^*|\, \|u^{n_k}\|_{L^2(\Omega)} \|\varphi\|_{L^2(\Omega)} \\
&\quad + \left| \int_\Omega (u^{n_k} - u^*)(q^* \varphi) \,\mathrm{d}x \right|.
\end{aligned}$$

The second term in the right-hand side converges towards zero because $q^*\varphi \in L^2(\Omega)$ and (4.6). For the first term, we infer $\|u^{n_k}\|_{L^2(\Omega)} \|\varphi\|_{L^2(\Omega)} < \infty$ for

all $k \in \mathbb{N}$ because $\|u^{n_k}\|_{L^2(\Omega)}$ is bounded independently of n_k and φ is in $L^2(\Omega)$, and because q^{n_k} converges towards q^*, the whole term converges towards zero.
Finally, we have to look at the term $\int_\Gamma \sigma u^{n_k}\varphi \mathrm{d}s$. From

$$\int_\Gamma \sigma u^{n_k}\varphi \mathrm{d}s - \int_\Gamma \sigma u^*\varphi \mathrm{d}s = \int_\Gamma (u^{n_k} - u^*)(\sigma\varphi)\mathrm{d}s,$$

$\sigma\varphi \in L^2(\Omega)$, and (4.9) we conclude that $e(q^{n_k}, u^{n_k}) \to e(q^*, u^*)$, and therefore $e(q^*, u^*) = 0$.
It is shown in [39] that any norm, such as the $L^2(\Gamma)$-norm in (4.2), is weakly lower semicontinuous. That means, if $u^n \rightharpoonup u^*$, of course $(u^n - u_\Gamma) \rightharpoonup (u^* - u_\Gamma)$ holds as well, and due to the definition of weak lower semicontinuity (see [39]) we obtain

$$\|u^* - u_\Gamma\|_{L^2(\Gamma)} \le \lim_{k\to\infty} \|u^{n_k} - u_\Gamma\|_{L^2(\Gamma)}.$$

Therefore, we observe

$$\inf_{(q,u)\in\mathcal{F}(\mathbf{P})} J(q,u) = \lim_{k\to\infty} J(q^{n_k}, u^{n_k}) \ge J(q^*, u^*),$$

so that (q^*, u^*) solves $(\mathbf{P})$. ∎

4.2 First-order necessary optimality conditions

The Lagrange function for $(\mathbf{P})$ is given by

$$\mathcal{L}(q,u,p,\lambda) = J(q,u) + \langle e(q,u), p\rangle_{H^1(\Omega)',H^1(\Omega)} + \lambda(q_a - q) \tag{4.10}$$

with the Lagrange multipliers $p \in H^1(\Omega)$ and $\lambda \in \mathbb{R}$.
To formulate first-order necessary optimality conditions for $(\mathbf{P})$ in terms of Lagrange multipliers we have to prove that J and e are Fréchet-differentiable and that $x^* = (q^*, u^*)$ is a regular point.

Definition 4.3 *Let B_1 and B_2 be Banach spaces and $f : B_1 \to B_2$. If there exists an operator $\mathcal{A} \in L(B_1, B_2)$ such that at some point $x \in B_1$*

$$\lim_{\|y\|_{B_1}\searrow 0} \frac{\|f(x+y) - f(x) - \mathcal{A}y\|_{B_2}}{\|y\|_{B_1}} = 0,$$

then $\mathcal{A}y$ is called the Fréchet-differential of $f(x)$ at x – written $\delta f(x;y)$. The operator $\mathcal{A}$ is called the Fréchet-derivative of $f(x)$ at x, and we write $\mathcal{A} = f'(x)$ and $\delta f(x;y) = f'(x)y$.

Remark 4.4 *On the Hilbert space $\mathbb{R} \times H^1(\Omega)$ we define the inner product*

$$\langle (q,u),(\tilde{q},\tilde{u})\rangle_{\mathbb{R}\times H^1(\Omega)} = \langle q,\tilde{q}\rangle_{\mathbb{R}} + \langle u,\tilde{u}\rangle_{H^1(\Omega)} = q\tilde{q} + \langle u,\tilde{u}\rangle_{H^1(\Omega)}$$

for $(q,u),(\tilde{q},\tilde{u}) \in \mathbb{R} \times H^1(\Omega)$ and its induced norm by

$$\|(q,u)\|^2_{\mathbb{R}\times H^1(\Omega)} = |q|^2 + \|u\|^2_{H^1(\Omega)}.$$

Lemma 4.5 *The operator J is Fréchet-differentiable.*

Proof. First of all we compute the directional derivative for J in the direction δu. Therefore we proceed

$$\begin{aligned}
\nabla_u J(q,u)\delta u =& \frac{d}{dt} J(q,u+t\,\delta u)|_{t=0} = \lim_{t\searrow 0} \frac{J(q,u+t\,\delta u) - J(q,u)}{t} \\
=& \lim_{t\searrow 0} \left(\frac{\frac{\alpha}{2}\int_\Gamma |u+t\,\delta u - u_\Gamma|^2 \mathrm{d}s + \frac{\kappa}{2}|q-q_d|^2}{t} \right. \\
& \left. - \frac{\frac{\alpha}{2}\int_\Gamma |u-u_\Gamma|^2 \mathrm{d}s + \frac{\kappa}{2}|q-q_d|^2}{t} \right) \\
=& \lim_{t\searrow 0} \frac{\frac{\alpha}{2}\int_\Gamma (t^2\,\delta u^2 + 2t\,\delta u(u-u_\Gamma))\mathrm{d}s}{t} \\
=& \lim_{t\searrow 0} \frac{\alpha}{2}\int_\Gamma (t\,\delta u^2 + 2\,\delta u(u-u_\Gamma))\mathrm{d}s = \alpha \int_\Gamma (u-u_\Gamma)\delta u\,\mathrm{d}s.
\end{aligned}$$

By Definition 4.3, J is Fréchet-differentiable if

$$\lim_{\|\delta u\|_{H^1(\Omega)}\searrow 0} \frac{|J(q,u+\delta u) - J(q,u) - \nabla_u J(q,u)\delta u|}{\|\delta u\|_{H^1(\Omega)}} = 0.$$

To prove that the directional derivative is also the Fréchet-derivative we

proceed

$$
\begin{aligned}
&|J(q,u+\delta u)-J(q,u)-\nabla_u J(q,u)\delta u| \\
=&\Big|\frac{\alpha}{2}\int_\Gamma |u+\delta u-u_\Gamma|^2\mathrm{d}s+\frac{\kappa}{2}|q-q_d|^2-\frac{\alpha}{2}\int_\Gamma |u-u_\Gamma|^2\mathrm{d}s-\frac{\kappa}{2}|q-q_d|^2 \\
&-\alpha\int_\Gamma (u-u_\Gamma)\delta u\,\mathrm{d}s\Big| \\
=&\Big|\frac{\alpha}{2}\int_\Gamma (|u-u_\Gamma|^2+|\delta u|^2+2(u-u_\Gamma)\delta u)\,\mathrm{d}s-\frac{\alpha}{2}\int_\Gamma |u-u_\Gamma|^2\,\mathrm{d}s \\
&-\alpha\int_\Gamma (u-u_\Gamma)\delta u\,\mathrm{d}s\Big| \\
=&\frac{\alpha}{2}\int_\Gamma |\delta u|^2\mathrm{d}s=\frac{\alpha}{2}\|\delta u\|^2_{L^2(\Gamma)}\le\frac{\alpha C_\Gamma^2}{2}\|\delta u\|^2_{H^1(\Omega)}(=O(\|\delta u\|^2_{H^1(\Omega)}))
\end{aligned}
$$

and therefore

$$
\begin{aligned}
0&\le\lim_{\|\delta u\|_{H^1(\Omega)}\searrow 0}\frac{|J(q,u+\delta u)-J(q,u)-\nabla_u J(q,u)\delta u|}{\|\delta u\|_{H^1(\Omega)}} \\
&\le\lim_{\|\delta u\|_{H^1(\Omega)}\searrow 0}\frac{\frac{\alpha C_\Gamma^2}{2}\|\delta u\|^2_{H^1(\Omega)}}{\|\delta u\|_{H^1(\Omega)}}=\lim_{\|\delta u\|_{H^1(\Omega)}\searrow 0}\frac{\alpha C_\Gamma^2}{2}\|\delta u\|_{H^1(\Omega)}=0.
\end{aligned}
$$

It follows that

$$
\lim_{\|\delta u\|_{H^1(\Omega)}\searrow 0}\frac{|J(q,u+\delta u)-J(q,u)-\nabla_u J(q,u)\delta u|}{\|\delta u\|_{H^1(\Omega)}}=0,
$$

thus J is Fréchet-differentiable with respect to u.
Since q is a scalar, by setting $\nabla_q J(q,u)\delta q=\kappa(q-q_d)\delta q$ we obtain

$$
\begin{aligned}
&|J(q+\delta q,u)-J(q,u)-\nabla_q J(q,u)\delta q|=\Big|\frac{\alpha}{2}\int_\Gamma |u-u_\Gamma|^2\mathrm{d}s+\frac{\kappa}{2}|q+\delta q-q_d|^2 \\
&-\frac{\alpha}{2}\int_\Gamma |u-u_\Gamma|^2\mathrm{d}s-\frac{\kappa}{2}|q-q_d|^2-\kappa(q-q_d)\delta q\Big| \\
=&\Big|\frac{\kappa}{2}|q-q_d|^2+\frac{\kappa}{2}|\delta q|^2+\kappa|q-q_d|\,|\delta q|-\frac{\kappa}{2}|q-q_d|^2-\kappa(q-q_d)\delta q\Big|=\frac{\kappa}{2}|\delta q|^2
\end{aligned}
$$

and therefore

$$
\begin{aligned}
&\lim_{|\delta q|\searrow 0}\frac{|J(q+\delta q,u)-J(q,u)-\nabla_q J(q,u)\delta q|}{|\delta q|} \\
=&\lim_{|\delta q|\searrow 0}\frac{\frac{\kappa}{2}|\delta q|^2}{|\delta q|}=\lim_{|\delta q|\searrow 0}\frac{\kappa}{2}|\delta q|=0.
\end{aligned}
$$

Thus, J is also Fréchet-differentiable with respect to q. ∎

Lemma 4.6 *The bilinear operator e is Fréchet-differentiable for all $(q,u) \in \mathbb{R} \times H^1(\Omega)$.*

Proof. The directional derivative $\nabla e(q,u) : \mathbb{R} \times H^1(\Omega) \to H^1(\Omega)'$ defined by

$$\nabla e(q,u)(\delta q, \delta u) = \frac{d}{dt} e(q + t\delta q, u + t\delta u)\big|_{t=0}$$

is given by

$$\begin{aligned}\langle \nabla e(q,u)(\delta q,\delta u), \varphi\rangle_{H^1(\Omega)',H^1(\Omega)} = &\int_\Omega c\nabla\delta u\cdot\nabla\varphi\,\mathrm{d}x + \int_\Omega \beta\cdot\nabla\delta u\varphi\,\mathrm{d}x \\ &+ \int_\Omega \delta q u\varphi\,\mathrm{d}x + \int_\Omega q\delta u\varphi\,\mathrm{d}x + \int_\Gamma \sigma\delta u\varphi\,\mathrm{d}s\end{aligned}$$

for $(q,u) \in \mathbb{R}\times H^1(\Omega)$, $\varphi \in H^1(\Omega)$ and for the direction $(\delta q, \delta u) \in \mathbb{R}\times H^1(\Omega)$. We observe

$$\begin{aligned}&\|e(q+\delta q, u+\delta u) - e(q,u) - \nabla e(q,u)(\delta q,\delta u)\|_{H^1(\Omega)'} \\ =&\sup_{\|\varphi\|_{H^1(\Omega)}=1} \langle e(q+\delta q,u+\delta u) - e(q,u) - \nabla e(q,u)(\delta q,\delta u),\varphi\rangle_{H^1(\Omega)',H^1(\Omega)} \\ =&\sup_{\|\varphi\|_{H^1(\Omega)}=1} \Big(\langle e(q+\delta q,u+\delta u),\varphi\rangle_{H^1(\Omega)',H^1(\Omega)} - \langle e(q,u),\varphi\rangle_{H^1(\Omega)',H^1(\Omega)} \\ &- \langle \nabla e(q,u)(\delta q,\delta u),\varphi\rangle_{H^1(\Omega)',H^1(\Omega)}\Big) \\ =&\sup_{\|\varphi\|_{H^1(\Omega)}=1} \int_\Omega \delta q\delta u\varphi\,\mathrm{d}x \le \sup_{\|\varphi\|_{H^1(\Omega)}=1} |\delta q|\|\delta u\|_{L^2(\Omega)}\|\varphi\|_{L^2(\Omega)} \\ \le&\sup_{\|\varphi\|_{H^1(\Omega)}=1} |\delta q|\|\delta u\|_{H^1(\Omega)}\|\varphi\|_{H^1(\Omega)} = |\delta q|\|\delta u\|_{H^1(\Omega)} \\ \le&\frac{1}{2}\big(|\delta q|^2 + \|\delta u\|^2_{H^1(\Omega)}\big) \le \frac{1}{2}\|(\delta q,\delta u)\|^2_{\mathbb{R}\times H^1(\Omega)},\end{aligned}$$

thus,

$$\begin{aligned}&\lim_{\|(\delta q,\delta u)\|_{\mathbb{R}\times H^1(\Omega)}\searrow 0} \frac{\|e(q+\delta q,u+\delta u) - e(q,u) - \nabla e(q,u)(\delta q,\delta u)\|_{H^1(\Omega)'}}{\|(\delta q,\delta u)\|_{\mathbb{R}\times H^1(\Omega)}} \\ \le&\lim_{\|(\delta q,\delta u)\|_{\mathbb{R}\times H^1(\Omega)}\searrow 0} \frac{1}{2}\|(\delta q,\delta u)\|_{\mathbb{R}\times H^1(\Omega)} = 0.\end{aligned}$$

Therefore, $\nabla e(q,u)(\delta q,\delta u)$ is the Fréchet-derivative of $e(q,u)$. ∎

Proposition 4.7 *Let the assumptions from Proposition 2.10 hold. Then the operator $\nabla e(q,u)$ is surjective for all $(q,u) \in Q_{ad} \times H^1(\Omega)$.*

Proof. For any $F \in H^1(\Omega)'$ we have to find $(\delta q, \delta u) \in \mathbb{R} \times H^1(\Omega)$ such that

$$\langle \nabla e(q,u)(\delta q, \delta u), \varphi \rangle_{H^1(\Omega)',H^1(\Omega)} = \langle F, \varphi \rangle_{H^1(\Omega)',H^1(\Omega)} \tag{4.11}$$

for all $\varphi \in H^1(\Omega)$.
We choose $\delta q = 0$. Then we obtain

$$\begin{aligned}\langle \nabla e(q,u)(0,\delta u), \varphi \rangle_{H^1(\Omega)',H^1(\Omega)} = &\int_\Omega c\nabla \delta u \cdot \nabla \varphi \,\mathrm{d}x + \int_\Omega \beta \cdot \nabla \delta u \varphi \,\mathrm{d}x \\ &+ \int_\Omega q \delta u \varphi + \int_\Gamma \sigma u \varphi \,\mathrm{d}s.\end{aligned}$$

Using the same arguments as in the proof of Proposition 2.10 we deduce that – provided that $\alpha_2(q_a) > 0$ holds – there exists a unique element δu such that (4.11) is fulfilled for all $(q,u) \in Q_{ad} \times H^1(\Omega)$. ∎

Remark 4.8 *Note that from the proof of Proposition 4.7 it follows that $\nabla_u e(q,u) : H^1(\Omega) \to H^1(\Omega)'$ is bijective.*

Now we are able to determine the Fréchet-derivatives of the Lagrange function with respect to the arguments q, u, and p and therefore formulate the first-order optimality conditions to the minimization problem (**P**):

Theorem 4.9 (Necessary first-order optimality conditions) *Let $x^* = (q^*, u^*)$ be a local solution to (**P**). Moreover, let the hypothesis of Proposition 2.10 be satisfied. Then there exist associated Lagrange multipliers $p^* \in H^1(\Omega)$ and $\lambda^* \geq 0$ such that the optimality condition*

$$\kappa(q^* - q_d)\delta q + \int_\Omega u^* p^* \delta q \,\mathrm{d}x - \lambda^* \delta q = 0 \tag{4.12}$$

holds for all $\delta q \in \mathbb{R}$, the dual/adjoint equation

$$\begin{aligned}&\alpha \int_\Gamma (u^* - u_\Gamma)\delta u \,\mathrm{d}s + \int_\Omega c\nabla p^* \cdot \nabla \delta u \,\mathrm{d}x \\ &+ \int_\Omega p^* \beta \cdot \nabla \delta u \,\mathrm{d}x + \int_\Omega q^* p^* \delta u \,\mathrm{d}x + \int_\Gamma \sigma p^* \delta u \,\mathrm{d}s = 0\end{aligned} \tag{4.13}$$

holds for all $\delta u \in H^1(\Omega)$*, and the state equation*

$$\begin{aligned}(4.14)\qquad &\int_\Omega c\nabla u^* \cdot \nabla \delta p \,\mathrm{d}x + \int_\Omega \beta \cdot \nabla u^* \delta p \,\mathrm{d}x + \int_\Omega q^* u^* \delta p \,\mathrm{d}x \\ &\quad - \int_\Omega f \delta p \,\mathrm{d}x + \int_\Gamma \sigma u^* \delta p \,\mathrm{d}s - \int_\Gamma g \delta p \,\mathrm{d}s = 0\end{aligned}$$

holds for all $\delta p \in H^1(\Omega)$*. Equations* (4.12)–(4.14) *are called Karush-Kuhn-Tucker (KKT) system for* (**P**)*.*

Proof. From Lemma 4.5 and Lemma 4.6 it follows that $\mathcal{L}$ is Fréchet-differentiable. From Proposition 4.7 and [35] it follows that there exist unique Lagrange multipliers $p^* \in H^1(\Omega)$ and $\lambda^* \geq 0$ such that

$$\nabla_q \mathcal{L}(q^*, u^*, p^*, \lambda^*)\delta q = 0 \text{ for all } \delta q \in \mathbb{R}, \tag{4.15}$$

$$\nabla_u \mathcal{L}(q^*, u^*, p^*, \lambda^*)\delta u = 0 \text{ for all } \delta u \in H^1(\Omega), \tag{4.16}$$

and

$$\nabla_p \mathcal{L}(q^*, u^*, p^*, \lambda^*)\delta p = 0 \text{ for all } \delta p \in H^1(\Omega) \tag{4.17}$$

hold. Let us build the derivative of $\mathcal{L}$ with respect to q first:

$$\nabla_q \mathcal{L}(q^*, u^*, p^*, \lambda^*)\delta q = \kappa(q^* - q_d)\delta q + \int_\Omega u^* p^* \delta q \,\mathrm{d}x - \lambda^* \delta q$$

for any $\delta q \in \mathbb{R}$. Because of (4.15) we obtain (4.12).
Similarly, when we differentiate $\mathcal{L}$ with respect to u we derive

$$\begin{aligned}\nabla_u \mathcal{L}(q^*, u^*, p^*, \lambda^*)\delta u =& \alpha \int_\Gamma (u^* - u_\Gamma)\delta u \,\mathrm{d}s + \int_\Omega c\nabla p^* \cdot \nabla \delta u \,\mathrm{d}x \\ &+ \int_\Omega p^* \beta \cdot \nabla \delta u \,\mathrm{d}x + \int_\Omega q^* p^* \delta u \,\mathrm{d}x + \int_\Gamma \sigma p^* \delta u \,\mathrm{d}s.\end{aligned}$$

Due to (4.16), for the optimal point x^* and the optimal Lagrange multipliers p^* and λ^*, this term must be equal to 0 for any direction $\delta u \in H^1(\Omega)$, i.e., in the strong form we obtain the elliptic boundary value problem

$$\begin{aligned}-c\Delta p^* - \beta \cdot \nabla p^* + q^* p^* &= 0 && \text{in } \Omega, \\ c\frac{\partial p^*}{\partial n} + (\sigma + \beta \cdot n)p^* &= \alpha(u_\Gamma - u^*) && \text{on } \Gamma.\end{aligned}$$

By differentiating $\mathcal{L}$ with respect to p, we obtain the weak formulation of the equality constraint:

$$\nabla_p \mathcal{L}(q^*, u^*, p^*, \lambda^*)\delta p = \int_\Omega c\nabla u^* \cdot \nabla \delta p \,\mathrm{d}x + \int_\Omega \beta \cdot \nabla u^* \delta p \,\mathrm{d}x + \int_\Omega q^* u^* \delta p \,\mathrm{d}x - \int_\Omega f\delta p \,\mathrm{d}x + \int_\Gamma \sigma u^* \delta p \,\mathrm{d}s - \int_\Gamma g\delta p \,\mathrm{d}s$$

for all $\delta p \in H^1(\Omega)$. Together with (4.17), this leads to (4.14). Note that the strong form of the state equation (4.14) is nothing else but (4.1). ■

Remark 4.10 *From*

$$\nabla_u \mathcal{L}(q^*, u^*, p^*, \lambda^*) = \nabla_u J(q^*, u^*) + \nabla_u e(q^*, u^*)^* p^* = 0$$

we deduce

$$\nabla_u e(q^*, u^*)^* p^* = -\nabla_u J(q^*, u^*),$$

where $\nabla_u e(q^*, u^*)^* : H^1(\Omega) \to H^1(\Omega)'$ *is the dual operator of* $\nabla_u e(q^*, u^*)$ *satisfying*

$$\langle \nabla_u e(q^*, u^*)^* p, \delta u \rangle_{H^1(\Omega)', H^1(\Omega)} = \langle \nabla_u e(q^*, u^*)\delta u, p \rangle_{H^1(\Omega)', H^1(\Omega)}$$

for all $p, \delta u \in H^1(\Omega)$.
From Remark 4.8 we know that $\nabla_u e(q^*, u^*)$ *is bijective, that means that it is also injective, thus* p^* *is uniquely determined.*

4.3 Augmentation of the inequality constraint

First of all we introduce the modified cost functional

$$J^\varrho_{\hat\lambda}(q, u) = J(q, u) + \frac{1}{2\varrho} \max\{0, \hat\lambda + \varrho(q_a - q)\}^2 \tag{4.18}$$

for $\varrho > 0$ and $\hat\lambda > 0$.

Lemma 4.11 *The inequality condition*

$$q_a - q^* \le 0 \tag{4.19}$$

together with the non-negativity condition

$$\lambda^* \geq 0 \tag{4.20}$$

and the complementarity condition

$$\lambda^*(q_a - q^*) = 0 \tag{4.21}$$

for optimal λ^* and q^* are equivalent to the equality

$$\lambda^* = \max\{0, \lambda^* + \varrho(q_a - q^*)\}. \tag{4.22}$$

Proof. Due to (4.21), at least one of the two inequality conditions (4.19) and (4.20) obviously must be active, that means that the respective equality holds true. To prove that (4.19), (4.20), and (4.21) imply (4.22), we consider three cases.
Let $\lambda^* = 0$ and $q_a - q^* = 0$ hold. Then, of course, (4.22) holds.
Let $\lambda^* > 0$ and $q_a - q^* = 0$ hold. We find that $\max\{0, \lambda^* + \varrho(q_a - q^*)\} = \max\{0, \lambda^* + 0\} = \lambda^*$, thus (4.22) holds.
Let $\lambda^* = 0$ and $q_a - q^* < 0$ hold. We observe $\max\{0, \lambda^* + \varrho(q_a - q^*)\} = \max\{0, 0 + \varrho(q_a - q^*)\} = 0$, therefore (4.22) holds.
Now we presume that (4.22) holds. It is obvious that (4.20) holds. Assume that $q_a - q^* > 0$ holds. Then we deduce for any $\varrho > 0$ that

$$\lambda^* = \max\{0, \lambda^* + \varrho(q_a - q^*)\} = \lambda^* + \varrho(q_a - q^*) > \lambda^*$$

By contradiction we find that $q_a - q^* > 0$ cannot hold, thus (4.19) holds.
Now let $\lambda^* > 0$. By assuming that $q_a - q^* < 0$ holds, we set $\varrho = \dfrac{-\lambda^*}{q_a - q^*} > 0$ and obtain

$$\lambda^* = \max\{0, \lambda^* + \varrho(q_a - q^*)\} = \max\left\{0, \lambda^* - \frac{\lambda^*}{q_a - q^*}(q_a - q^*)\right\} = 0,$$

this is in the contradiction to the assumption that $\lambda^* > 0$. Thus, $q_a - q^* = 0$ and therefore (4.21) holds.
Next we investigate the case where $q_a - q^* < 0$ holds. Assuming that $\lambda^* > 0$ and setting $\varrho = \dfrac{-\lambda^*}{q_a - q^*} > 0$ again, we deduce

$$\lambda^* = \max\{0, \lambda^* + \varrho(q_a - q^*)\} = \max\left\{0, \lambda^* - \frac{\lambda^*}{q_a - q^*}(q_a - q^*)\right\} = 0.$$

This is a contradiction again, thus $\lambda^* > 0$ cannot hold. Therefore, (4.21) must hold. ■

With our new cost functional $J^{\varrho}_{\hat{\lambda}}$ we have already integrated the inequality constraint by the penalization term $\frac{1}{2\varrho}\max\{0, \hat{\lambda} + \varrho(q_a - q)\}^2$. Thus, we obtain the following optimal control problem

$$(\mathbf{P}^{\varrho}_{\hat{\lambda}}) \qquad \min J^{\varrho}_{\hat{\lambda}}(q,u) \text{ s.t. } (q,u) \in \mathbb{R} \times H^1(\Omega) \text{ satisfies (2.1).}$$

It follows as in the proof of Theorem 4.2 that $(\mathbf{P}^{\varrho}_{\hat{\lambda}})$ has a local solution.
Of course, this new minimization problem changes our KKT system, if $\hat{\lambda} + \varrho(q_a - q) > 0$.

The optimization problem $(\mathbf{P}^{\varrho}_{\hat{\lambda}})$ itself is solved by a globalized SQP method. This method is discussed in Section 4.3.1.
A solution to $(\mathbf{P}^{\varrho}_{\hat{\lambda}})$ is also a solution to the original minimization problem $(\mathbf{P})$ for sufficiently large ϱ, i.e. there exists a $\underline{\varrho} > 0$ such that a solution to $(\mathbf{P}^{\varrho}_{\hat{\lambda}})$ is a solution to $(\mathbf{P})$ for $\varrho \geq \underline{\varrho}$. A proof to this claim can be found in [9], for instance.
Summarizing these ideas we introduce the following algorithm:

Algorithm 4.1 (Augmented Lagrangian method) (1) *Choose $\lambda^0 \geq 0$, $\varrho_0 > 0$, and the augmentation factor $\beta^{\varrho} > 1$. Set $k = 0$. Choose an appropriate stopping criterion.*

(2) *Determine a solution $x^{k+1} = (q^{k+1}, u^{k+1})$ of $(\mathbf{P}^{\varrho}_{\hat{\lambda}})$ with $\varrho = \varrho_k$ and $\hat{\lambda} = \lambda^k$ and its associated Lagrange multiplier p^{k+1} by applying Algorithm 4.2 (globalized SQP method).*

(3) *Update the Lagrange multiplier by $\lambda^{k+1} = \max\left\{0, \lambda^k + \varrho_k(q_a - q^k)\right\}$.*

(4) *Unless the chosen stopping rule is satisfied, set $\varrho_{k+1} = \beta^{\varrho}\varrho_k$, $k = k+1$, and continue with step* (2).

The augmented Lagrange function is of the form

$$\begin{aligned}\mathcal{L}^{\varrho}_{\hat{\lambda}}(q,u,p) =& J^{\varrho}_{\hat{\lambda}}(q,u) + \left\langle e(q,u),p\right\rangle_{H^1(\Omega)',H^1(\Omega)}\\ =& \frac{\alpha}{2}\int_\Gamma |u-u_\Gamma|^2 \mathrm{d}s + \frac{\kappa}{2}|q-q_d|^2 + \frac{1}{2\varrho}\max\{0,\hat{\lambda}+\varrho(q_a-q)\}^2\\ &+ \int_\Omega c\nabla u\cdot\nabla p\,\mathrm{d}x + \int_\Omega \beta\cdot\nabla up\,\mathrm{d}x + \int_\Omega qup\,\mathrm{d}x - \int_\Omega fp\,\mathrm{d}x\\ &+ \int_\Gamma \sigma up\,\mathrm{d}s - \int_\Gamma gp\,\mathrm{d}s.\end{aligned}$$

We observe that $\nabla_q\mathcal{L}^{\varrho}_{\hat{\lambda}}(q,u,p)\delta q$ is the only partial derivative that is different compared to the partial derivatives of $\mathcal{L}(q,u,p,\lambda)$. We obtain

$$\nabla_q\mathcal{L}^{\varrho}_{\hat{\lambda}}(q,u,p)\delta q = \kappa(q-q_d)\delta q + \int_\Omega up\delta q\,\mathrm{d}x - \max\{0,\hat{\lambda}+\varrho(q_a-q)\}\delta q.$$

In the Hessian of $\mathcal{L}^{\varrho}_{\hat{\lambda}}(q,u,p)$ the only entry that might differ from its corresponding entry in $\nabla^2\mathcal{L}(q,u,p,\lambda)$ is the one where we calculate $\nabla^2_{(q,q)}\mathcal{L}^{\varrho}_{\hat{\lambda}}(q,u,p)$. Obviously this entry remains unchanged only if $\lambda+\varrho(q_a-q)\le 0$ holds. If $\lambda+\varrho(q_a-q)>0$ holds, we obtain $\nabla^2_{(q,q)}\mathcal{L}^{\varrho}_{\hat{\lambda}}(q,u,p)=\kappa+\varrho$.

4.3.1 SQP method for $(\mathbf{P}^{\varrho}_{\hat{\lambda}})$

We solve the minimization problem $(\mathbf{P}^{\varrho}_{\hat{\lambda}})$ in each (outer) iteration by the method of sequential quadratic programming (SQP) – see, e.g., [37] for a detailed analysis of this method. Applied to our optimization problem this yields the following algorithm:

Algorithm 4.2 (Globalized SQP method) (1) *Set $q^0=q^k$, $u^0=u^k$ and $p^0=p^k$. Set $i=0$. Choose an appropriate stopping criterion. Set $\varrho=\varrho_k$ and $\hat{\lambda}=\lambda^k$.*

(2) *For q^i,u^i,p^i calculate the Hessian $\nabla^2\mathcal{L}^{\varrho}_{\hat{\lambda}}(q^i,u^i,p^i)$ and the Fréchet-derivative $\nabla\mathcal{L}^{\varrho}_{\hat{\lambda}}(q^i,u^i,p^i)$ with $\hat{\lambda}=\lambda^k+\varrho(q_a-q^i)$ for each SQP iterate $x^i=(q^i,u^i)$.*

(3) *Compute the solution of the linear equation system $\nabla^2\mathcal{L}^{\varrho}_{\hat{\lambda}}(q^i,u^i,p^i)\Delta\hat{x} = -\nabla\mathcal{L}^{\varrho}_{\hat{\lambda}}(q^i,u^i,p^i)$ where $\Delta\hat{x}=(\Delta q,\Delta u,\Delta p)^T$.*

(4) *If* $\nabla^2_{xx}\mathcal{L}^{\varrho}_{\hat{\lambda}}(q^i, u^i, p^i)$ *is not coercive, modify the Hessian (see Section 4.3.2) and go back to step* (3).

(5) *Perform a L1-line search in the direction* $\Delta\hat{x}$ *(see Section 4.3.3) to obtain the step size* s.

(6) *Unless the chosen stopping rule is satisfied, set* $q^{i+1} = q^i + s\Delta q$, $u^{i+1} = u^i + s\Delta u$, $p^{i+1} = p^i + s\Delta p$, *set* $i = i + 1$, *and continue with step* (2). *If the stopping criterion is satisfied, set* $q^{k+1} = q^i$, $u^{k+1} = u^i$, *and* $p^{k+1} = p^i$ *and continue with step* (3) *in Algorithm 4.1.*

Since the SQP method is a second-order method, we require to compute the Hessian of the Lagrangian. Therefore we formulate second-order optimality conditions. Second-order conditions for finite-dimensional problems are given in Theorem A.20 and Theorem A.21. For equivalent results in infinite-dimensional problems we refer the reader to [35].

In order to formulate second-order optimality conditions, of course it is necessary that the cost function J and the bilinear operator e are twice Fréchet-differentiable. The proof that these functions are indeed twice Fréchet-differentiable is straightforward and very similar to the proofs of Theorems 4.5 and 4.6.
Moreover, we have already shown in Proposition 4.7 that the operator $\nabla e(q, u)$ is surjective.

For numerical purposes (see Section 4.5) we will investigate sufficient second-order optimality conditions for the generalized Lagrange function which is given by

$$\begin{aligned}\mathcal{L}^{\varrho,\gamma}_{\hat{\lambda}}(q,u,p) =& J^{\varrho}_{\hat{\lambda}}(q,u) + \left\langle e^{\gamma}(q,u), p\right\rangle_{H^1(\Omega)',H^1(\Omega)} \\ =& \frac{\alpha}{2}\int_{\Gamma}|u-u_{\Gamma}|^2 ds + \frac{\kappa}{2}|q-q_d|^2 + \frac{1}{2\varrho}\max\{0, \hat{\lambda}+\varrho(q_a-q)\}^2 \\ &+ \int_{\Omega} c\nabla u\cdot\nabla p\,\mathrm{d}x + \int_{\Omega}\beta\cdot\nabla up\,\mathrm{d}x + \int_{\Omega}\gamma qup\,\mathrm{d}x - \int_{\Omega} fp\,\mathrm{d}x \\ &+ \int_{\Gamma}\sigma up\,\mathrm{d}s - \int_{\Gamma} gp\,\mathrm{d}s.\end{aligned}$$

Note that for $\gamma = 1$ we have $\mathcal{L}^{\varrho,\gamma}_{\hat{\lambda}} \equiv \mathcal{L}^{\varrho}_{\hat{\lambda}}$.

In order to ensure the sufficient second-order optimality condition we make use of the following two lemmas.

Lemma 4.12 *Let $x^* = (q^*, u^*)$ solve (**P**) and let $(\delta q, \delta u) \in \operatorname{Ker}(\nabla e(q^*, u^*))$. Then*

$$\|\delta u\|_{H^1(\Omega)} \leq C_u |\delta q|$$

holds with $C_u := \dfrac{\|u^*\|_{L^2(\Gamma)}}{\alpha_2(q_a)}$.

Proof. Due to the definition of $\operatorname{Ker}(\nabla e(q^*, u^*))$ we have $\nabla e(q^*, u^*)(\delta q, \delta u) = 0$. Therefore

$$\begin{aligned} \langle \nabla e(q^*, u^*)(\delta q, \delta u), \varphi \rangle_{H^1(\Omega)', H^1(\Omega)} &= \int_\Omega c \nabla \delta u \cdot \nabla \varphi \, \mathrm{d}x \\ + \int_\Omega \beta \cdot \nabla \delta u \varphi \, \mathrm{d}x + \int_\Omega \delta q u^* \varphi \, \mathrm{d}x + \int_\Omega q^* \delta u \varphi \, \mathrm{d}x + \int_\Gamma \sigma \delta u \varphi \, \mathrm{d}s &= 0 \end{aligned} \tag{4.23}$$

holds for all $\varphi \in H^1(\Omega)$. We find that (4.23) is the variational equation of the elliptic differential equation

$$-c\Delta \delta u + \beta \cdot \nabla \delta u + q^* \, \delta u = -\delta q \, u^* \qquad \text{in } \Omega, \tag{4.24a}$$

$$c \frac{\partial \delta u}{\partial n} + \sigma \delta u = 0 \qquad \text{on } \Gamma. \tag{4.24b}$$

This differential equation corresponds to the state equation of the equality constraint $e(q^*, \delta u) = 0$, with $f = -\delta q \, u^*$ and $g = 0$. Thus, the functional $F \in H^1(\Omega)'$ as introduced in (2.2) is given by

$$\langle F, \varphi \rangle_{H^1(\Omega)', H^1(\Omega)} = \int_\Omega (-\delta q u^*) \varphi \, \mathrm{d}x.$$

From Remark 2.11 it follows that for any $q \geq q_a$ we have

$$\alpha_2(q) \geq \alpha_2(q_a) = \min \left\{ \frac{c}{2}, q_a - \frac{\|\beta\|^2_{\mathbb{R}^d}}{2c} \right\} + \min\{0, \sigma C_\Gamma^2\} > 0.$$

Hence, the inequality

$$B(\delta u, \delta u; q) \geq \alpha_2(q_a) \|\delta u\|^2_{H^1(\Omega)}$$

holds true for any $q \geq q_a$. Due to Proposition 2.10 there exists a unique solution $\delta u \in H^1(\Omega)$ to (4.24) satisfying

$$B(\delta u, \phi; q) = \langle F, \phi \rangle_{H^1(\Omega)', H^1(\Omega)} \text{ for all } \phi \in H^1(\Omega),$$

thus

$$\langle F, \delta u\rangle_{H^1(\Omega)',H^1(\Omega)} = B(\delta u, \delta u; q) \geq \alpha_2(q_a)\|\delta u\|^2_{H^1(\Omega)}$$

holds. By defining the norm of a functional $F \in H^1(\Omega)'$ as

$$\|F\|_{H^1(\Omega)'} := \sup_{\|\varphi\|_{H^1(\Omega)}=1} \langle F, \varphi\rangle_{H^1(\Omega)',H^1(\Omega)}$$

we deduce

$$\|\delta u\|^2_{H^1(\Omega)} \leq \frac{1}{\alpha_2(q_a)}\langle F, \delta u\rangle_{H^1(\Omega)',H^1(\Omega)} \leq \frac{1}{\alpha_2(q_a)}\|F\|_{H^1(\Omega)'}\|\delta u\|_{H^1(\Omega)},$$

therefore

$$\begin{aligned}
\|\delta u\|_{H^1(\Omega)} \leq & \frac{1}{\alpha_2(q_a)} \sup_{\|\varphi\|_{H^1(\Omega)}=1} \langle F, \varphi\rangle_{H^1(\Omega)',H^1(\Omega)} \\
= & \frac{1}{\alpha_2(q_a)} \sup_{\|\varphi\|_{H^1(\Omega)}=1} \left(- \int_\Omega \delta q\, u^*\, \varphi \,\mathrm{d}x \right) \\
\leq & \frac{1}{\alpha_2(q_a)} \sup_{\|\varphi\|_{H^1(\Omega)}=1} \Big(|\delta q| \left\langle |u^*|, |\varphi| \right\rangle_{L^2(\Omega)} \Big) \\
\leq & \frac{1}{\alpha_2(q_a)} \sup_{\|\varphi\|_{H^1(\Omega)}=1} \Big(|\delta q| \, \|u^*\|_{L^2(\Omega)} \|\varphi\|_{L^2(\Omega)} \Big) \\
\leq & \frac{1}{\alpha_2(q_a)} \sup_{\|\varphi\|_{H^1(\Omega)}=1} \Big(|\delta q| \, \|u^*\|_{L^2(\Omega)} \|\varphi\|_{H^1(\Omega)} \Big) \\
= & \frac{1}{\alpha_2(q_a)} |\delta q| \, \|u^*\|_{L^2(\Omega)} = C_u |\delta q|,
\end{aligned}$$

which gives the claim. ∎

Remark 4.13 *For the optimization problem* $(\mathbf{P}^{\varrho}_{\hat{\lambda}})$ *the optimality condition* (4.12) *is replaced by*

$$\kappa(q^* - q_d)\delta q + \int_\Omega up\delta q \,\mathrm{d}x - \max\{0, \hat{\lambda} + \varrho(q_a - q^*)\}\delta q = 0. \tag{4.25}$$

The other two optimality conditions, namely (4.13) *and* (4.14)*, remain unchanged.*

Lemma 4.14 *Let (q^*, u^*, p^*) be a solution of the first-order necessary optimality conditions* (4.25)*,* (4.13)*, and* (4.14)*. Then*

$$\|p^*\|_{H^1(\Omega)} \leq \frac{C_\Gamma}{\hat{\alpha}_2(q_a)} \|\alpha(u_\Gamma - u^*)\|_{L^2(\Gamma)}$$

holds with $\hat{\alpha}_2(q_a) = \min\left\{\frac{c}{2}, q - \frac{\|\beta\|^2_{\mathbb{R}^d}}{2c}\right\} + \min\{0, (\sigma + \beta \cdot n)C_\Gamma^2\} > 0.$

Proof. The strong form of the adjoint equation (4.13) is given by

$$\begin{aligned} -c\Delta p^* - \beta \cdot \nabla p^* + q^* p^* &= 0 && \text{in } \Omega, \\ c\frac{\partial p^*}{\partial n} + (\sigma + \beta \cdot n)p^* &= \alpha(u_\Gamma - u^*) && \text{on } \Gamma. \end{aligned}$$

We define the bilinear operator $\hat{B}(\cdot\,, \cdot\,; q)$ for any fixed $q \geq q_a$ by

$$\hat{B}(p^*, \phi; q) = \int_\Omega c\nabla p^* \cdot \nabla \phi \,\mathrm{d}x + \int_\Omega p^* \beta \cdot \nabla \phi \,\mathrm{d}x + \int_\Omega q^* p^* \phi \,\mathrm{d}x + \int_\Gamma \sigma p^* \phi \,\mathrm{d}s.$$

Analogous to the proof of Proposition 2.10 we infer that for $\hat{\alpha}_2(q_a) = \min\left\{\frac{c}{2}, q_a - \frac{\|\beta\|^2_{\mathbb{R}^d}}{2c}\right\} + \min\{0, (\sigma + \beta \cdot n)C_\Gamma^2\} > 0$ and $\hat{F} \in H^1(\Omega)'$ given by

$$\langle \hat{F}, \varphi \rangle_{H^1(\Omega)', H^1(\Omega)} = \int_\Gamma \alpha(u_\Gamma - u^*)\varphi \,\mathrm{d}s \text{ for } \varphi \in H^1(\Omega)$$

the inequality

$$\hat{B}(p^*, p^*; q) \geq \hat{\alpha}_2(q_a)\|p^*\|^2_{H^1(\Omega)}$$

holds. Moreover, we have

$$\hat{B}(p^*, \phi; q) = \langle \hat{F}, \phi \rangle_{H^1(\Omega)', H^1(\Omega)} \text{ for all } \phi \in H^1(\Omega).$$

We deduce

$$\hat{\alpha}_2(q_a)\|p^*\|^2_{H^1(\Omega)} \leq \hat{B}(p^*, p^*; q) = \langle \hat{F}, p^* \rangle_{H^1(\Omega)', H^1(\Omega)} \leq \|\hat{F}\|_{H^1(\Omega)'} \|p^*\|_{H^1(\Omega)},$$

thus we find

$$\begin{aligned}\|p^*\|_{H^1(\Omega)} \leq & \frac{1}{\hat{\alpha}_2(q_a)}\|\hat{F}\|_{H^1(\Omega)'} = \frac{1}{\hat{\alpha}_2(q_a)} \sup_{\|\varphi\|_{H^1(\Omega)}=1} \langle \hat{F}, \varphi\rangle_{H^1(\Omega)',H^1(\Omega)} \\ = & \frac{1}{\hat{\alpha}_2(q_a)} \sup_{\|\varphi\|_{H^1(\Omega)}=1} \int_\Gamma \alpha(u_\Gamma - u^*)\varphi \,\mathrm{d}s \\ \leq & \frac{1}{\hat{\alpha}_2(q_a)} \sup_{\|\varphi\|_{H^1(\Omega)}=1} \big(\|\alpha(u_\Gamma - u^*)\|_{L^2(\Gamma)}\|\varphi\|_{L^2(\Gamma)}\big) \\ \leq & \frac{1}{\hat{\alpha}_2(q_a)} \sup_{\|\varphi\|_{H^1(\Omega)}=1} \big(\|\alpha(u_\Gamma - u^*)\|_{L^2(\Gamma)}C_\Gamma\|\varphi\|_{H^1(\Omega)}\big) \\ = & \frac{C_\Gamma}{\hat{\alpha}_2(q_a)}\|\alpha(u_\Gamma - u^*)\|_{L^2(\Gamma)},\end{aligned}$$

which gives the claim. ∎

In [35] the following sufficient second-order optimality condition for infinite-dimensional problems is proved.

Theorem 4.15 *Let (q^*, u^*, p^*) be a solution to the first order optimality conditions* (4.25), (4.13), *and* (4.14). *If $\nabla^2_{xx}\mathcal{L}^{\varrho,\gamma}_{\hat{\lambda}}(q^*, u^*, p^*)$ is coercive on* $\mathrm{Ker}(\nabla e(q^*, u^*))$, *then (q^*, u^*, p^*) solves $(\mathbf{P}^\varrho_{\hat{\lambda}})$.*

The next theorem gives a sufficient condition that the assumption of Theorem 4.15 is satisfied for $(\mathbf{P}^\varrho_{\hat{\lambda}})$.

Theorem 4.16 *Let (q^*, u^*, p^*) be a solution to the first order optimality conditions* (4.25), (4.13), *and* (4.14). *If*

$$\gamma < \frac{\hat{\alpha}_2(q_a)\min\left\{\frac{\kappa}{2}, \frac{\kappa}{2C_u^2}\right\}}{C_\Gamma\|\alpha(u_\Gamma - u^*)\|_{L^2(\Gamma)}} \tag{4.26}$$

holds, then $\nabla^2_{xx}\mathcal{L}^{\varrho,\gamma}_{\hat{\lambda}}(q^*, u^*, p^*)$ is coercive on $\mathrm{Ker}(\nabla e(q^*, u^*))$, *thus (q^*, u^*, p^*) solves $(\mathbf{P}^\varrho_{\hat{\lambda}})$.*

Proof. We have

$$\nabla^2_{xx}\mathcal{L}^{\varrho,\gamma}_{\hat{\lambda}}(q^*, u^*, p^*)\big((\delta q, \delta u), (\delta q, \delta u)\big) = \kappa|\delta q|^2 + 2\gamma\delta q\int_\Gamma p\,\delta u\,\mathrm{d}s + \alpha\|\delta u\|^2_{L^2(\Gamma)}$$

for $(\delta q, \delta u) \in \mathbb{R} \times H^1(\Omega)$ and need to show that for (4.26) the coercivity of $\nabla^2_{xx}\mathcal{L}^{\varrho,\gamma}_{\hat{\lambda}}(q^*, u^*, p^*)$ on $\mathrm{Ker}(\nabla e(q^*, u^*))$ is guaranteed, i.e., that

$$\nabla^2_{xx}\mathcal{L}^{\varrho,\gamma}_{\hat{\lambda}}(q^*, u^*, p^*)\big((\delta q, \delta u), (\delta q, \delta u)\big) \geq \eta \|(\delta q, \delta u)\|^2_{\mathbb{R}\times H^1(\Omega)} \tag{4.27}$$

holds for $(\delta q, \delta u) \in \mathrm{Ker}(\nabla e(q^*, u^*))$ and a constant $\eta > 0$.
Let us first consider the case where $\gamma = 0$ holds. Clearly (4.26) holds. Because of Lemma 4.12 we obtain

$$\begin{aligned}
\nabla^2_{xx}\mathcal{L}^{\varrho,0}_{\hat{\lambda}}(q^*, u^*, p^*)\big((\delta q, \delta u), (\delta q, \delta u)\big) =& \kappa|\delta q|^2 + \alpha\|\delta u\|^2_{L^2(\Gamma)} \\
=& \frac{\kappa}{2}|\delta q|^2 + \frac{\kappa}{2}|\delta q|^2 + \alpha\|\delta u\|^2_{L^2(\Gamma)} \\
\geq& \frac{\kappa}{2}|\delta q|^2 + \frac{\kappa}{2C_u^2}\|\delta u\|^2_{H^1(\Omega)} \\
\geq& \min\left\{\frac{\kappa}{2}, \frac{\kappa}{2C_u^2}\right\}\|(\delta q, \delta u)\|^2_{\mathbb{R}\times H^1(\Omega)}.
\end{aligned}$$

Thus, (4.27) is satisfied with $\eta := \min\left\{\frac{\kappa}{2}, \frac{\kappa}{2C_u^2}\right\} > 0$.
Now suppose that $\gamma > 0$. Using Lemma 4.14 we deduce

$$\begin{aligned}
&\nabla^2_{xx}\mathcal{L}^{\varrho,\gamma}_{\hat{\lambda}}(q^*, u^*, p^*)\big((\delta q, \delta u), (\delta q, \delta u)\big) \\
\geq& \frac{\kappa}{2}|\delta q|^2 + \frac{\kappa}{2C_u^2}\|\delta u\|^2_{H^1(\Omega)} - 2\gamma\int_\Omega |\delta q|\,|p^*|\,|\delta u|\,\mathrm{d}x \\
\geq& \frac{\kappa}{2}|\delta q|^2 + \frac{\kappa}{2C_u^2}\|\delta u\|^2_{H^1(\Omega)} - 2\gamma|\delta q|\,\|p^*\|_{L^2(\Omega)}\,\|\delta u\|_{L^2(\Omega)} \\
\geq& \frac{\kappa}{2}|\delta q|^2 + \frac{\kappa}{2C_u^2}\|\delta u\|^2_{H^1(\Omega)} - 2\gamma|\delta q|\,\|p^*\|_{H^1(\Omega)}\,\|\delta u\|_{H^1(\Omega)} \\
\geq& \frac{\kappa}{2}|\delta q|^2 + \frac{\kappa}{2C_u^2}\|\delta u\|^2_{H^1(\Omega)} - 2\gamma\frac{C_\Gamma}{\hat{\alpha}_2(q_a)}\|\alpha(u_\Gamma - u^*)\|_{L^2(\Gamma)}|\delta q|\,\|\delta u\|_{H^1(\Omega)} \\
\geq& \frac{\kappa}{2}|\delta q|^2 + \frac{\kappa}{2C_u^2}\|\delta u\|^2_{H^1(\Omega)} - 2\gamma\frac{C_\Gamma}{\hat{\alpha}_2(q_a)}\|\alpha(u_\Gamma - u^*)\|_{L^2(\Gamma)}\frac{1}{2}\big(|\delta q|^2 + \|\delta u\|^2_{H^1(\Omega)}\big) \\
\geq& \left(\min\left\{\frac{\kappa}{2}, \frac{\kappa}{2C_u^2}\right\} - \gamma\frac{C_\Gamma}{\hat{\alpha}_2(q_a)}\|\alpha(u_\Gamma - u^*)\|_{L^2(\Gamma)}\right)\|(\delta q, \delta u)\|^2_{\mathbb{R}\times H^1(\Omega)}.
\end{aligned}$$

Thus, if (4.26) holds, the constant

$$\hat{\eta} := \min\left\{\frac{\kappa}{2}, \frac{\kappa}{2C_u^2}\right\} - \gamma\frac{C_\Gamma}{\hat{\alpha}_2(q_a)}\|\alpha(u_\Gamma - u^*)\|_{L^2(\Gamma)}$$

is larger than zero, and (4.27) holds for $\eta = \hat{\eta}$. ∎

Remark 4.17 *Due to Theorem 4.16, coercivity of* $\nabla^2_{xx}\mathcal{L}^{\varrho,\gamma}_{\hat{\lambda}}(q^*,u^*,p^*)$ *is guaranteed if*

$$\gamma\|\alpha(u_\Gamma - u^*)\|_{L^2(\Gamma)} < \frac{\hat{\alpha}_2(q_a)}{C_\Gamma}\min\left\{\frac{\kappa}{2}, \frac{\kappa}{2C_u^2}\right\}$$

holds, i.e., the term $\gamma\|\alpha(u_\Gamma - u^*)\|_{L^2(\Gamma)}$ *must be sufficiently small. Thus, if* u^* *is sufficiently close to the measurements* u_Γ *on the boundary* Γ *in the* L^2*-norm, the damping parameter* γ *(compare Section 4.3.2) apparently can be set equal to one.*

The next theorem ensures local convergence of the SQP method with a convergence rate 2.

Theorem 4.18 (Convergence theorem) *Let* (q^*,u^*) *be a solution to* $(\mathbf{P}^{\varrho}_{\hat{\lambda}})$*. Moreover, let* $\nabla e(q^*,u^*)$ *be surjective, and let* $\nabla^2_{xx}\mathcal{L}^{\varrho}_{\hat{\lambda}}(q^*,u^*,p^*)$ *be coercive on* $\mathrm{Ker}(\nabla e(q^*,u^*))$*. Then, since the SQP method is locally equivalent with the Newton method applied to first-order optimality conditions, the convergence of the SQP method we apply in our problem is local quadratic, i.e., there exists a* $\rho > 0$ *and a* $C > 0$ *such that*

$$\begin{aligned}&\|(q^{i+1},u^{i+1},p^{i+1}) - (q^*,u^*,p^*)\|_{\mathbb{R}\times H^1(\Omega)\times H^1(\Omega)}\\ \leq\, &C\|(q^i,u^i,p^i) - (q^*,u^*,p^*)\|^2_{\mathbb{R}\times H^1(\Omega)\times H^1(\Omega)} \text{ for all } i\in\mathbb{N},\end{aligned}$$

if for the starting iterate $\|(q^0,u^0,p^0) - (q^*,u^*,p^*)\|_{\mathbb{R}\times H^1(\Omega)\times H^1(\Omega)} < \rho$ *holds.*

Remark 4.19 *The indefinite and symmetric linear equation system in step* (3) *of Algorithm 4.2 can be solved, e.g., by means of the LU decomposition or by an iterative method like GMRES. For the GMRES algorithm it is useful to apply preconditioners to* $\nabla^2\mathcal{L}^{\varrho}_{\hat{\lambda}}(q^i,u^i,p^i)$ *– see [6],[7], and [8] for a detailed review and analysis of preconditioners.*

4.3.2 Damping of the Hessian

With $\nabla^2_{xx}\mathcal{L}^{\varrho}_{\hat{\lambda}}(q,u,p)$ we denote the Hessian consisting of the second partial derivatives with respect to the variables $x=(q,u)$, only. That means that $\nabla^2_{xx}\mathcal{L}^{\varrho}_{\hat{\lambda}}(q,u,p)$ is of the form

$$\nabla^2_{xx}\mathcal{L}^{\varrho}_{\hat{\lambda}}(q,u,p) = \begin{pmatrix} \kappa+\varrho & p\\ p & \alpha\end{pmatrix}$$

if $\lambda + \varrho(q_a - q) > 0$ holds, and of the form

$$\nabla^2_{xx}\mathcal{L}^{\varrho}_{\hat{\lambda}}(q,u,p) = \begin{pmatrix} \kappa & p \\ p & \alpha \end{pmatrix}$$

if $\lambda + \varrho(q_a - q) \leq 0$ holds. Recall that the operator $\nabla^2_{xx}\mathcal{L}^{\varrho}_{\hat{\lambda}}(q,u,p)$ needs to be coercive on $\mathrm{Ker}\nabla e(q,u)$ in order to obtain a descent direction (see, e.g., [37]). This condition is not necessarily fulfilled for all iterates $\hat{x} = (q,u,p)$. In order to evade this problem we aim to damp the entries of $\nabla^2_{xx}\mathcal{L}^{\varrho}_{\hat{\lambda}}(q,u,p)$ that possibly cause the non-coercivity of the Hessian. These entries are the partial derivatives $\nabla^2_{(q,u)}\mathcal{L}^{\varrho}_{\hat{\lambda}}(q,u,p)$ and $\nabla^2_{(u,q)}\mathcal{L}^{\varrho}_{\hat{\lambda}}(q,u,p)$. We modify the operator $\nabla^2_{xx}\mathcal{L}^{\varrho}_{\hat{\lambda}}(q,u,p)$ a little and obtain the damped Hessian $\nabla^2_{xx}\mathcal{L}^{\varrho,\gamma}_{\hat{\lambda}}(q,u,p)$:

$$\nabla^2_{xx}\mathcal{L}^{\varrho,\gamma}_{\hat{\lambda}}(q,u,p) = \begin{pmatrix} \kappa(+\varrho) & \gamma p \\ \gamma p & \alpha \end{pmatrix},$$

see [18].
Numerically, we require

$$\nabla^2_{xx}\mathcal{L}^{\varrho,\gamma}_{\hat{\lambda}}(q^*,u^*,p^*)\big((\Delta q,\Delta u),(\Delta q,\Delta u)\big) \geq \eta\big(|\Delta q|^2 + \|\Delta u\|^2_{H^1(\Omega)}\big) \tag{4.28}$$

for the direction $(\Delta q, \Delta u)$ as computed in step (3) of Algorithm 4.3.1, where η is a fixed value larger than zero. In any iteration i we have

$$\begin{aligned}\nabla^2_{xx}\mathcal{L}^{\varrho,\gamma}_{\hat{\lambda}}(q^i,u^i,p^i)\big((\Delta q^i,\Delta u^i),(\Delta q^i,\Delta u^i)\big) =& \kappa|\Delta q^i|^2 + 2\gamma\Delta q^i\int_\Gamma p^i\,\Delta u^i\,\mathrm{d}s \\ &+ \alpha\|\Delta u^i\|^2_{L^2(\Gamma)}.\end{aligned}$$

For $\gamma = 0$, we obtain a positive definite operator $\nabla^2_{xx}\mathcal{L}^{\varrho,0}_{\hat{\lambda}}(q^i,u^i,p^i)$ because of the positiveness of α and the positiveness of κ. Yet, we aim to approximate $\nabla^2_{xx}\mathcal{L}^{\varrho}_{\hat{\lambda}}(q^i,u^i,p^i)$ sufficiently well. Therefore, we try to find a value for γ between 0 and 1 that ensures (4.28). Note that for

$$\gamma \leq C_\gamma := \frac{\eta\big(|\Delta q^i|^2 + \|\Delta u^i\|^2_{H^1(\Omega)}\big) - \kappa|\Delta q^i|^2 - \alpha\|\Delta u^i\|^2_{H^1(\Omega)}}{2\Delta q^i\int_\Gamma p^i\Delta u^i\,\mathrm{d}s}$$

we obviously obtain (4.28). Thus, one possibility is to set

$$\gamma^i := \max\big\{0, \min\{\xi C_\gamma, 1\}\big\}$$

in each SQP-iteration i, with $\xi = 0.9$, for instance.

4.3.3 Line search

The globalized augmented Lagrange SQP method still does not necessarily need to converge, even if the iteration sequence appears to yield positive definite Hessian matrices without exception. Therefore we try to overcome this unpleasant state by applying an appropriate line search.
One possible line search technique is the Armijo-rule (see, e.g., [18]) which demands that for the step size s the following inequality holds:

$$\varphi^i(s) - \varphi^i(0) \leq \varepsilon s(\bar{\varphi}^i(1) - \bar{\varphi}^i(0)) \tag{4.29}$$

for a fixed constant $\varepsilon \in [10^{-4}, 10^{-3}]$, where in each iteration i the function φ^i is defined for $x^i = (q^i, u^i)$ by

$$\varphi^i(s) = J(x^i + s\Delta x^i) + \mu \|e(x^i + s\Delta x^i)\|_{H^1(\Omega)} \text{ for } s \in [0,1]$$

and its linearization $\bar{\varphi}^i$ is defined by

$$\bar{\varphi}^i(s) = J(x^i) + s\nabla J(x^i)\Delta x^i + \mu \|e(x^i) + s\nabla e(x^i)\Delta x^i\|_{H^1(\Omega)} \text{ for } s \in [0,1]$$

for a penalty parameter $\mu \geq 0$.
Note that $\nabla e(x^i)\Delta x^i = -e(x^i)$ holds (because of the third row of the KKT system), and therefore we have

$$\bar{\varphi}^i(s) = J(x^i) + s\nabla J(x^i)\Delta x^i + \mu \|e(x^i)\|_{H^1(\Omega)}(1 - s) \text{ for } s \in [0,1]$$

The penalty term $\mu\|e(x^i)\|_{H^1(\Omega)}$ in the so-called merit function ('exact penalty function', see [42]) can be motivated by the fact that we want to trade off the descent of the cost functional J and the measuring of the violation of the equality constraint $e(x^i) = 0$. For a sufficiently large parameter μ the descent of the function φ^i is guaranteed.
Numerically, one has to augment the parameter μ until the condition

$$\bar{\varphi}^i(1) - \bar{\varphi}^i(0) < 0$$

holds. This value for μ is inserted into the functions φ^i and $\bar{\varphi}^i$. Now we can start the actual Armijo-type line search. That is, we decrease the step size s – for instance by one half – until (4.29) is fulfilled.

4.4 Galerkin approximation of the SQP algorithm

In our numerical experiments we approximate the state u by a linear combination of finite elements (see Section 2.3) or by a linear combination of POD basis functions which we compute as described in Section 3.3.
Throughout this work, we will denote the FE based solution by u^h and the POD based solution by u^ℓ. The coefficients to the respective basis functions (the FE basis functions are denoted by φ_i for $i = 1, ..., n_{FE}$ and the POD basis functions by ψ_i for $i = 1, ..., \ell$) are denoted by u_i^h and u_i^ℓ, that means

$$u^h(x) = \sum_{i=1}^{n_{FE}} \mathrm{u}_i^h \varphi_i(x)$$

and

$$u^\ell(x) = \sum_{i=1}^{\ell} \mathrm{u}_i^\ell \psi_i(x).$$

The same can be done with the Lagrange multiplier $p \in H^1(\Omega)$, thus we define the FE-based approximation of p by

$$p^h(x) = \sum_{i=1}^{n_{FE}} \mathrm{p}_i^h \varphi_i(x)$$

and the POD-based approximation by

$$p^\ell(x) = \sum_{i=1}^{\ell} \mathrm{p}_i^\ell \psi_i(x),$$

where p_i^h and p_i^ℓ are the coefficients to the respective Galerkin ansatz functions.

We want to discretize the continuous optimality conditions (4.12)–(4.14) for the problem (**P**) by a finite-dimensional nonlinear system.
Let us take a look at the L^2-inner product of the functions u^h and p^h first:

$$\begin{aligned}\int_\Omega u^h(x)p^h(x)\,\mathrm{d}x &= \int_\Omega \sum_{i=1}^{n_{FE}} \mathrm{u}_i^h \varphi_i(x) \sum_{j=1}^{n_{FE}} \mathrm{p}_j^h \varphi_j(x)\,\mathrm{d}x \\ &= \sum_{i=1}^{n_{FE}} \sum_{j=1}^{n_{FE}} \mathrm{u}_i^h \int_\Omega \varphi_i(x)\varphi_j(x)\,\mathrm{d}x\,\mathrm{p}_j^h = \sum_{i=1}^{n_{FE}} \sum_{j=1}^{n_{FE}} \mathrm{u}_i^h \mathrm{M}_{ij}^h\,\mathrm{p}_j^h.\end{aligned}$$

By M^h we define the symmetric and positive semidefinite FE-based mass matrix whose entries M^h_{ij} are the L^2-inner products of the FE basis functions φ_i and φ_j for $1 \leq i, j \leq n_{FE}$ (compare (3.15) in Section 3.3). Thus, the FE-discretization of the first optimality condition (4.12) leads to

$$\kappa(q - q_d) + (\mathrm{u}^h)^T \mathrm{M}^h \mathrm{p}^h - \lambda = 0, \tag{4.30}$$

where u^h and p^h are the vectors containing the coefficients $\{\mathrm{u}^h_i\}_{i=1}^{n_{FE}}$ and $\{\mathrm{p}^h_i\}_{i=1}^{n_{FE}}$ in the FE approximations, respectively.

Let us introduce the symmetric and positive semidefinite matrix $\mathrm{BD}^h \in \mathbb{R}^{n_{FE} \times n_{FE}}$ by

$$\mathrm{BD}^h_{ij} = \int_\Gamma \varphi_i(x)\varphi_j(x)\,\mathrm{d}s \text{ for } 1 \leq i, j \leq n_{FE}.$$

Note that $\mathrm{BD}^h_{ij} = \langle \varphi_i(x), \varphi_j(x) \rangle_{L^2(\Gamma)}$ holds for $1 \leq i, j \leq n_{FE}$. Hence, we can write the term $\alpha \int_\Gamma (u^h - u^h_\Gamma)\delta u^h \,\mathrm{d}s$ as

$$\begin{aligned}
\alpha \int_\Gamma (u^h - u^h_\Gamma)\delta u^h \,\mathrm{d}s &= \alpha \int_\Gamma \sum_{i=1}^{n_{FE}} (\mathrm{u}^h_i - (\mathrm{u}^h_\Gamma)_i)\varphi_i(x) \sum_{j=1}^{n_{FE}} \delta \mathrm{u}^h_j \varphi_j(x)\,\mathrm{d}s \\
&= \alpha \sum_{i=1}^{n_{FE}} \sum_{j=1}^{n_{FE}} (\mathrm{u}^h_i - (\mathrm{u}^h_\Gamma)_i) \int_\Gamma \varphi_i(x)\varphi_j(x)\,\mathrm{d}s\, \delta \mathrm{u}^h_j \\
&= \alpha (\mathrm{u}^h - \mathrm{u}^h_\Gamma)^T \mathrm{BD}^h \delta \mathrm{u}^h,
\end{aligned}$$

where $\delta \mathrm{u}^h$ is the vector containing the coefficients $\{\delta \mathrm{u}^h_i\}_{i=1}^{n_{FE}}$ of the FE basis functions, again.
Furthermore, we define the symmetric matrix and positive semidefinite matrix $\hat{\mathrm{S}}^h \in \mathbb{R}^{n_{FE} \times n_{FE}}$ by

$$\hat{\mathrm{S}}^h_{ij} = \int_\Omega \int_\Omega \nabla \varphi_i(x) \cdot \nabla \varphi_j(x)\,\mathrm{d}x \text{ for } 1 \leq i, j \leq n_{FE}.$$

Note that $\hat{\mathrm{S}}^h = \mathrm{S}^h - \mathrm{M}^h$ holds, where S^h stands for the stiffness matrix as introduced in (3.16).

Then we obtain

$$\int_\Omega c\nabla p^h \cdot \nabla \delta u^h \,\mathrm{d}x = c\int_\Omega \nabla\Big(\sum_{i=1}^{n_{FE}} \mathrm{p}_i^h \varphi_i(x)\Big)\cdot\nabla\Big(\sum_{j=1}^{n_{FE}} \delta\mathrm{u}_j^h \varphi_j(x)\Big)\,\mathrm{d}x$$
$$= \sum_{i=1}^{n_{FE}}\sum_{j=1}^{n_{FE}} c\mathrm{p}_i^h \int_\Omega \nabla\varphi_i(x)\cdot\nabla\varphi_j(x)\,\mathrm{d}x\,\delta\mathrm{u}_j^h = (\mathrm{p}^h)^T c\,\hat{\mathrm{S}}^h \delta\mathrm{u}^h.$$

Defining the non-symmetric matrix $\mathrm{BE}^h \in \mathbb{R}^{n_{FE}\times n_{FE}}$ by

$$\mathrm{BE}_{ij}^h = \int_\Omega \varphi_i(x)\beta\cdot\nabla\varphi_j(x)\,\mathrm{d}x \text{ for } 1\le i,j\le n_{FE}$$

we obtain

$$\int_\Omega p^h\beta\cdot\nabla\delta u^h\,\mathrm{d}x = \int_\Omega \sum_{i=1}^{n_{FE}} \mathrm{p}_i^h\varphi_i(x)\beta\cdot\nabla\Big(\sum_{j=1}^{n_{FE}}\delta\mathrm{u}_j^h\varphi_j(x)\Big)\,\mathrm{d}x$$
$$= \sum_{i=1}^{n_{FE}}\sum_{j=1}^{n_{FE}} \mathrm{p}_i^h\int_\Omega \varphi_i(x)\beta\cdot\nabla\varphi_j(x)\,\mathrm{d}x\,\delta\mathrm{u}_j^h = (\mathrm{p}^h)^T\mathrm{BE}^h\delta\mathrm{u}^h.$$

Furthermore, we observe

$$\int_\Omega qp^h\delta u^h\,\mathrm{d}x = \int_\Omega q\sum_{i=1}^{n_{FE}}\mathrm{p}_i^h\varphi_i(x)\sum_{j=1}^{n_{FE}}\delta\mathrm{u}_j^h\varphi_j(x)$$
$$= q\sum_{i=1}^{n_{FE}}\sum_{j=1}^{n_{FE}}\mathrm{p}_i^h\int_\Omega\varphi_i(x)\varphi_j(x)\,\mathrm{d}x\,\delta\mathrm{u}_j^h = q(\mathrm{p}^h)^T\mathrm{M}^h\delta\mathrm{u}^h$$

and

$$\int_\Gamma \sigma p^h\delta u^h\,\mathrm{d}s = \int_\Gamma \sigma\sum_{i=1}^{n_{FE}}\mathrm{p}_i^h\varphi_i(x)\sum_{j=1}^{n_{FE}}\delta\mathrm{u}_j^h\varphi_j(x)\,\mathrm{d}s$$
$$= \sigma\sum_{i=1}^{n_{FE}}\sum_{j=1}^{n_{FE}}\mathrm{p}_i^h\int_\Gamma\varphi_i(x)\varphi_j(x)\,\mathrm{d}s\,\delta\mathrm{u}_j^h = \sigma(\mathrm{p}^h)^T\mathrm{BD}^h\delta\mathrm{u}^h.$$

The FE-discretization of the optimality condition (4.13) is thus given by

$$\alpha\mathrm{BD}^h(\mathrm{u}^h-\mathrm{u}_\Gamma^h) + (c\,\hat{\mathrm{S}}^h + (\mathrm{BE}^h)^T + q\mathrm{M}^h + \sigma\mathrm{BD}^h)\mathrm{p}^h = 0. \tag{4.31}$$

We define the vectors $\mathrm{f}^h \in \mathbb{R}^{n_{FE}}$ and $\mathrm{g}^h \in \mathbb{R}^{n_{FE}}$ by

$$\mathrm{f}_i^h = \int_\Omega f(x)\varphi_i(x)\,\mathrm{d}x \text{ for } 1 \le i \le n_{FE}$$

and

$$\mathrm{g}_i^h = \int_\Gamma g(x)\varphi_i(x)\,\mathrm{d}x \text{ for } 1 \le i \le n_{FE}.$$

Similar to (4.30) and (4.31), we apply the FE-discretization to the optimality condition (4.14) and obtain

$$(c\hat{\mathrm{S}}^h + \mathrm{BE}^h + q\mathrm{M}^h + \sigma\mathrm{BD}^h)\mathrm{u}^h - \mathrm{f}^h - \mathrm{g}^h = 0. \tag{4.32}$$

Next we turn to the POD Galerkin approximation of the optimality system (4.12)–(4.14).
The POD basis functions $\{\psi_i^h\}_{i=1}^{\ell}$ themselves can be written as linear combinations of the finite elements $\{\varphi_i\}_{i=1}^{n_{FE}}$, i.e., for $i \in \{1, ...\ell\}$:

$$\psi_i^h(x) = \sum_{l=1}^{n_{FE}} \mathrm{U}_{li}\varphi_l(x).$$

The resulting matrix containing the FE coefficients of the POD basis functions is denoted by $\mathrm{U} = ((\mathrm{U}_{ij})) \in \mathbb{R}^{n_{FE}\times\ell}$.

When we look at the L^2-inner product of the POD-based functions u^ℓ and p^ℓ we obtain

$$\begin{aligned}\int_\Omega u^\ell(x)p^\ell(x)\,\mathrm{d}x &= \int_\Omega \sum_{i=1}^{\ell} \mathrm{u}_i^\ell\psi_i(x) \sum_{j=1}^{\ell} \mathrm{p}_j^\ell\psi_j(x)\,\mathrm{d}x \\ &= \sum_{i=1}^{\ell}\sum_{j=1}^{\ell} \mathrm{u}_i^\ell \int_\Omega \psi_i(x)\psi_j(x)\,\mathrm{d}x\,\mathrm{p}_j^\ell.\end{aligned}$$

Setting

$$\begin{aligned}\mathrm{M}_{ij}^\ell &= \int_\Omega \psi_i(x)\psi_j(x)\,\mathrm{d}x = \int_\Omega \sum_{l=1}^{n_{FE}}\sum_{k=1}^{n_{FE}} \mathrm{U}_{li}\varphi_l(x)\mathrm{U}_{kj}\varphi_k(x)\,\mathrm{d}x \\ &= \sum_{l=1}^{n_{FE}}\sum_{k=1}^{n_{FE}} \mathrm{U}_{li}\mathrm{U}_{kj}\int_\Omega \varphi_l(x)\varphi_k(x)\,\mathrm{d}x = \sum_{l=1}^{n_{FE}}\sum_{k=1}^{n_{FE}} \mathrm{U}_{li}\mathrm{U}_{kj}\mathrm{M}_{lk}^h = (\mathrm{U}^T\mathrm{M}^h\mathrm{U})_{ij},\end{aligned}$$

we observe that

$$\mathrm{M}^\ell = \mathrm{U}^T \mathrm{M}^h \mathrm{U} \in \mathbb{R}^{\ell \times \ell}$$

holds. Analogously, we compute the matrices

$$\mathrm{BD}^\ell = \mathrm{U}^T \mathrm{BD}^h \mathrm{U} \in \mathbb{R}^{\ell \times \ell},$$

$$\hat{\mathrm{S}}^\ell = \mathrm{U}^T \hat{\mathrm{S}}^h \mathrm{U} \in \mathbb{R}^{\ell \times \ell},$$

$$\mathrm{BE}^\ell = \mathrm{U}^T \mathrm{BE}^h \mathrm{U} \in \mathbb{R}^{\ell \times \ell},$$

as well as the vectors

$$\mathrm{f}^\ell = \mathrm{U}^T \mathrm{f}^h \in \mathbb{R}^\ell$$

and

$$\mathrm{g}^\ell = \mathrm{U}^T \mathrm{g}^h \in \mathbb{R}^\ell.$$

Then, the POD-discretization of the optimality conditions (4.12)–(4.14) leads to

$$\kappa(q - q_d) + (\mathrm{u}^\ell)^T \mathrm{M}^\ell \mathrm{p}^\ell - \lambda = 0, \tag{4.33}$$

$$\alpha \mathrm{BD}^\ell (\mathrm{u}^\ell - \mathrm{u}_\Omega^\ell) + (c\,\hat{\mathrm{S}}^\ell + (\mathrm{BE}^\ell)^T + q\mathrm{M}^\ell + \sigma \mathrm{BD}^\ell)\mathrm{p}^\ell = 0, \tag{4.34}$$

and

$$(c\hat{\mathrm{S}}^\ell + \mathrm{BE}^\ell + q\mathrm{M}^\ell + \sigma \mathrm{BD}^\ell)\mathrm{u}^\ell - \mathrm{f}^\ell - \mathrm{g}^\ell = 0. \tag{4.35}$$

Note that these optimality conditions are generally of a smaller dimension ($\ell \ll n_{FE}$) than those in (4.30)–(4.32).

Remark 4.20 *The symmetric matrices* M^h, $\hat{\mathrm{S}}^h$, BD^h, *and the generally non-symmetric matrix* BE^h *can be computed with a software package like* FEMLAB. *The corresponding POD matrices* M^ℓ, $\hat{\mathrm{S}}^\ell$, BD^ℓ, *and* BE^ℓ *can be derived as described above. The same holds for the vectors* f^h, g^h, *as well as* f^ℓ, g^ℓ.

The discretization of the Hessian is done analogously as for the first-order optimality conditions. Eventually we obtain the Hessian

$$\begin{pmatrix} \kappa & \mathrm{p}^T\mathrm{M} & \mathrm{u}^T\mathrm{M} \\ \mathrm{M}^T\mathrm{p} & \alpha\mathrm{BD} & (c\hat{\mathrm{S}} + \mathrm{BE} + q\mathrm{M} + \sigma\mathrm{BD})^T \\ \mathrm{M}^T\mathrm{u} & c\hat{\mathrm{S}} + \mathrm{BE} + q\mathrm{M} + \sigma\mathrm{BD} & 0 \end{pmatrix},$$

where the matrices M, $\hat{\mathrm{S}}$, BD, and BE and the vectors u and p stand for those matrices and vectors that arise in the discretization either regarding the FE basis or the POD basis. They should be replaced by M^h, $\hat{\mathrm{S}}^h$, BD^h, BE^h, u^h, and p^h if we worked with the finite element basis, and by M^ℓ, $\hat{\mathrm{S}}^\ell$, BD^ℓ, BE^ℓ, u^ℓ, and p^ℓ if we settled for the POD basis.

4.5 Numerical results in parameter estimation

We apply Algorithm 4.1 introduced in Section 4.3 in order to solve the parameter estimation problem. The FE-discretized problem is implemented numerically in MATLAB 6.1, again, using routines from FEMLAB 2.2. In the programming code we can choose if we want to make use of the POD method for calculating the solution or to settle for the FE method as introduced in Section 2.3. We will show that the usage of the POD method is indeed reasonable in the problem we investigate in this work.

Because in our specific problem we usually do not have any given a-priori estimation for q^* (the parameter to be estimated) we should use a rather small value for κ. What we really care about is that the solution u^* is really close to the data u_Γ on the boundary, so we should give a relatively large value to the weight α.

The linear system in step (3) of Algorithm 4.3.1 can be solved by a LU-factorization for the POD-discretized problem, if ℓ is sufficiently small. Since the eigenvalues decay quite rapidly in the problems we investigated in this work, a small value for ℓ already guarantees a good approximation of the state u (compare Section 3.4). For the FE-discretized problem, we need to apply the GMRES method because the linear system is of a much bigger dimension and the LU-factorization alone would take much computing time. As preconditioners for the Hessian we use an incomplete LU-factorization.

4.5.1 Parameter estimation with the POD method and the FE method

Run 4.1 *Let the setting be as in Run 3.1. Our aim is to estimate the ideal coefficient* $q_{ideal} = 25$ *from a measurement* u_d *that is given by the boundary values of the FE solution* $u^h_{ideal} = u^h(q_{ideal})$ *to* (3.21), *i.e.,* $u_\Gamma = u_d = u^h_{ideal}|_\Gamma$.

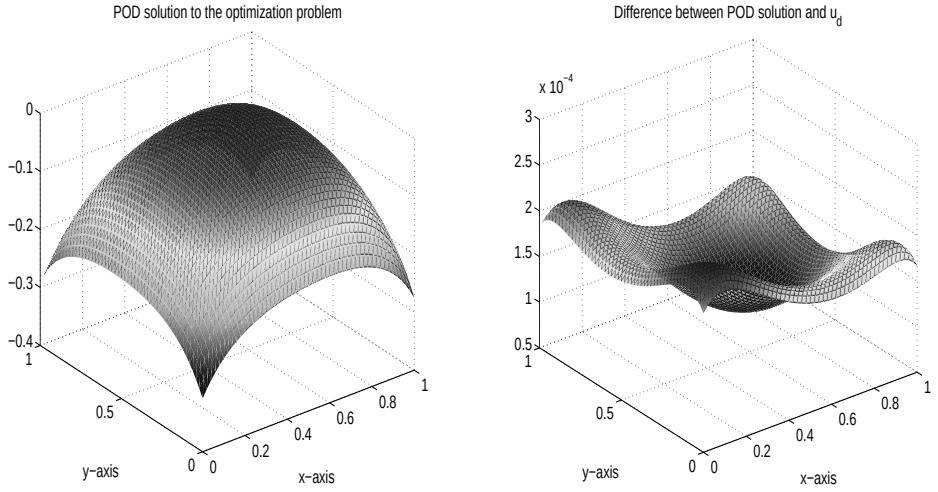

Figure 4.1: Run 4.1: Solution u^* to our augmented SQP algorithm (left plot) and error between the solution and the exact data (right plot).

Let $q_a = 1$ *be the lower bound, i.e.,* $Q_{ad} = \{q \in \mathbb{R} : q \geq 1\}$*, and the regularization parameters be* $\kappa = 0.001$ *and* $\alpha = 10000$ *in* (4.2)*. Moreover, we choose* $q_d = q_a$ *as our initial guess, since we presume that we do not have any a-priori knowledge of the optimal parameter* q_{ideal}*.*
We initialize Algorithm 4.1 with $q^0 = q_d$ *and* $u^0 = u_d$*.*
The resulting optimal POD state $u^\ell := u^*$*, which is given by the solution of Algorithm 4.1 when we are using the POD approximation, is presented in Figure 4.1 (left plot). In the right plot of Figure 4.1 we see that* u^ℓ *is close to the exact data* u^h_{ideal}*. Numerically, we observe*

$$\frac{\|u^h_{ideal} - u^\ell\|_{L^2(\Omega)}}{\|u^h_{ideal}\|_{L^2(\Omega)}} \approx 0.0014 \text{ and } \frac{\|u^h_{ideal} - u^\ell\|_{H^1(\Omega)}}{\|u^h_{ideal}\|_{H^1(\Omega)}} \approx 4.82 \cdot 10^{-4}.$$

Moreover, we obtain $q^* = 24.9493$ *as the solution to our parameter identification problem. Thus, we calculate a relative error of*

$$\frac{|q^* - q_{ideal}|}{|q_{ideal}|} \approx 0.002 = 0.2\%.$$

The performance of Algorithm 4.1 using Algorithm 4.2 as the inner iteration method is shown in Table 4.1. In the first outer iteration we need 6 SQP-iterations – in the first SQP-iteration the step size is decreased to $s = 1/16$*,*

and in the second SQP-iteration coercivity does not hold, so the damping parameter γ is reduced to 0.3677. *The needed computing time of the SQP*

Outer iteration	SQP method	q^k
$k = 1$	6 iterations	24.9489971
$k = 2$	3 iterations	24.9492551
$k = 3$	2 iterations	24.9492513

Table 4.1: Run 4.1: Performance of Algorithm 4.1 with Algorithm 4.2 as the inner iteration method for the POD-discretized problem.

solver is only 1.27 *seconds.*

Next we apply the FE approximation of Algorithm 4.1 to the same problem. The performance of the algorithm is shown in Table 4.2. Since the discrete

Outer iteration	SQP method	q^k
$k = 1$	7 iterations	24.9497474
$k = 2$	2 iterations	24.9497847

Table 4.2: Run 4.1: Performance of Algorithm 4.1 with Algorithm 4.2 as the inner iteration method for the FE-discretized problem.

optimal control problem is of a much higher dimension than in the previous case with POD approximations, the computation of the solution now takes 151.65 *seconds. We obtain a solution of* $q^* = 24.9498$, *thus we observe the following relative errors (by setting* $u^h := u^*$*):*

$$\frac{\|u^h_{ideal} - u^h\|_{L^2(\Omega)}}{\|u^h_{ideal}\|_{L^2(\Omega)}} \approx 0.0014, \quad \frac{\|u^h_{ideal} - u^h\|_{H^1(\Omega)}}{\|u^h_{ideal}\|_{H^1(\Omega)}} \approx 4.77 \cdot 10^{-4}$$

and

$$\frac{|q^* - q_{ideal}|}{|q_{ideal}|} \approx 0.002 = 0.2\%.$$

We test the algorithm on another example, based on the setting of Run 3.3 in Section 3.5.1.

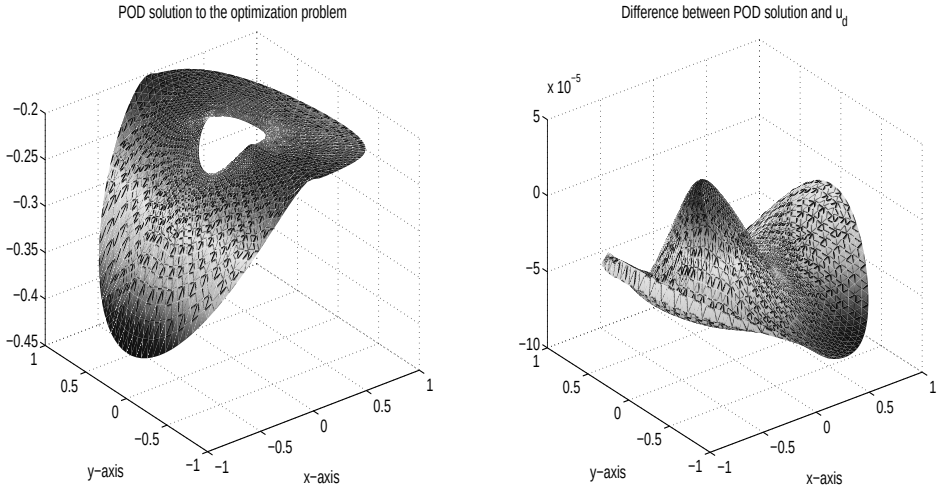

Figure 4.2: Run 4.2: Solution u^* to our augmented SQP algorithm (left plot) and error between the solution and the exact data (right plot).

Run 4.2 *Again we aim to estimate the ideal coefficient, which is now given by* $q_{ideal} = 2$*, from a measurement* u_d *that is given by the boundary values of the FE solution* $u^h_{ideal} = u^h(q_{ideal})$*.*
Let $q_a = 1$ *be the lower bound and the regularization parameters be* $\kappa = 0.001$ *and* $\alpha = 10000$ *in* (4.2)*, as in the previous run. Moreover, we choose* $q_d = q_a$ *as our initial guess and initialize Algorithm 4.1 with* $q^0 = q_d$ *and* $u^0 = u_d$*.*

The optimal POD state and the difference between the POD state and exact data are shown in Figure 4.2.
The relative errors are

$$\frac{\|u^h_{ideal} - u^\ell\|_{L^2(\Omega)}}{\|u^h_{ideal}\|_{L^2(\Omega)}} \approx 1.93 \cdot 10^{-4}, \ \frac{\|u^h_{ideal} - u^\ell\|_{H^1(\Omega)}}{\|u^h_{ideal}\|_{H^1(\Omega)}} \approx 4.15 \cdot 10^{-4},$$

and

$$\frac{|q^* - q_{ideal}|}{|q_{ideal}|} \approx 2.49 \cdot 10^{-4}.$$

The performance of Algorithm 4.1 is shown in Table 4.3. The computation of the solution takes 1.81 *seconds.*

Outer iteration	SQP method	q^k
$k = 1$	11 iterations	2.0005028
$k = 2$	3 iterations	2.0004985
$k = 3$	3 iterations	2.0004981

Table 4.3: Run 4.2: Performance of Algorithm 4.1 with Algorithm 4.2 as the inner iteration method for the POD-discretized problem.

Again we apply the FE-discretized SQP solver to the same problem and obtain – after a computation time of 240.16 *seconds and only one augmented Lagrange iteration (with* 14 *SQP iterations) – a solution of* $q^* = 2.00001$. *Thus, the relative error is only*

$$\frac{|q^* - q_{ideal}|}{|q_{ideal}|} \approx 5.1 \cdot 10^{-6}.$$

Moreover, we observe

$$\frac{\|u^h_{ideal} - u^h\|_{L^2(\Omega)}}{\|u^h_{ideal}\|_{L^2(\Omega)}} \approx 1.38 \cdot 10^{-6} \text{ and } \frac{\|u^h_{ideal} - u^h\|_{H^1(\Omega)}}{\|u^h_{ideal}\|_{H^1(\Omega)}} \approx 5.93 \cdot 10^{-6}.$$

4.5.2 Parameter estimation for noisy data with the POD method and the FE method

As we mentioned in the beginning of Section 4, in many practical applications one gets measurements which are corrupted due to measurement noise. In this section we investigate the case where we do not have an exact representation for the data, but (little) perturbed measurements. In this work we admit random noise in the size of 5% to the exact data u^h_{ideal}.

Run 4.3 *The setting is the same as in Run 4.1, except that the noisy data is given by* $u_\Gamma = u_d = (1+0.05\varepsilon)u^h_{ideal}\big|_\Gamma$, *where* $\varepsilon : \Omega \to [-1, 1]$ *is a mapping with random function values. Nevertheless, we obtain a solution for* q^* *that is close to* q_{ideal} *for any random variable. For one possible random noise vector we present our results here. The relative errors are*

$$\frac{\|u^h_{ideal} - u^\ell\|_{L^2(\Omega)}}{\|u^h_{ideal}\|_{L^2(\Omega)}} \approx 0.0036, \quad \frac{\|u^h_{ideal} - u^\ell\|_{H^1(\Omega)}}{\|u^h_{ideal}\|_{H^1(\Omega)}} \approx 0.0012,$$

$$\frac{\|u_d - u^\ell\|_{L^2(\Omega)}}{\|u_d\|_{L^2(\Omega)}} \approx 0.0207, \ \frac{\|u_d - u^\ell\|_{H^1(\Omega)}}{\|u_d\|_{H^1(\Omega)}} \approx 0.5117,$$

and

$$\frac{|q^* - q_{ideal}|}{|q_{ideal}|} \approx 0.0051 = 0.51\%.$$

Note that the function u^ℓ, which is part of the solution to our optimization problem, is in fact closer to u^h_{ideal} than our initial (noisy) data u_d, which has a relative error to u^h_{ideal} of 0.0204 in the L^2-norm.
The relative error between the solution and the noisy data in the H^1-norm is large, of course, due to measurement noise. The performance of the algorithm, which takes only 0.42 seconds of computing time, is presented in Table 4.4.

Outer iteration	SQP method	q^k
$k = 1$	6 iterations	24.8723549
$k = 2$	3 iterations	24.8725878
$k = 3$	2 iterations	24.8725841

Table 4.4: Run 4.3: Performance of Algorithm 4.1 with Algorithm 4.2 as the inner iteration method for the POD-discretized problem.

In order to possibly improve the results in our parameter identification problem, one option is to use the a-priori solution q^* from above and restart Algorithm 4.1 with $q_d = q^*$ and $q^0 = q^*$. Moreover, we reduce the regularization κ to 0.0001. We hope to find a solution q^{**} in the second cycle that is closer to q_{ideal}.
Setting $q^0 = 24.8726$ and rerunning the algorithm, we obtain an optimal parameter $q^{**} = 24.9227$. Thus, the new error of the estimated parameter is only

$$\frac{|q^* - q_{ideal}|}{|q_{ideal}|} \approx 0.0031 = 0.31\%.$$

The relative errors of the new optimal state $u^\ell := u^{**}$ are now

$$\frac{\|u^h_{ideal} - u^\ell\|_{L^2(\Omega)}}{\|u^h_{ideal}\|_{L^2(\Omega)}} \approx 0.0022, \ \frac{\|u^h_{ideal} - u^\ell\|_{H^1(\Omega)}}{\|u^h_{ideal}\|_{H^1(\Omega)}} \approx 7.3345 \cdot 10^{-4},$$

and

$$\frac{\|u_d - u^\ell\|_{L^2(\Omega)}}{\|u_d\|_{L^2(\Omega)}} \approx 0.0205, \ \frac{\|u_d - u^\ell\|_{H^1(\Omega)}}{\|u_d\|_{H^1(\Omega)}} \approx 0.5117.$$

If we repeat this procedure and set $q_d = q^{**}$ *and* $q^0 = q^{**}$*, the results become hardly better. For instance, we obtain* $q^{***} = 24.9228$ *and, for* $u^\ell := u^{***}$*,*

$$\frac{\|u^h_{ideal} - u^\ell\|_{H^1(\Omega)}}{\|u^h_{ideal}\|_{H^1(\Omega)}} \approx 7.3335 \cdot 10^{-4},$$

in the third cycle.

Again, the FE-discretization of Algorithm 4.1 is tested for the same problem with noisy data. After a computing time of 170.8 *seconds we obtain* $q^* = 24.8516$*, thus*

$$\frac{|q^* - q_{ideal}|}{|q_{ideal}|} \approx 0.0059 = 0.59\%.$$

Moreover, setting $u^h := u^*$*, we observe*

$$\frac{\|u^h_{ideal} - u^h\|_{L^2(\Omega)}}{\|u^h_{ideal}\|_{L^2(\Omega)}} \approx 0.0042, \ \frac{\|u^h_{ideal} - u^h\|_{H^1(\Omega)}}{\|u^h_{ideal}\|_{H^1(\Omega)}} \approx 0.0014,$$

and

$$\frac{\|u_d - u^h\|_{L^2(\Omega)}}{\|u_d\|_{L^2(\Omega)}} \approx 0.0208, \ \frac{\|u_d - u^h\|_{H^1(\Omega)}}{\|u_d\|_{H^1(\Omega)}} \approx 0.5117.$$

When we are restarting Algorithm 4.1 with $q_d = q^*$ *and* $q^0 = q^*$*, we obtain* $q^{**} = 24.901$ *after* 93.24 *seconds of computing time, and for* $u^h := u^{**}$ *we observe*

$$\frac{\|u^h_{ideal} - u^h\|_{L^2(\Omega)}}{\|u^h_{ideal}\|_{L^2(\Omega)}} \approx 0.0028, \ \frac{\|u^h_{ideal} - u^h\|_{H^1(\Omega)}}{\|u^h_{ideal}\|_{H^1(\Omega)}} \approx 9.3737 \cdot 10^{-4},$$

$$\frac{\|u_d - u^h\|_{L^2(\Omega)}}{\|u_d\|_{L^2(\Omega)}} \approx 0.0206, \ \frac{\|u_d - u^h\|_{H^1(\Omega)}}{\|u_d\|_{H^1(\Omega)}} \approx 0.5117,$$

and

$$\frac{|q^{**} - q_{ideal}|}{|q_{ideal}|} \approx 0.004 = 0.4\%.$$

As in the POD discretized algorithm, in a third cycle we hardly observe better results.

A Appendix

A.1 Basic linear algebra

Definition A.1 *We define the Euclidian scalar product* $\langle \cdot, \cdot \rangle_{\mathbb{R}^m}$ *by*

$$\langle x, y \rangle_{\mathbb{R}^m} = \sum_{i=1}^{m} x_i y_i$$

and the respective induced norm $\| \cdot \|_{\mathbb{R}^m}$ *by*

$$\|x\|_{\mathbb{R}^m} = \sqrt{\langle x, x \rangle_{\mathbb{R}^m}} = \sqrt{\sum_{i=1}^{m} x_i^2}.$$

We often write simply $x \cdot y$ *instead of* $\langle x, y \rangle_{\mathbb{R}^m}$ *when it is clear that* x *and* y *are vectors of the dimension* m*, for any* $m \in \mathbb{R}$*.*

Lemma A.2 (Young's inequality) *For all* $a, b \in \mathbb{R}$ *and every* $\epsilon > 0$

$$ab \leq \epsilon a^2 + \frac{b^2}{4\epsilon}$$

holds.

Definition A.3 (Frobenius norm) *The so called Frobenius norm* $\| \cdot \|_F$ *of a matrix* $A \in \mathbb{R}^{m \times n}$ *is defined by*

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}|^2}.$$

Lemma A.4 *Let us introduce the trace of a matrix* $M \in \mathbb{R}^{n \times n}$ *by* $\operatorname{trace}(M) = \sum_{j=1}^{n} M_{jj}$*. Then the equality*

$$\|A\|_F = \sqrt{\operatorname{trace}(A^T A)} \tag{A.1}$$

holds for any $A \in \mathbb{R}^{m \times n}$*.*

Proof. We have

$$\text{trace}(A^TA) = \sum_{j=1}^{n}(A^TA)_{jj} = \sum_{j=1}^{n}\sum_{i=1}^{m}(A^T)_{ji}A_{ij} = \sum_{j=1}^{n}\sum_{i=1}^{m}A_{ij}A_{ij} = \sum_{j=1}^{n}\sum_{i=1}^{m}A_{ij}^2,$$

and therefore

$$\sqrt{\text{trace}(A^TA)} = \sqrt{\sum_{j=1}^{n}\sum_{i=1}^{m}A_{ij}^2} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}A_{ij}^2} = \|A\|_F$$

holds. ■

Corollary A.5 *Let $U \in \mathbb{R}^{m\times m}$ be an orthonormal matrix, that means that $U^TU = I$ holds, where I denotes the identity matrix. Then $\|A\|_F = \|UA\|_F$ holds for all matrices $A \in \mathbb{R}^{m\times n}$.*

Proof. By applying (A.1) twice we have

$$\|UA\|_F^2 = \text{trace}((UA)^T(UA)) = \text{trace}(A^TU^TUA) = \text{trace}(A^TA) = \|A\|_F^2,$$

thus the assertion holds. ■

Lemma A.6 *For a symmetric positive definite matrix $M \in \mathbb{R}^{m\times m}$ we have the eigenvalue decomposition*

$$M = QDQ^T,$$

where $D = \text{diag}(\eta_1, ..., \eta_m) \in \mathbb{R}^{m\times m}$ with $\eta_i > 0$ for all $i \in \{1, ..., m\}$. Moreover, $Q \in \mathbb{R}^{m\times m}$ is orthogonal.

Definition A.7 *For any $\alpha \in \mathbb{R}$ we define*

$$M^\alpha = Q\text{diag}(\eta_1^\alpha, ..., \eta_m^\alpha)Q^T,$$

where the matrix $Q \in \mathbb{R}^{m\times m}$ as well as the positive values η_i, $i = 1, ..., m$, are introduced in Lemma A.6.

Lemma A.8 *The following properties hold:*

$$(M^\alpha)^{-1} = M^{-\alpha},$$

$$M^{\alpha+\beta} = M^\alpha M^\beta \text{ for all } \alpha, \beta \in \mathbb{R},$$

$$(M^\alpha)^T = M^\alpha.$$

Setting $\alpha = \frac{1}{2}$, we have

$$M^{\frac{1}{2}} = Q\text{diag}(\sqrt{\eta_1}, ..., \sqrt{\eta_m})Q^T.$$

Definition A.9 *The Kronecker symbol δ_{ik} is defined by $\delta_{ik} = 1$ if $i = k$ and $\delta_{ik} = 0$ if $i \neq k$.*

A.2 Basic functional analysis

Definition A.10 *A collection $\mathcal{M}$ of subsets of $\mathbb{R}^d$ is called a σ-algebra if*

$$\begin{aligned}
&\emptyset, \mathbb{R}^d \in \mathcal{M}, \\
&A \in \mathcal{M} \text{ implies } (\mathbb{R}^d - A) \in \mathcal{M}, \text{ and} \\
&\text{if } \{A_k\}_{k=1}^{\infty} \subset \mathcal{M}, \text{ then } \bigcup_{k=1}^{\infty} A_k, \bigcap_{k=1}^{\infty} A_k \in \mathcal{M}.
\end{aligned}$$

Definition A.11 *Let $f : \mathbb{R}^d \to \mathbb{R}$. We say f is a measurable function if*

$$f^{-1}(U) \in \mathcal{M}$$

for each open subset $U \subset \mathbb{R}$.

Definition A.12 *The adjoint operator $\mathcal{A}^* : H \to V$ to a given linear and bounded operator $\mathcal{A} : V \to H$ is defined as*

$$\langle \mathcal{A}u, v\rangle_H = \langle u, \mathcal{A}^* v\rangle_V$$

for any $u \in V$ and $v \in H$.

Lemma A.13 (Cauchy-Schwarz inequality) *Let V be a Hilbert space. Then*

$$\langle u, v\rangle_V \leq \|u\|_V \|v\|_V$$

holds for all $u, v \in V$.

Lemma A.14 (Trace theorem) *Assume that Ω is bounded and $\Gamma = \partial\Omega$ is C^1. Then there exists a bounded linear operator*

$$\tau_\Gamma : H^1(\Omega) \to L^2(\Gamma)$$

such that

$$\tau_\Gamma u = u|_\Gamma$$

and

$$\|\tau_\Gamma u\|_{L^2(\Gamma)} \leq C_\Gamma \|u\|_{H^1(\Omega)}$$

for each $u \in H^1(\Omega)$, with a constant $C_\Gamma > 0$, named the trace constant, depending on Ω.

A.3 Optimization theory

Consider a finite-dimensional optimization problem of the form

$$\min J(x) \text{ for } x \in \mathbb{R}^n \quad \text{s.t. } e(x) = 0 \in \mathbb{R}^m. \tag{A.2}$$

A solution to (A.2) is called an optimal solution to the optimization problem.

Definition A.15 (Feasible points) *The set of feasible points is given by*

$$\mathcal{F} = \{x \in \mathbb{R}^n : e(x) = 0\}.$$

Definition A.16 (Partial derivative) *Let $x \in X^n$ and $y \in Y^m$ with $n, m \in \mathbb{N}$. The term $\nabla_x J(x, y)$ denotes the vector containing the entries $\frac{\partial J(x,y)}{\partial x_i}$ for $1 \leq i \leq n$, only.*

Definition A.17 (Hessian matrix) *The matrix of the second partial derivatives of a function $J : X^n \to \mathbb{R}$ containing the entries $\frac{\partial^2 J(x)}{\partial x_i \partial x_j}$ at $(\nabla^2 J(x))_{ij}$ for $1 \leq i, j \leq n$, is called the Hessian matrix and denoted by $\nabla^2 J(x)$.*

Again, if $x \in X^n$ and $y \in Y^m$ with $n, m \in \mathbb{N}$, the matrix $\nabla^2_{xx} J(x, y)$ contains the entries $\frac{\partial^2 J(x,y)}{\partial x_i \partial x_j}$ for $1 \leq i, j \leq n$, only.

Definition A.18 (Regular points) *The point $x^* \in \mathcal{F}$ is a regular point if the (weak) gradients $\nabla e_1(x^*), ..., \nabla e_m(x^*)$ are linearly independent.*

Thus, x^* is regular if the matrix

$$\nabla e(x^*) = \begin{pmatrix} \nabla e_1(x^*)^T \\ \vdots \\ \nabla e_m(x^*)^T \end{pmatrix}$$

is surjective.

Definition A.19 (Local solution) *The point $x^* \in \mathcal{F}$ is a (strict) local solution to* (A.2) *if and only if*

$$J(x) \geq J(x^*) \quad (J(x) > J(x^*)) \text{ for all } x \in (U \cap \mathcal{F}) \backslash \{x^*\},$$

where U is a neighborhood of x^*.

Theorem A.20 (First-order necessary optimality condition) *Suppose that $x^* \in \mathbb{R}^n$ is a local solution to* (A.2) *and let x^* be regular. Then there exists a unique vector of so called Lagrange-multipliers $\lambda^* = (\lambda_1^*, \ldots, \lambda_m^*)^T$ satisfying the first-order necessary optimality conditions (KKT conditions)*

$$\nabla J(x^*) + \nabla e(x^*)^T \lambda^* = 0 \in \mathbb{R}^n$$

and

$$e(x^*) = 0 \in \mathbb{R}^m.$$

Introducing the Lagrange function $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ by

$$\mathcal{L}(x, \lambda) = J(x) + e(x)^T \lambda = J(x) + \langle e(x), \lambda \rangle_{\mathbb{R}^m} \text{ for } (x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^m,$$

we observe that the first-order necessary optimality conditions for (A.2) can be written as

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = 0 \text{ and } \nabla_\lambda \mathcal{L}(x^*, \lambda^*) = 0.$$

Theorem A.21 (Second-order sufficient optimality condition) *Suppose that $x^* \in \mathcal{F}$ is a regular point and let x^* together with $\lambda^* \in \mathbb{R}^m$ satisfy*

$$\nabla J(x^*) + \nabla e(x^*)^T \lambda^* = 0 \in \mathbb{R}^n$$

and

$$e(x^*) = 0 \in \mathbb{R}^m.$$

Moreover, let the matrix

$$\nabla_{xx} \mathcal{L}(x^*, \lambda^*) = \nabla_{xx} J(x^*, \lambda^*) + \nabla_{xx} e(x^*)^T \lambda^* = \nabla_{xx} J(x^*, \lambda^*) + \sum_{i=1}^{m} \lambda_i^* \nabla_{xx} e_i(x^*)$$

be positive definite on $\mathrm{Ker}(\nabla e(x^*))$*, i.e., $v^T \nabla_{xx} \mathcal{L}(x^*, \lambda^*) v > 0$ for all $v \in \mathbb{R}^n$ which satisfy $\nabla e(x^*) v = 0$. Then x^* is a strict local solution to* (A.2).

References

[1] A. C. Antoulas, D. C Sorensen, and S. Gugercin. A survey of model reduction methods for large-scale systems. *Structured matrices in mathematics, computer science, and engineering I.* Proceedings of an AMS-IMS-SIAM joint summer research conference, University of Colorado, Boulder, CO, USA, June 27-July 1, 1999, Providence, RI: American Mathematical Society (AMS). Contemp. Math. 280, 193-219, 2001.

[2] J. A. Atwell and B. B. King. Reduced order controllers for spatially distributed systems via proper orthogonal decomposition. *SIAM Journal on Scientific Computing*, 26:128-151, 2004.

[3] H. T. Banks, M. L. Joyner, B. Winchesky, and W. P. Winfree. Nondestructive evaluation using a reduced-order computational methodology. *Inverse Problems*, 16:1-17, 2000.

[4] H. T. Banks, R. C. H. del Rosario, and R. C. Smith. Reduced order model feedback control design: numerical implementation in a thin shell model. Technical Report CRSC-TR98-27, North Carolina State University, 1998.

[5] M. Barrault, Y. Maday, N. C. Nguyen, and A. T. Patera. An empirical interpolation method: application to efficient reduced-basis discretization of partial differential equations. *Comptes Rendus de'l Académie des Sciences Paris*, Ser. I 339:667-672, 2004.

[6] A. Battermann. Preconditioning of Karush-Kuhn-Tucker Systems arising in optimal control problems. Diploma Thesis, Blacksburg, Virginia, 1996.

[7] A. Battermann and M. Heinkenschloss. Preconditioners for Karush-Kuhn-Tucker matrices arising in the optimal control of distributed systems. *Fast Solution Methods for Discretized Optimization Problems*, K.H. Hoffmann, R.H.W. Hoppe, V. Schulz, eds. Birkhäuser Verlag, Basel, 2001, pp. 15-32.

[8] A. Battermann and E. W. Sachs. Block preconditioners for KKT Systems in PDE-governed optimal control problems. *Fast Solution Methods for Discretized Optimization Problems*, K.H. Hoffmann, R.H.W. Hoppe, V. Schulz, eds. Birkhäuser Verlag, Basel, 2001, pp. 1-18.

[9] D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, Mass., 1995.

[10] D. Braess. *Finite Elemente.* Springer Verlag, Berlin, 1997.

[11] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Elements.* Springer Verlag, Berlin, 1994.

[12] R. Dautray and J. L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology*, volume 2. Springer Verlag, Berlin, 1988.

[13] P. Deuflhard and A. Hohmann. *Numerische Mathematik I.* Walter de Gruyter, Berlin, 2002.

[14] L. C. Evans. *Partial Differential Equations.* American Mathematical Society, Providence, R.I., 1999.

[15] K. Fukuda. *Introduction to Statistical Recognition.* Academic Press, New York, 1990.

[16] T. Gänzler, S. Volkwein, and M. Weiser. SQP methods for parameter identification problems arising in hyperthermia. To appear in *Optimization Methods and Software, Special Issue on Parameter Estimation and Experimental Design*, 2006.

[17] H. Goering, H. G. Roos, and L. Tobiska. *Finite-Element-Methode.* Akademie Verlag, Berlin, 1993.

[18] M. Hintermüller. On a globalized augmented Lagrangian-SQP algorithm for nonlinear optimal control problems with box constraints. *Fast Solution Methods for Discretized Optimization Problems*, K.H. Hoffmann, R.H.W. Hoppe, V. Schulz, eds. Birkhäuser Verlag, Basel, 2001, pp. 139-153.

[19] M. Hinze and S. Volkwein. Error estimates for abstract linear-quadratic optimal control problems using proper orthogonal decomposition. University of Graz, Institute for Mathematics and Scientific Computing, Report No. 2/2005, submitted.

[20] M. Hinze and S. Volkwein. Proper Orthogonal Decomposition surrogate models for nonlinear dynamical systems: error estimates and suboptimal control. *In Reduction of Large-Scale Systems*, P. Benner, V. Mehrmann, D. C. Sorensen (eds.), Lecture Notes in Computational Science and Engineering, Vol. 45, 261-306, 2005.

[21] P. Holmes, J. L. Lumley, and G. Berkooz. *Turbulence, coherent structures, dynamical systems and symmetry.* Cambridge University Press, 1996.

[22] K. Ito and K. Kunisch. Augmented Lagrangian-SQP methods in Hilbert spaces and application to control in the coefficient problems. *SIAM Journal on Optimization*, 6:96-125, 1996.

[23] K. Ito and K. Kunisch. Augmented Lagrangian-SQP methods for nonlinear optimal control problems of tracking type. *SIAM Journal on Optimization*, 34:874-891, 1996.

[24] T. Kato. *Perturbation Theory.* Springer Verlag, Berlin, 1984.

[25] M. Kahlbacher and S. Volkwein. Galerkin proper orthogonal decomposition methods for parameter dependent elliptic systems. University of Graz, Institute for Mathematics and Scientific Computing, Report No. 6/2006, submitted.

[26] M. Kahlbacher and S. Volkwein. Model reduction by proper orthogonal decomposition for estimation of scalar parameters in elliptic PDEs. *Proceedings of European Conference on Computational Fluid Dynamics (ECCOMAS CFD), P. Wesseling, E. Onate, and J. Periaux (eds.)*, Egmont aan Zee, 2006.

[27] K. Kunisch and S. Volkwein. Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics. *SIAM Journal on Numerical Analysis*, 40:492-515, 2002.

[28] K. Kunisch and S. Volkwein. Augmented Lagrangian-SQP techniques and their approximations. *Optimization methods in partial differential equations, Contemporary Mathematics*, 209:147-159, 1997.

[29] K. Kunisch and S. Volkwein. Galerkin proper orthogonal decomposition methods for parabolic problems. *Numerische Mathematik*, 90:117-148, 2001.

[30] K. Kunisch and S. Volkwein. Control of Burgers' equation by a reduced order approach using proper orthogonal decomposition. *Journal of Optimization Theory and Applications*, 102:345-371, 1999.

[31] K. Kunisch and S. Volkwein. Proper orthogonal decomposition for optimality systems. University of Graz, Institute for Mathematics and Scientific Computing, Report No. 10/2006, submitted.

[32] H. V. Ly and H. T. Tran. Modelling and control of physical processes using proper orthogonal decomposition. *Mathematical and Computer Modeling*, 33:223-236, 2001.

[33] L. Machiels, Y. Maday, and A. T. Patera. Output bounds for reduced-order approximations of elliptic partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 190:3413-3426, 2001.

[34] Y. Maday and E. M. Rønquist. A reduced-basis element method. *Journal of Scientific Computing*, 17, 1-4, 2002.

[35] H. Maurer and J. Zowe. First and second order necessary and sufficient optimality conditions for infinite-dimensional programming problems. *Mathematical Programming*, 16, 98–110, 1979.

[36] H. Müller and S. Volkwein. Model reduction by proper orthogonal decomposition for lambda-omega systems. *Proceedings of European Conference on Computational Fluid Dynamics (ECCOMAS CFD), P. Wesseling, E. Onate, and J. Periaux (eds.)*, Egmont aan Zee, 2006.

[37] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Verlag, Berlin, 1999.

[38] L. Qi and F. Sun. A nonsmooth version of Newton's method. *Mathematical Programming*, 58:353-367, 1993.

[39] M. Reed and B. Simon. *Methods of Modern Mathematical Physics. I: Functional Analysis*. Academic Press, San Diego, CA, 1980.

[40] E. W. Sachs and S. Volkwein. Augmented Lagrangian-SQP methods with Lipschitz-continuous Lagrange multiplier updates. *SIAM Journal on Numerical Analysis*, 40:233-353, 2002.

[41] L. Sirovich. Turbulence and the dynamics of coherent structures, parts I-III. *Quarterly of Applied Mathematics*, XLV:561-590, 1987.

[42] S. Volkwein. A globalized SQP method for the optimal control of laser surface hardening. University of Graz, Institute for Mathematics and Scientific Computing, SFB-Preprint No. 272, June 2003.

[43] S. Volkwein. Basic functional analysis for the optimization of partial differential equations. Script, January 2003. http://www.uni-graz.at/imawww/volkwein/preliminaries.pdf

[44] S. Volkwein. Optimal control of a phase-field model using proper orthogonal decomposition. *Z. Angew. Math. Mech.*, 81:83-97, 2001.