Chapter 1

POD for linear-quadratic Optimal Control

M. Gubisch and S. Volkwein[*]

1.1 ▪ Introduction

Optimal control problems for partial differential equation are often hard to tackle numerically because their discretization leads to very large scale optimization problems. Therefore, different techniques of model reduction were developed to approximate these problems by smaller ones that are tractable with less effort.

Balanced truncation [2, 64, 79] is one well studied model reduction technique for state-space systems. This method utilizes the solutions to two Lyapunov equations, the so-called controllability and observability Gramians. The balanced truncation method is based on transforming the state-space system into a balanced form so that its controllability and observability Gramians become diagonal and equal. Moreover, the states that are difficult to reach or to observe, are truncated. The advantage of this method is that it preserves the asymptotic stability in the reduced-order system. Furthermore, a-priori error bounds are available. Recently, the theory of balanced truncation model reduction was extended to descriptor systems; see, e.g., [48] and [21].

Recently the application of *reduced-order models* to linear time varying and nonlinear systems, in particular to nonlinear control systems, has received an increasing amount of attention. The reduced-order approach is based on projecting the dynamical system onto subspaces consisting of basis elements that contain characteristics of the expected solution. This is in contrast to, e.g., finite element techniques (see, e.g., [7], where the basis elements of the subspaces do not relate to the physical properties of the system that they approximate. The *reduced basis* (RB) method, as developed in [19, 54] and [31], is one such reduced-order method, where the basis elements correspond to the dynamics of expected control regimes. Let us refer to the [14, 23, 49, 53] for the successful use of reduced basis method in PDE constrained optimization problems. Currently, *Proper orthogonal decomposition* (POD) is probably the mostly used and most successful model reduction technique for nonlinear optimal control problems, where the basis functions contain in-

[*]The authors gratefully acknowledges support by the German Science Fund DFG grant VO 1658/2-1 *A-posteriori-POD Error Estimators for Nonlinear Optimal Control Problems governed by Partial Differential Equations*. The first author is further supported by the Landesgraduiertenförderung of Baden-Wurttemberg.

formation from the solutions of the dynamical system at pre-specified time-instances, so-called snapshots; see, e.g., [9, 29, 67, 75]. Due to a possible linear dependence or almost linear dependence the snapshots themselves are not appropriate as a basis. Hence a singular value decomposition is carried out and the leading generalized eigenfunctions are chosen as a basis, referred to as the POD basis. POD is successfully used in a variety of fields including fluid dynamics, coherent structures [1, 3] and inverse problems [6]. Moreover in [5] POD is successfully applied to compute reduced-order controllers. The relationship between POD and balancing was considered in [44, 61, 77]. An error analysis for nonlinear dynamical systems in finite dimensions were carried out in [58] and a missing point estimation in models described by POD was studied in [4].

Reduced order models are used in PDE-constrained optimization in various ways; see, e.g., [27, 63] for a survey. In optimal control problems it is sometimes necessary to compute a feedback control law instead of a fixed optimal control. In the implementation of these feedback laws models of reduced-order can play an important and very useful role, see [5, 43, 46, 59]. Another useful application is the use in optimization problems, where a PDE solver is part of the function evaluation. Obviously, thinking of a gradient evaluation or even a step-size rule in the optimization algorithm, an expensive function evaluation leads to an enormous amount of computing time. Here, the reduced-order model can replace the system given by a PDE in the objective function. It is quite common that a PDE can be replaced by a five- or ten-dimensional system of ordinary differential equations. This results computationally in a very fast method for optimization compared to the effort for the computation of a single solution of a PDE. There is a large amount of literature in engineering applications in this regard, we mention only the papers [47, 50]. Recent applications can also be found in finance using the RB model [56] and the POD model [62, 65] in the context of calibration for models in option pricing.

In the present work we use POD for deriving low order models of dynamical systems. These low order models then serve as surrogates for the dynamical system in the optimization process. We consider a linear-quadratic optimal control problem in an abstract setting and prove error estimates for the POD Galerkin approximations of the optimal control. This is achieved by combining techniques from [11, 12, 25] and [38, 40]. For nonlinear problems we refer the reader to [27, 55, 63]. However, unless the snapshots are generating a sufficiently rich state space or are computed from the exact (unknown) optimal controls, it is not a-priorly clear how far the optimal solution of the POD problem is from the exact one. On the other hand, the POD method is a universal tool that is applicable also to problems with time-dependent coefficients or to nonlinear equations. Moreover, by generating snapshots from the real (large) model, a space is constructed that inhibits the main and relevant physical properties of the state system. This, and its ease of use makes POD very competitive in practical use, despite of a certain heuristic flavor. In this work, we review results for a POD a-posteriori analysis; see, e.g., [71] and [20, 33, 34, 68, 69, 74, 76]. We use a fairly standard perturbation method to deduce how far the suboptimal control, computed on the basis of the POD model, is from the (unknown) exact one. This idea turned out to be very efficient in our examples. It is able to compensate for the lack of a priori analysis for POD methods. Let us also refer to the papers [13, 18, 49], where a-posteriori error bounds are computed for linear-quadratic optimal control problems approximated by the reduced basis method.

The manuscript is organized in the following manner: In Section 1.2 we introduce the method of POD in real, separable Hilbert spaces and discuss its relationship to the singular value decomposition. We distinguish between two versions of the POD method: the discrete and the continuous one. Reduced-order modelling with POD is carried out in Section 1.3. The error between the exact solution and its POD approximation is investigated

by an a-priori error analysis. In Section 1.4 we study quadratic optimal control problems governed by linear evolution problems and bilateral inequality constraints. These problems are infinite dimensional, convex optimization problems. Their optimal solutions are characterised by first-order optimality conditions. POD Galerkin discretizations of the optimality conditions are introduced and analysed. By an a-priori error analysis the error of the exact optimal control and its POD suboptimal approximation is estimated. For the error control in the numerical realisations we make use of an a-posteriori error analysis, which turns out to be very efficient in our numerical examples, which are presented in Section 1.5.

1.2 ▪ The POD method

Throughout we suppose that X is a real Hilbert space endowed with the inner product $\langle\cdot,\cdot\rangle_X$ and the associated induced norm $\|\cdot\|_X = \langle\cdot,\cdot\rangle_X^{1/2}$. Furthermore, we assume that X is *separable*, i.e., X has a countable dense subset. This implies that X possesses a countable orthonormal basis; see, e.g., [60, p. 47]. For the POD method in complex Hilbert spaces we refer to [73], for instance.

1.2.1 ▪ The discrete variant of the POD method

For fixed $n, \wp \in \mathbb{N}$ let the so-called *snapshots* $y_1^k, \dots, y_n^k \in X$ be given for $1 \le k \le \wp$. To avoid a trivial case we suppose that at least one of the y_j^k's is nonzero. Then, we introduce the finite dimensional, linear subspace

$$\mathcal{V}^n = \operatorname{span}\left\{ y_j^k \,|\, 1 \le j \le n \text{ and } 1 \le k \le \wp \right\} \subset X \tag{1.1}$$

with dimension $d^n \in \{1, \dots, n\wp\} < \infty$. We call the set $\mathcal{V}^n$ *snapshot subspace*.

Remark 1.1. Later we will focus on the following application: Let $0 \le t_1 < t_2 < \dots < t_n \le T$ be a given time grid in the interval $[0, T]$. To simplify of the presentation, the time grid is assumed to be equidistant with step-size $\Delta t = T/(n-1)$, i.e., $t_j = (j-1)\Delta t$. For nonequidistant grids we refer the reader to [40, 39]. Suppose that we have trajectories $y^k \in C([0,T];X)$, $1 \le k \le \wp$. Here, the Banach space $C([0,T];X)$ contains all functions $\varphi : [0,T] \to X$, which are continuous on $[0,T]$; see, e.g., [70, p. 142]. Let the snapshots be given as $y_j^k = y^k(t_j) \in X$ or $y_j^k \approx y^k(t_j) \in X$. In Sections 1.3 and 1.4 we will choose trajectories as solutions to evolution problems. ◊

In Section 1.2.3 we consider the case, where the number n is varied. Therefore, we emphasize this dependence by using the super index n. We distinguish two cases:

1) The separable Hilbert space X has finite dimension m. Then, X is isomorphic to $\mathbb{R}^m$; see, e.g., [60, p. 47]. We set $\mathcal{I} = \{1, \dots, m\}$. Clearly, we have $d^n \le \min(n\wp, m)$.

2) Since X is separable, each orthonormal basis of X has countably many elements. In this case X is isomorphic to the set ℓ_2 of sequences $\{x_i\}_{i\in\mathbb{N}}$ of real numbers which satisfy $\sum_{i=1}^{\infty} |x_i|^2 < \infty$; see [60, p. 47], for instance. Then, we define $\mathcal{I} = \mathbb{N}$.

The method of POD consists in choosing an orthonormal set $\{\psi_i\}_{i=1}^{\ell}$ in X such that for every $\ell \in \{1, \dots, d^n\}$ the mean square error between the $n\wp$ elements y_j^k and their

corresponding ℓ-th partial Fourier sum is minimized on average:

$$\begin{cases} \min \sum_{k=1}^{\wp} \sum_{j=1}^{n} \alpha_j^n \left\| y_j^k - \sum_{i=1}^{\ell} \langle y_j^k, \psi_i \rangle_X \psi_i \right\|_X^2 \\ \text{s.t. } \{\psi_i\}_{i=1}^{\ell} \subset X \text{ and } \langle \psi_i, \psi_j \rangle_X = \delta_{ij},\ 1 \le i, j \le \ell, \end{cases} \tag{$\mathbf{P}_n^\ell$}$$

where the α_j^n's denote positive weighting parameters. Here, the symbol δ_{ij} denotes the Kronecker symbol satisfying $\delta_{ii} = 1$ and $\delta_{ij} = 0$ for $i \neq j$. An optimal solution $\{\bar{\psi}_i^n\}_{i=1}^{\ell}$ to ($\mathbf{P}_n^\ell$) is called a *POD basis* of *rank* ℓ, which can be extended to a complete orthonormal basis $\{\psi_i\}_{i \in \mathscr{I}}$ in the Hilbert space X. Notice that

$$\begin{aligned} &\left\| y_j^k - \sum_{i=1}^{\ell} \langle y_j^k, \psi_i \rangle_X \psi_i \right\|_X^2 \\ &= \left\langle y_j^k - \sum_{i=1}^{\ell} \langle y_j^k, \psi_i \rangle_X \psi_i, y_j^k - \sum_{l=1}^{\ell} \langle y_j^k, \psi_l \rangle_X \psi_l \right\rangle_X \\ &= \|y_j^k\|_X^2 - 2 \sum_{i=1}^{\ell} \langle y_j^k, \psi_i \rangle_X^2 + \sum_{i=1}^{\ell} \sum_{l=1}^{\ell} \langle y_j^k, \psi_i \rangle_X \langle y_j^k, \psi_l \rangle_X \langle \psi_i, \psi_l \rangle_X \\ &= \|y_j^k\|_X^2 - \sum_{i=1}^{\ell} \langle y_j^k, \psi_i \rangle_X^2 \end{aligned} \tag{1.2}$$

holds for any set $\{\psi_i\}_{i=1}^{\ell} \subset X$ satisfying $\langle \psi_i, \psi_j \rangle_X = \delta_{ij}$. Thus, ($\mathbf{P}_n^\ell$) is equivalent with the maximization problem

$$\begin{cases} \max \sum_{k=1}^{\wp} \sum_{j=1}^{n} \alpha_j^n \sum_{i=1}^{\ell} \langle y_j^k, \psi_i \rangle_X^2 \\ \text{s.t. } \{\psi_i\}_{i=1}^{\ell} \subset X \text{ and } \langle \psi_i, \psi_j \rangle_X = \delta_{ij},\ 1 \le i, j \le \ell. \end{cases} \tag{$\hat{\mathbf{P}}_n^\ell$}$$

Suppose that $\{\psi_i\}_{i \in \mathscr{I}}$ is a complete orthonormal basis in X. Since X is separable, any $y_j^k \in X$, $1 \le j \le n$ and $1 \le k \le \wp$, can be written as

$$y_j^k = \sum_{i \in \mathscr{I}} \langle y_j^k, \psi_i \rangle_X \psi_i \tag{1.3}$$

and the (probably infinite) sum converges for all snapshots (even for all elements in X). Thus, the POD basis $\{\bar{\psi}_i^n\}_{i=1}^{\ell}$ of rank ℓ maximizes the absolute values of the first ℓ Fourier coefficients $\langle y_j^k, \psi_i \rangle_X$ for all $n\wp$ snapshots y_j^k in an average sense. Let us recall the following definition for linear operators in Banach spaces.

Definition 1.2. *Let $\mathscr{B}_1$, $\mathscr{B}_2$ be two real Banach spaces. The operator $\mathscr{T} : \mathscr{B}_1 \to \mathscr{B}_2$ is called a* linear, bounded operator *if these conditions are satisfied:*

1) $\mathscr{T}(\alpha u + \beta v) = \alpha \mathscr{T} u + \beta \mathscr{T} v$ *for all* $\alpha, \beta \in \mathbb{R}$ *and* $u, v \in \mathscr{B}_1$.

2) *There exists a constant* $c > 0$ *such that* $\|\mathscr{T} u\|_{\mathscr{B}_2} \le c \|u\|_{\mathscr{B}_1}$ *for all* $u \in \mathscr{B}_1$.

The set of all linear, bounded operators from $\mathscr{B}_1$ to $\mathscr{B}_2$ is denoted by $\mathscr{L}(\mathscr{B}_1, \mathscr{B}_2)$ which is a Banach space equipped with the operator norm [60, pp. 69-70]

$$\|\mathscr{T}\|_{\mathscr{L}(\mathscr{B}_1,\mathscr{B}_2)} = \sup_{\|u\|_{\mathscr{B}_1}=1} \|\mathscr{T}u\|_{\mathscr{B}_2} \quad \text{for } \mathscr{T} \in \mathscr{L}(\mathscr{B}_1, \mathscr{B}_2).$$

If $\mathscr{B}_1 = \mathscr{B}_2$ holds, we briefly write $\mathscr{L}(\mathscr{B}_1)$ instead of $\mathscr{L}(\mathscr{B}_1, \mathscr{B}_2)$. The dual mapping $\mathscr{T}' : \mathscr{B}_2' \to \mathscr{B}_1'$ *of an operator $\mathscr{T} \in \mathscr{L}(\mathscr{B}_1, \mathscr{B}_2)$ is defined as*

$$\langle \mathscr{T}' f, u \rangle_{\mathscr{B}_1', \mathscr{B}_1} = \langle f, \mathscr{T}u \rangle_{\mathscr{B}_2', \mathscr{B}_2} \quad \text{for all } (u, f) \in \mathscr{B}_1 \times \mathscr{B}_2',$$

where, for instance, $\langle \cdot, \cdot \rangle_{\mathscr{B}_1', \mathscr{B}_1}$ denotes the dual pairing of the space $\mathscr{B}_1$ with its dual space $\mathscr{B}_1' = \mathscr{L}(\mathscr{B}_1, \mathbb{R})$.

Let $\mathscr{H}_1$ and $\mathscr{H}_2$ denote two real Hilbert spaces. For a given $\mathscr{T} \in \mathscr{L}(\mathscr{H}_1, \mathscr{H}_2)$ the *adjoint operator* $\mathscr{T}^\star : \mathscr{H}_2 \to \mathscr{H}_1$ is uniquely defined by

$$\langle \mathscr{T}^\star v, u \rangle_{\mathscr{H}_1} = \langle v, \mathscr{T}u \rangle_{\mathscr{H}_2} = \langle \mathscr{T}u, v \rangle_{\mathscr{H}_2} \quad \text{for all } (u, v) \in \mathscr{H}_1 \times \mathscr{H}_2.$$

Let $\mathscr{J}_i : \mathscr{H}_i \to \mathscr{H}_i'$, $i = 1, 2$, denote the Riesz isomorphisms satisfying

$$\langle u, v \rangle_{\mathscr{H}_i} = \langle \mathscr{J}_i u, v \rangle_{\mathscr{H}_i', \mathscr{H}_i} \quad \text{for all } v \in \mathscr{H}_i.$$

Then, we have the representation $\mathscr{T}^\star = \mathscr{J}_1^{-1} \mathscr{T}' \mathscr{J}_2$; see [70, p. 186]. Moreover, $(\mathscr{T}^\star)^\star = \mathscr{T}$ for every $\mathscr{T} \in \mathscr{L}(\mathscr{H}_1, \mathscr{H}_2)$. If $\mathscr{T} = \mathscr{T}^\star$ holds, $\mathscr{T}$ is said to be *selfadjoint*. The operator $\mathscr{T} \in \mathscr{L}(\mathscr{H}_1, \mathscr{H}_2)$ is called *nonnegative* if $\langle \mathscr{T}u, u \rangle_{\mathscr{H}_2} \geq 0$ for all $u \in \mathscr{H}_1$. Finally, $\mathscr{T} \in \mathscr{L}(\mathscr{H}_1, \mathscr{H}_2)$ is called *compact* if for every bounded sequence $\{u_n\}_{n\in\mathbb{N}} \subset \mathscr{H}_1$ the sequence $\{\mathscr{T}u_n\}_{n\in\mathbb{N}} \subset \mathscr{H}_2$ contains a convergent subsequence.

Now we turn to $(\mathbf{P}_n^\ell)$ and $(\hat{\mathbf{P}}_n^\ell)$. We make use of the following lemma.

Lemma 1.3. *Let X be a (separable) real Hilbert space and $y_1^k, \ldots, y_n^k \in X$ are given snapshots for $1 \leq k \leq \wp$. Define the linear operator $\mathscr{R}^n : X \to X$ as follows:*

$$\mathscr{R}^n \psi = \sum_{k=1}^{\wp} \sum_{j=1}^{n} \alpha_j^n \langle \psi, y_j^k \rangle_X \, y_j^k \quad \text{for } \psi \in X \tag{1.4}$$

with positive weights $\alpha_1^n, \ldots, \alpha_n^n$. Then, $\mathscr{R}^n$ is a compact, nonnegative and selfadjoint operator.

Proof. It is clear that $\mathscr{R}^n$ is a linear operator. From

$$\|\mathscr{R}^n \psi\|_X \leq \sum_{k=1}^{\wp} \sum_{j=1}^{n} \alpha_j^n \left| \langle \psi, y_j^k \rangle_X \right| \|y_j^k\|_X \quad \text{for } \psi \in X$$

and the Cauchy-Schwarz inequality [60, p. 38]

$$\left| \langle \varphi, \phi \rangle_X \right| \leq \|\varphi\|_X \|\phi\|_X \quad \text{for } \varphi, \phi \in X$$

we conclude that $\mathscr{R}^n$ is bounded. Since $\mathscr{R}^n \psi \in \mathscr{V}^n$ holds for all $\psi \in X$, the range of $\mathscr{R}^n$ is finite dimensional. Thus, $\mathscr{R}^n$ is a finite rank operator which is compact; see [60,

p. 199]. Next we show that $\mathscr{R}^n$ is nonnegative. For that purpose we choose an arbitrary element $\psi \in X$ and consider

$$\langle \mathscr{R}^n \psi, \psi \rangle_X = \sum_{k=1}^{\wp} \sum_{j=1}^{n} \alpha_j^n \langle \psi, y_j^k \rangle_X \langle y_j^k, \psi \rangle_X = \sum_{k=1}^{\wp} \sum_{j=1}^{n} \alpha_j^n \langle \psi, y_j^k \rangle_X^2 \geq 0.$$

Thus, $\mathscr{R}^n$ is nonnegative. For any $\psi, \tilde{\psi} \in X$ we derive

$$\begin{aligned} \langle \mathscr{R}^n \psi, \tilde{\psi} \rangle_X &= \sum_{k=1}^{\wp} \sum_{j=1}^{n} \alpha_j^n \langle \psi, y_j^k \rangle_X \langle y_j^k, \tilde{\psi} \rangle_X = \sum_{k=1}^{\wp} \sum_{j=1}^{n} \alpha_j^n \langle \tilde{\psi}, y_j^k \rangle_X \langle y_j^k, \psi \rangle_X \\ &= \langle \mathscr{R}^n \tilde{\psi}, \psi \rangle_X = \langle \psi, \mathscr{R}^n \tilde{\psi} \rangle_X. \end{aligned}$$

Thus, $\mathscr{R}^n$ is selfadjoint. □

Next we recall some important results from the spectral theory of operators (on infinite dimensional spaces). We begin with the following definition; see [60, Section VI.3].

Definition 1.4. *Let $\mathscr{H}$ be a real Hilbert space and $\mathscr{T} \in \mathscr{L}(\mathscr{H})$.*

1) *A complex number λ belongs to the* resolvent set $\rho(\mathscr{T})$ *if $\lambda \mathscr{I} - \mathscr{T}$ is a bijection with a bounded inverse. Here, $\mathscr{I} \in \mathscr{L}(\mathscr{H})$ stands for the identity operator. If $\lambda \notin \rho(\mathscr{T})$, then λ is an element of the spectrum $\sigma(\mathscr{T})$ of $\mathscr{T}$.*

2) *Let $u \neq 0$ be a vector with $\mathscr{T} u = \lambda u$ for some $\lambda \in \mathbb{C}$. Then, u is said to be an* eigenvector *of $\mathscr{T}$. We call λ the corresponding* eigenvalue*. If λ is an eigenvalue, then $\lambda \mathscr{I} - \mathscr{T}$ is not injective. This implies $\lambda \in \sigma(\mathscr{T})$. The set of all eigenvalues is called the* point spectrum *of $\mathscr{T}$.*

We will make use of the next two essential theorems for compact operators; see [60, p. 203].

Theorem 1.5 (Riesz-Schauder). *Let $\mathscr{H}$ be a real Hilbert space and $\mathscr{T} : \mathscr{H} \to \mathscr{H}$ a linear, compact operator. Then the spectrum $\sigma(\mathscr{T})$ is a discrete set having no limit points except perhaps 0. Furthermore, the space of eigenvectors corresponding to each nonzero $\lambda \in \sigma(\mathscr{T})$ is finite dimensional.*

Theorem 1.6 (Hilbert-Schmidt). *Let $\mathscr{H}$ be a real separable Hilbert space and $\mathscr{T} : \mathscr{H} \to \mathscr{H}$ a linear, compact, selfadjoint operator. Then, there is a sequence of eigenvalues $\{\lambda_i\}_{i \in \mathscr{I}}$ and of an associated complete orthonormal basis $\{\psi_i\}_{i \in \mathscr{I}} \subset X$ satisfying*

$$\mathscr{T} \psi_i = \lambda_i \psi_i \quad \text{and} \quad \lambda_i \to 0 \text{ as } i \to \infty.$$

Since X is a separable real Hilbert space and $\mathscr{R}^n : X \to X$ is a linear, compact, nonnegative, selfadjoint operator (see Lemma 1.3), we can utilize Theorems 1.5 and 1.6: There exist a complete countable orthonormal basis $\{\bar{\psi}_i^n\}_{i \in \mathscr{I}}$ and a corresponding sequence of real eigenvalues $\{\bar{\lambda}_i^n\}_{i \in \mathscr{I}}$ satisfying

$$\mathscr{R}^n \bar{\psi}_i^n = \bar{\lambda}_i^n \bar{\psi}_i^n, \quad \bar{\lambda}_1^n \geq \ldots \geq \bar{\lambda}_{d^n} > \bar{\lambda}_{d^n+1} = \ldots = 0. \tag{1.5}$$

The spectrum of $\mathscr{R}^n$ is a pure point spectrum except for possibly 0. Each nonzero eigenvalue of $\mathscr{R}^n$ has finite multiplicity and 0 is the only possible accumulation point of the spectrum of $\mathscr{R}^n$.

Remark 1.7. From (1.4), (1.5) and $\|\psi\|_X = 1$ we infer that

$$\begin{aligned}\sum_{k=1}^{\wp}\sum_{j=1}^{n}\alpha_j^n \left\langle y_j^k, \bar{\psi}_i^n\right\rangle_X^2 &= \left\langle \sum_{k=1}^{\wp}\sum_{j=1}^{n}\alpha_j^n \left\langle y_j^k, \bar{\psi}_i^n\right\rangle_X y_j^k, \bar{\psi}_i^n\right\rangle_X \\ &= \left\langle \mathscr{R}^n\bar{\psi}_i^n, \bar{\psi}_i^n\right\rangle_X = \bar{\lambda}_i^n \quad \text{for any } i \in \mathscr{I}.\end{aligned} \tag{1.6}$$

In particular, it follows that

$$\sum_{k=1}^{\wp}\sum_{j=1}^{n}\alpha_j^n \left\langle y_j^k, \bar{\psi}_i^n\right\rangle_X^2 = 0 \quad \text{for all } i > d^n. \tag{1.7}$$

Since $\{\bar{\psi}_i^n\}_{i\in\mathscr{I}}$ is a complete orthonormal basis and $\|y_j^k\|_X < \infty$ holds for $1 \le k \le \wp$, $1 \le j \le n$, we derive from (1.6) and (1.7) that

$$\begin{aligned}\sum_{k=1}^{\wp}\sum_{j=1}^{n}\alpha_j^n \|y_j^k\|_X^2 &= \sum_{k=1}^{\wp}\sum_{j=1}^{n}\alpha_j^n \sum_{\nu\in\mathscr{I}}\left\langle y_j^k, \bar{\psi}_\nu^n\right\rangle_X^2 \\ &= \sum_{\nu\in\mathscr{I}}\sum_{k=1}^{\wp}\sum_{j=1}^{n}\alpha_j^n \left\langle y_j^k, \bar{\psi}_\nu^n\right\rangle_X^2 = \sum_{i\in\mathscr{I}}\bar{\lambda}_i^n = \sum_{i=1}^{d^n}\bar{\lambda}_i^n.\end{aligned} \tag{1.8}$$

By (1.8) the (probably infinite) sum $\sum_{i\in\mathscr{I}}\bar{\lambda}_i^n$ is bounded. It follows from (1.2) that the objective of $(\mathbf{P}_n^\ell)$ can be written as

$$\begin{aligned}&\sum_{k=1}^{\wp}\sum_{j=1}^{n}\alpha_j^n \left\| y_j^k - \sum_{i=1}^{\ell}\left\langle y_j^k, \bar{\psi}_i^n\right\rangle_X \bar{\psi}_i^n\right\|_X^2 \\ &\quad = \sum_{i=1}^{d^n}\bar{\lambda}_i^n - \sum_{k=1}^{\wp}\sum_{j=1}^{n}\alpha_j^n \sum_{i=1}^{\ell}\left\langle y_j^k, \bar{\psi}_i^n\right\rangle_X^2\end{aligned} \tag{1.9}$$

which we will use in the proof of Theorem 1.8. ◊

Now we can formulate the main result for $(\mathbf{P}_n^\ell)$ and $(\hat{\mathbf{P}}_n^\ell)$.

Theorem 1.8. *Let X be a separable real Hilbert space, $y_1^k, \dots, y_n^k \in X$ for $1 \le k \le \wp$ and $\mathscr{R}^n : X \to X$ be defined by* (1.4)*. Suppose that $\{\bar{\lambda}_i^n\}_{i\in\mathscr{I}}$ and $\{\bar{\psi}_i^n\}_{i\in\mathscr{I}}$ denote the nonnegative eigenvalues and associated orthonormal eigenfunctions of $\mathscr{R}^n$ satisfying* (1.5)*. Then, for every $\ell \in \{1, \dots, d^n\}$ the first ℓ eigenfunctions $\{\bar{\psi}_i^n\}_{i=1}^\ell$ solve $(\mathbf{P}_n^\ell)$ and $(\hat{\mathbf{P}}_n^\ell)$. Moreover, the value of the cost evaluated at the optimal solution $\{\bar{\psi}_i^n\}_{i=1}^\ell$ satisfies*

$$\sum_{k=1}^{\wp}\sum_{j=1}^{n}\alpha_j^n \left\| y_j^k - \sum_{i=1}^{\ell}\left\langle y_j^k, \bar{\psi}_i^n\right\rangle_X \bar{\psi}_i^n\right\|_X^2 = \sum_{i=\ell+1}^{d^n}\bar{\lambda}_i^n \tag{1.10}$$

and

$$\sum_{k=1}^{\wp}\sum_{j=1}^{n}\alpha_j^n \sum_{i=1}^{\ell}\left\langle y_j^k, \bar{\psi}_i^n\right\rangle_X^2 = \sum_{i=1}^{\ell}\bar{\lambda}_i^n. \tag{1.11}$$

Proof. We prove the claim for $(\hat{\mathbf{P}}_n^\ell)$ by finite induction over $\ell \in \{1,\dots,d^n\}$.

1) The base case: Let $\ell = 1$ and $\psi \in X$ with $\|\psi\|_X = 1$. Since $\{\bar{\psi}_\nu^n\}_{\nu\in\mathscr{I}}$ is a complete orthonormal basis in X, we have the representation

$$\psi = \sum_{\nu\in\mathscr{I}} \langle \psi, \bar{\psi}_\nu^n\rangle_X \, \bar{\psi}_\nu^n. \tag{1.12}$$

Inserting this expression for ψ in the objective of $(\hat{\mathbf{P}}_n^\ell)$ we find that

$$\begin{aligned}
\sum_{k=1}^{\wp}\sum_{j=1}^{n} \alpha_j^n \, \langle y_j^k, \psi\rangle_X^2 &= \sum_{k=1}^{\wp}\sum_{j=1}^{n} \alpha_j^n \Big\langle y_j^k, \sum_{\nu\in\mathscr{I}} \langle \psi, \bar{\psi}_\nu^n\rangle_X \, \bar{\psi}_\nu^n \Big\rangle_X^2 \\
&= \sum_{k=1}^{\wp}\sum_{j=1}^{n} \alpha_j^n \sum_{\nu\in\mathscr{I}}\sum_{\mu\in\mathscr{I}} \Big(\langle y_j^k, \langle \psi, \bar{\psi}_\nu^n\rangle_X \, \bar{\psi}_\nu^n\rangle_X \langle y_j^k, \langle \psi, \bar{\psi}_\mu^n\rangle_X \, \bar{\psi}_\mu^n\rangle_X \Big) \\
&= \sum_{k=1}^{\wp}\sum_{j=1}^{n} \alpha_j^n \sum_{\nu\in\mathscr{I}}\sum_{\mu\in\mathscr{I}} \Big(\langle y_j^k, \bar{\psi}_\nu^n\rangle_X \langle y_j^k, \bar{\psi}_\mu^n\rangle_X \langle \psi, \bar{\psi}_\nu^n\rangle_X \langle \psi, \bar{\psi}_\mu^n\rangle_X \Big) \\
&= \sum_{\nu\in\mathscr{I}}\sum_{\mu\in\mathscr{I}} \Big(\Big\langle \sum_{k=1}^{\wp}\sum_{j=1}^{n} \alpha_j^n \langle y_j^k, \bar{\psi}_\nu^n\rangle_X \, y_j^k, \bar{\psi}_\mu^n \Big\rangle_X \langle \psi, \bar{\psi}_\nu^n\rangle_X \langle \psi, \bar{\psi}_\mu^n\rangle_X \Big).
\end{aligned}$$

Utilizing (1.4), (1.5) and $\|\bar{\psi}_\nu^n\|_X = 1$ we find that

$$\begin{aligned}
\sum_{k=1}^{\wp}\sum_{j=1}^{n} \alpha_j^n \, \langle y_j^k, \psi\rangle_X^2 &= \sum_{\nu\in\mathscr{I}}\sum_{\mu\in\mathscr{I}} \Big(\langle \bar{\lambda}_\nu^n \bar{\psi}_\nu^n, \bar{\psi}_\mu^n\rangle_X \langle \psi, \bar{\psi}_\nu^n\rangle_X \langle \psi, \bar{\psi}_\mu^n\rangle_X \Big) \\
&= \sum_{\nu\in\mathscr{I}} \bar{\lambda}_\nu^n \, \langle \psi, \bar{\psi}_\nu^n\rangle_X^2.
\end{aligned}$$

From $\bar{\lambda}_1^n \geq \bar{\lambda}_\nu^n$ for all $\nu \in \mathscr{I}$ and (1.6) we infer that

$$\begin{aligned}
\sum_{\nu\in\mathscr{I}} \bar{\lambda}_\nu^n \, \langle \psi, \bar{\psi}_\nu^n\rangle_X^2 \leq \bar{\lambda}_1^n \sum_{\nu\in\mathscr{I}} \langle \psi, \bar{\psi}_\nu^n\rangle_X^2 &= \bar{\lambda}_1^n \, \|\psi\|_X^2 = \bar{\lambda}_1^n \\
&= \sum_{k=1}^{\wp}\sum_{j=1}^{n} \alpha_j^n \, \langle y_j^k, \bar{\psi}_1^n\rangle_X^2,
\end{aligned}$$

i.e., $\bar{\psi}_1^n$ solves $(\hat{\mathbf{P}}_n^\ell)$ for $\ell = 1$ and (1.11) holds. This gives the base case. Notice that (1.9) and (1.11) imply (1.10).

2) The induction hypothesis: Now we suppose that

$$\begin{cases} \text{for any } \ell \in \{1,\dots,d^n - 1\} \text{ the set } \{\bar{\psi}_i^n\}_{i=1}^\ell \subset X \text{ solve } (\hat{\mathbf{P}}_n^\ell) \\ \text{and } \displaystyle\sum_{k=1}^{\wp}\sum_{j=1}^{n} \alpha_j^n \sum_{i=1}^{\ell} \langle y_j^k, \bar{\psi}_i^n\rangle_X^2 = \sum_{i=1}^{\ell} \bar{\lambda}_i^n. \end{cases} \tag{1.13}$$

3) The induction step: We consider

$$\begin{cases} \max \displaystyle\sum_{k=1}^{\wp}\sum_{j=1}^{n} \alpha_j^n \sum_{i=1}^{\ell+1} \langle y_j^k, \psi_i\rangle_X^2 \\ \text{s.t. } \{\psi_i\}_{i=1}^{\ell+1} \subset X \text{ and } \langle \psi_i, \psi_j\rangle_X = \delta_{ij},\ 1 \leq i,j \leq \ell+1. \end{cases} \tag{$\hat{\mathbf{P}}_n^{\ell+1}$}$$

By (1.13) the elements $\{\bar{\psi}_i^n\}_{i=1}^{\ell}$ maximize the term

$$\sum_{k=1}^{\wp}\sum_{j=1}^{n}\alpha_j^n\sum_{i=1}^{\ell}\langle y_j^k,\psi_i\rangle_X^2.$$

Thus, $(\hat{\mathbf{P}}_n^{\ell+1})$ is equivalent with

$$\begin{cases}\max\sum_{k=1}^{\wp}\sum_{j=1}^{n}\alpha_j^n\langle y_j^k,\psi\rangle_X^2\\ \text{s.t. } \psi\in X \text{ and } \|\psi\|_X=1,\ \langle\psi,\bar{\psi}_i^n\rangle_X=0,\ 1\le i\le\ell.\end{cases}\tag{1.14}$$

Let $\psi\in X$ be given satisfying $\|\psi\|_X=1$ and $\langle\psi,\bar{\psi}_i^n\rangle_X=0$ for $i=1\ldots,\ell$. Then, using the representation (1.12) and $\langle\psi,\bar{\psi}_i^n\rangle_X=0$ for $i=1\ldots,\ell$, we derive as above

$$\sum_{k=1}^{\wp}\sum_{j=1}^{n}\alpha_j^n\langle y_j^k,\psi\rangle_X^2=\sum_{\nu\in\mathscr{I}}\bar{\lambda}_\nu^n\langle\psi,\bar{\psi}_\nu^n\rangle_X^2=\sum_{\nu>\ell}\bar{\lambda}_\nu^n\langle\psi,\bar{\psi}_\nu^n\rangle_X^2.$$

From $\bar{\lambda}_{\ell+1}^n\ge\bar{\lambda}_\nu^n$ for all $\nu\ge\ell+1$ and (1.6) we conclude that

$$\begin{aligned}\sum_{k=1}^{\wp}\sum_{j=1}^{n}\alpha_j^n\langle y_j^k,\psi\rangle_X^2&\le\bar{\lambda}_{\ell+1}^n\sum_{\nu>\ell}\langle\psi,\bar{\psi}_\nu^n\rangle_X^2\le\bar{\lambda}_{\ell+1}^n\sum_{\nu\in\mathscr{I}}\langle\psi,\bar{\psi}_\nu^n\rangle_X^2\\&=\bar{\lambda}_{\ell+1}^n\|\psi\|_X^2=\bar{\lambda}_{\ell+1}^n=\sum_{k=1}^{\wp}\sum_{j=1}^{n}\alpha_j^n\langle y_j^k,\bar{\psi}_{\ell+1}^n\rangle_X^2.\end{aligned}$$

Thus, $\bar{\psi}_{\ell+1}^n$ solves (1.14), which implies that $\{\bar{\psi}_i^n\}_{i=1}^{\ell+1}$ is a solution to $(\hat{\mathbf{P}}_n^{\ell+1})$ and

$$\sum_{k=1}^{\wp}\sum_{j=1}^{n}\alpha_j^n\sum_{i=1}^{\ell+1}\langle y_j^k,\bar{\psi}_i^n\rangle_X^2=\sum_{i=1}^{\ell+1}\bar{\lambda}_i^n.$$

Again, (1.9) and (1.11) imply (1.10).

It follows that the claim is proved. □

Remark 1.9. Theorem 1.8 can also be proved by using the theory of nonlinear programming; see [29, 73], for instance. In this case $(\hat{\mathbf{P}}_n^{\ell})$ is considered as an equality constrained optimization problem. Applying a Lagrangian framework it turns out that (1.5) are first-order necessary optimality conditions for $(\hat{\mathbf{P}}_n^{\ell})$. ◊

For the application of POD to concrete problems the choice of ℓ is certainly of central importance for applying POD. It appears that no general a-priori rules are available. Rather the choice of ℓ is based on heuristic considerations combined with observing the ratio of the modeled to the "total energy" contained in the snapshots $y_1^k,\ldots,y_n^k$, $1\le k\le\wp$, which is expressed by

$$\mathscr{E}(\ell)=\frac{\sum_{i=1}^{\ell}\bar{\lambda}_i^n}{\sum_{i=1}^{d^n}\bar{\lambda}_i^n}\in[0,1].$$

Utilizing (1.8) we have

$$\mathscr{E}(\ell) = \frac{\sum_{i=1}^{\ell} \bar{\lambda}_i^n}{\sum_{k=1}^{\wp} \sum_{j=1}^{n} \alpha_j^n \, \|y_j^k\|_X^2},$$

i.e., the computation of the eigenvalues $\{\bar{\lambda}_i\}_{i=\ell+1}^{d}$ is not necessary. This is utilized in numerical implementations when iterative eigenvalue solver are applied like, e.g., the Lanczos method; see [2, Chapter 10], for instance.

In the following we will discuss three examples which illustrate that POD is strongly related to the singular value decomposition of matrices.

Remark 1.10 (POD in Euclidean space $\mathbb{R}^m$; see [37]). Suppose that $X = \mathbb{R}^m$ with $m \in \mathbb{N}$ and $\wp = 1$ hold. Then we have n snapshot vectors $y_1, \ldots, y_n$ and introduce the rectangular matrix $Y = [y_1 | \ldots | y_n] \in \mathbb{R}^{m \times n}$ with rank $d^n \leq \min(m, n)$. Choosing $\alpha_j^n = 1$ for $1 \leq j \leq n$ problem $(\mathbf{P}_n^\ell)$ has the form

$$\begin{cases} \min \sum_{j=1}^{n} \left\| y_j - \sum_{i=1}^{\ell} (y_j^\top \psi_i) \psi_i \right\|_{\mathbb{R}^m}^2 \\ \text{s.t. } \{\psi_i\}_{i=1}^{\ell} \subset \mathbb{R}^m \text{ and } \psi_i^\top \psi_j = \delta_{ij}, \; 1 \leq i, j \leq \ell, \end{cases} \tag{1.15}$$

where $\|\cdot\|_{\mathbb{R}^m}$ stands for the Euclidean norm in $\mathbb{R}^m$ and "$\top$" denotes the transpose of a given vector (or matrix). From

$$\left(\mathscr{R}^n \psi\right)_i = \Big(\sum_{j=1}^{n} (y_j^\top \psi) y_j \Big)_i = \sum_{j=1}^{n} \sum_{l=1}^{m} Y_{lj} \psi_l Y_{ij} = \left(Y Y^\top \psi\right)_i, \quad \psi \in \mathbb{R}^m,$$

for each component $1 \leq i \leq m$ we infer that (1.5) leads to the symmetric $m \times m$ eigenvalue problem

$$Y Y^\top \bar{\psi}_i^n = \bar{\lambda}_i^n \bar{\psi}_i^n, \quad \bar{\lambda}_1^n \geq \ldots \geq \bar{\lambda}_{d^n}^n > \bar{\lambda}_{d^n+1}^n = \ldots = \bar{\lambda}_m^n = 0. \tag{1.16}$$

Recall that (1.16) can be solved by utilizing the singular value decomposition (SVD) [51]: There exist real numbers $\bar{\sigma}_1^n \geq \bar{\sigma}_2^n \geq \ldots \geq \bar{\sigma}_{d^n}^n > 0$ and orthogonal matrices $\Psi \in \mathbb{R}^{m \times m}$ with column vectors $\{\bar{\psi}_i^n\}_{i=1}^m$ and $\Phi \in \mathbb{R}^{n \times n}$ with column vectors $\{\bar{\phi}_i^n\}_{i=1}^n$ such that

$$\Psi^\top Y \Phi = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} =: \Sigma \in \mathbb{R}^{m \times n}, \tag{1.17}$$

where $D = \operatorname{diag}(\bar{\sigma}_1^n, \ldots, \bar{\sigma}_{d^n}^n) \in \mathbb{R}^{d \times d}$ and the zeros in (1.17) denote matrices of appropriate dimensions. Moreover the vectors $\{\bar{\psi}_i^n\}_{i=1}^d$ and $\{\bar{\phi}_i^n\}_{i=1}^d$ satisfy

$$Y \bar{\phi}_i^n = \bar{\sigma}_i^n \bar{\psi}_i^n \quad \text{and} \quad Y^\top \bar{\psi}_i^n = \bar{\sigma}_i^n \bar{\phi}_i^n \quad \text{for } i = 1, \ldots, d^n. \tag{1.18}$$

They are eigenvectors of $Y Y^\top$ and $Y^\top Y$, respectively, with eigenvalues $\bar{\lambda}_i^n = (\bar{\sigma}_i^n)^2 > 0$, $i = 1, \ldots, d^n$. The vectors $\{\bar{\psi}_i^n\}_{i=d^n+1}^m$ and $\{\bar{\phi}_i^n\}_{i=d^n+1}^n$ (if $d^n < m$ respectively $d^n < n$) are eigenvectors of $Y Y^\top$ and $Y^\top Y$ with eigenvalue 0. Consequently, in the case $n < m$ one can determine the POD basis of rank ℓ as follows: Compute the eigenvectors $\bar{\phi}_1^n, \ldots, \bar{\phi}_\ell^n \in \mathbb{R}^n$ by solving the symmetric $n \times n$ eigenvalue problem

$$Y^\top Y \bar{\phi}_i^n = \bar{\lambda}_i^n \bar{\phi}_i^n \quad \text{for } i = 1, \ldots, \ell$$

and set, by (1.18),

$$\bar{\psi}_i^n = \frac{1}{(\bar{\lambda}_i^n)^{1/2}} Y \bar{\phi}_i^n \quad \text{for } i = 1, \dots, \ell.$$

For historical reasons this method of determing the POD-basis is sometimes called the *method of snapshots*; see [67]. On the other hand, if $m < n$ holds, we can obtain the POD basis by solving the $m \times m$ eigenvalue problem (1.16). If the matrix Y is badly scaled, we should avoid to build the matrix product $YY^\top$ (or $Y^\top Y$). In this case the SVD turns out to be more stable for the numerical computation of the POD basis of rank ℓ. ◊

Remark 1.11 (POD in $\mathbb{R}^m$ with weighted inner product). As in Remark 1.10 we choose $X = \mathbb{R}^m$ with $m \in \mathbb{R}^m$ and $\wp = 1$. Let $W \in \mathbb{R}^{m \times m}$ be a given symmetric, positive definite matrix. We supply $\mathbb{R}^m$ with the weighted inner product

$$\langle \psi, \tilde{\psi} \rangle_W = \psi^\top W \tilde{\psi} = \langle \psi, W\tilde{\psi} \rangle_{\mathbb{R}^m} = \langle W\psi, \tilde{\psi} \rangle_{\mathbb{R}^m} \quad \text{for } \psi, \tilde{\psi} \in \mathbb{R}^m.$$

Then, problem $(\mathbf{P}_n^\ell)$ has the form

$$\begin{cases} \min \sum_{j=1}^n \alpha_j^n \Big\| y_j - \sum_{i=1}^\ell \langle y_j, \psi_i \rangle_W \psi_i \Big\|_W^2 \\ \text{s.t. } \{\psi_i\}_{i=1}^\ell \subset \mathbb{R}^m \text{ and } \langle \psi_i, \psi_j \rangle_W = \delta_{ij},\ 1 \le i, j \le \ell. \end{cases}$$

As in Remark 1.10 we introduce the matrix $Y = [y_1 | \dots | y_n] \in \mathbb{R}^{m \times n}$ with rank $d^n \le \min(m, n)$. Moreover, we define the diagonal matrix $D = \operatorname{diag}(\alpha_1^n, \dots, \alpha_n^n) \in \mathbb{R}^{n \times n}$. We find that

$$\begin{aligned} \big(\mathscr{R}^n \psi\big)_i &= \Big(\sum_{j=1}^n \alpha_j^n \langle y_j, \psi \rangle_W y_j \Big)_i = \sum_{j=1}^n \sum_{l=1}^m \sum_{\nu=1}^m \alpha_j^n Y_{lj} W_{l\nu} \psi_\nu Y_{ij} \\ &= \big(YDY^\top W \psi\big)_i \quad \text{for } \psi \in \mathbb{R}^m, \end{aligned}$$

for each component $1 \le i \le m$. Consequently, (1.5) leads to the eigenvalue problem

$$YDY^\top W \bar{\psi}_i^n = \bar{\lambda}_i^n \bar{\psi}_i^n, \quad \bar{\lambda}_1^n \ge \dots \ge \bar{\lambda}_{d^n}^n > \bar{\lambda}_{d^n+1}^n = \dots = \bar{\lambda}_m^n = 0. \tag{1.19}$$

Since W is symmetric and positive definite, W possesses an eigenvalue decomposition of the form $W = QBQ^\top$, where $B = \operatorname{diag}(\beta_1, \dots, \beta_m)$ contains the eigenvalues $\beta_1 \ge \dots \ge \beta_m > 0$ of W and $Q \in \mathbb{R}^{m \times m}$ is an orthogonal matrix. We define

$$W^r = Q \operatorname{diag}(\beta_1^r, \dots, \beta_m^r) Q^\top \quad \text{for } r \in \mathbb{R}.$$

Note that $(W^r)^{-1} = W^{-r}$ and $W^{r+s} = W^r W^s$ for $r, s \in \mathbb{R}$. Moreover, we have

$$\langle \psi, \tilde{\psi} \rangle_W = \langle W^{1/2}\psi, W^{1/2}\tilde{\psi} \rangle_{\mathbb{R}^m} \quad \text{for } \psi, \tilde{\psi} \in \mathbb{R}^m$$

and $\|\psi\|_W = \|W^{1/2}\psi\|_{\mathbb{R}^m}$ for $\psi \in \mathbb{R}^m$. Analogously, the matrix $D^{1/2}$ is defined. Inserting $\psi_i^n = W^{1/2}\bar{\psi}_i^n$ in (1.19), multiplying (1.19) by $W^{1/2}$ from the left and setting $\hat{Y} = W^{1/2} Y D^{1/2}$ yield the symmetric $m \times m$ eigenvalue problem

$$\hat{Y}\hat{Y}^\top \psi_i^n = \bar{\lambda}_i^n \psi_i^n, \quad 1 \le i \le \ell.$$

Note that

$$\hat{Y}^\top \hat{Y} = D^{1/2} Y^\top W Y D^{1/2} \in \mathbb{R}^{n\times n}. \tag{1.20}$$

Thus, the POD basis $\{\bar{\psi}_i^n\}_{i=1}^\ell$ of rank ℓ can also be computed by the methods of snapshots as follows: First solve the symmetric $n \times n$ eigenvalue problem

$$\hat{Y}^\top \hat{Y} \phi_i^n = \bar{\lambda}_i^n \phi_i^n,\ 1 \le i \le \ell \quad \text{and} \quad \langle \phi_i^n, \phi_j^n \rangle_{\mathbb{R}^n} = \delta_{ij},\ 1 \le i,j \le \ell.$$

Then we set (by using the SVD of $\hat{Y}$)

$$\bar{\psi}_i^n = W^{-1/2} \psi_i^n = \frac{1}{\bar{\sigma}_i^n} W^{-1/2} \hat{Y} \phi_i^n = \frac{1}{\bar{\sigma}_i^n} Y D^{1/2} \phi_i^n, \quad 1 \le i \le \ell. \tag{1.21}$$

Note that

$$\langle \bar{\psi}_i^n, \bar{\psi}_j^n \rangle_W = (\bar{\psi}_i^n)^\top W \bar{\psi}_j^n = \frac{1}{\bar{\sigma}_i^n \bar{\sigma}_j^n} (\phi_i^n)^\top \underbrace{D^{1/2} Y^\top W Y D^{1/2}}_{=\hat{Y}^\top \hat{Y}} \phi_j^n = \delta_{ij}$$

for $1 \le i,j \le \ell$. Thus, the POD basis $\{\bar{\psi}_i^n\}_{i=1}^\ell$ of rank ℓ is orthonormal in $\mathbb{R}^m$ with respect to the inner product $\langle \cdot, \cdot \rangle_W$. We observe from (1.20) and (1.21) that the computation of $W^{1/2}$ and $W^{-1/2}$ is not required. For applications, where W is not just a diagonal matrix, the method of snapshots turns out to be more attractive with respect to the computational costs even if $m > n$ holds. ◊

Remark 1.12 (POD in $\mathbb{R}^m$ with multiple snapshots). Let us discuss the more general case $\wp = 2$ in the setting of Remark 1.11. The extension for $\wp > 2$ is straightforward. We introduce the matrix $Y = [y_1^1 | \dots | y_n^1 | y_1^2 | \dots | y_n^2] \in \mathbb{R}^{m\times(n\wp)}$ with rank $d^n \le \min(m, n\wp)$. Then we find

$$\begin{aligned}\mathscr{R}^n \psi &= \sum_{j=1}^n \Big(\alpha_j^n \langle y_j^1, \psi \rangle_W y_j^1 + \alpha_j^n \langle y_j^2, \psi \rangle_W y_j^2 \Big) \\ &= Y \underbrace{\begin{pmatrix} D & 0 \\ 0 & D \end{pmatrix}}_{=:\tilde{D} \in \mathbb{R}^{(n\wp)\times(n\wp)}} Y^\top W \psi = Y \tilde{D} Y^\top W \psi \quad \text{for } \psi \in \mathbb{R}^m.\end{aligned}$$

Hence, (1.5) corresponds to the eigenvalue problem

$$Y \tilde{D} Y^\top W \bar{\psi}_i^n = \bar{\lambda}_i^n \bar{\psi}_i^n, \quad \bar{\lambda}_1^n \ge \dots \ge \bar{\lambda}_{d^n}^n > \bar{\lambda}_{d^n+1}^n = \dots = \bar{\lambda}_m^n = 0. \tag{1.22}$$

Setting $\psi_i^n = W^{1/2} \bar{\psi}_i^n$ in (1.22) and multiplying by $W^{1/2}$ from the left yield

$$W^{1/2} Y \tilde{D} Y^\top W^{1/2} \psi_i^n = \bar{\lambda}_i^n \psi_i^n. \tag{1.23}$$

Let $\hat{Y} = W^{1/2} Y \tilde{D}^{1/2} \in \mathbb{R}^{m\times(n\wp)}$. Using $W^\top = W$ as well as $\tilde{D}^\top = \tilde{D}$ we infer from (1.23) that the POD basis $\{\bar{\psi}_i^n\}_{i=1}^\ell$ of rank ℓ is given by the symmetric $m \times m$ eigenvalue problem

$$\hat{Y} \hat{Y}^\top \psi_i^n = \bar{\lambda}_i^n \psi_i^n,\ 1 \le i \le \ell, \quad \text{and} \quad \langle \psi_i^n, \psi_j^n \rangle_{\mathbb{R}^m} = \delta_{ij},\ 1 \le i,j \le \ell$$

and $\bar{\psi}_i^n = W^{-1/2} \psi_i^n$. Note that

$$\hat{Y}^\top \hat{Y} = \tilde{D}^{1/2} Y^\top W Y \tilde{D}^{1/2} \in \mathbb{R}^{(n\wp)\times(n\wp)}.$$

Thus, the POD basis of rank ℓ can also be computed by the methods of snapshots as follows: First solve the symmetric $(n\wp)\times(n\wp)$ eigenvalue problem

$$\hat{Y}^\top \hat{Y}\phi_i^n = \bar{\lambda}_i^n \phi_i,\ 1\le i\le \ell \quad \text{and} \quad \langle \phi_i^n, \phi_j^n\rangle_{\mathbb{R}^{n\wp}} = \delta_{ij},\ 1\le i,j\le \ell.$$

Then we set (by SVD)

$$\bar{\psi}_i^n = W^{-1/2}\psi_i^n = \frac{1}{\bar{\sigma}_i^n} W^{-1/2}\hat{Y}\phi_i^n = \frac{1}{\bar{\sigma}_i^n} Y\tilde{D}^{1/2}\phi_i^n$$

for $1\le i\le \ell$. $\Diamond$

1.2.2 ▪ The continuous variant of the POD method

As in Remark 1.1 tet $0\le t_1<t_2<\ldots<t_n\le T$ be a given time grid in the interval $[0,T]$ with equidistant with step-size $\Delta t = T/(n-1)$, i.e., $t_j=(j-1)\Delta t$. Suppose that we have trajectories $y^k\in C([0,T];X)$, $1\le k\le \wp$. Let the snapshots be given as $y_j^k = y^k(t_j)\in X$ or $y_j^k\approx y^k(t_j)\in X$. Then, the snapshot subspace $\mathcal{V}^n$ introduced in (1.1) depends on the chosen time instances $\{t_j\}_{j=1}^n$. Consequently, the POD basis $\{\bar{\psi}_i^n\}_{i=1}^\ell$ of rank ℓ as well as the corresponding eigenvalues $\{\bar{\lambda}_i^n\}_{i=1}^\ell$ depend also on the time instances (which has already been indicated by the superindex n). Moreover, we have not discussed so far what is the motivation to introduce the positive weights $\{\alpha_j^n\}_{j=1}^n$ in $(\mathbf{P}_n^\ell)$. For this reason we proceed by investigating the following two questions:

- How to choose good time instances for the snapshots?
- What are appropriate positive weights $\{\alpha_j^n\}_{j=1}^n$?

To address these two questions we will introduce a *continuous version* of POD. In Section 1.2.1 we have introduced the operator $\mathcal{R}^n$ in (1.4). By $\{\bar{\psi}_i^n\}_{i\in\mathcal{I}}$ and $\{\bar{\lambda}_i^n\}_{i\in\mathcal{I}}$ we have denoted the eigenfunctions and eigenvalues for $\mathcal{R}^n$ satisfying (1.5). Moreover, we have set $d^n=\dim\mathcal{V}^n$ for the dimension of the snapshot set. Let us now introduce the snapshot set by

$$\mathcal{V} = \operatorname{span}\left\{ y^k(t)\,|\, t\in[0,T] \text{ and } 1\le k\le \wp\right\}\subset X$$

with dimension $d\le\infty$. For any $\ell\le d$ we are interested in determining a POD basis of rank ℓ which minimizes the mean square error between the trajectories y^k and the corresponding ℓ-th partial Fourier sums on average in the time interval $[0,T]$:

$$\left\{\begin{array}{l} \min \displaystyle\sum_{k=1}^{\wp}\int_0^T \Big\| y^k(t) - \sum_{i=1}^{\ell}\langle y^k(t),\psi_i\rangle_X\,\psi_i\Big\|_X^2\,\mathrm{d}t \\ \text{s.t. } \{\psi_i\}_{i=1}^\ell\subset X \text{ and } \langle\psi_i,\psi_j\rangle_X=\delta_{ij},\ 1\le i,j\le\ell. \end{array}\right. \tag{$\mathbf{P}^\ell$}$$

An optimal solution $\{\bar{\psi}_i\}_{i=1}^\ell$ to $(\mathbf{P}^\ell)$ is called a *POD basis of rank* ℓ. Analogous to $(\hat{\mathbf{P}}_n^\ell)$ we can – instead of $(\mathbf{P}^\ell)$ – consider the problem

$$\left\{\begin{array}{l} \max \displaystyle\sum_{k=1}^{\wp}\int_0^T \sum_{i=1}^{\ell}\langle y^k(t),\psi_i\rangle_X^2\,\mathrm{d}t \\ \text{s.t. } \{\psi_i\}_{i=1}^\ell\subset X \text{ and } \langle\psi_i,\psi_j\rangle_X=\delta_{ij},\ 1\le i,j\le\ell. \end{array}\right. \tag{$\hat{\mathbf{P}}^\ell$}$$

A solution to ($\mathbf{P}^\ell$) and to ($\hat{\mathbf{P}}^\ell$) can be characterized by an eigenvalue problem for the linear integral operator $\mathscr{R} : X \to X$ given as

$$\mathscr{R}\psi = \sum_{k=1}^{\wp} \int_0^T \langle y^k(t), \psi \rangle_X \, y^k(t) \, \mathrm{d}t \quad \text{for } \psi \in X. \tag{1.24}$$

For the given real Hilbert space X we denote by $L^2(0,T;X)$ the Hilbert space of square integrable functions $t \mapsto \varphi(t) \in X$ so that [70, p. 143]

- the mapping $t \mapsto \varphi(t)$ is measurable for $t \in [0,T]$ and
- $\|\varphi\|_{L^2(0,T;X)} = \Big(\int_0^T \|\varphi(t)\|_X^2 \, \mathrm{d}t \Big)^{1/2} < \infty.$

Recall that $\varphi : [0,T] \to X$ is called *measurable* if there exists a sequence $\{\varphi_n\}_{n\in\mathbb{N}}$ of step functions $\varphi_n : [0,T] \to X$ satisfying $\varphi(t) = \lim_{n\to\infty} \varphi_n(t)$ for almost all $t \in [0,T]$. The standard inner product on $L^2(0,T;X)$ is given by

$$\langle \varphi, \psi \rangle_{L^2(0,T;X)} = \int_0^T \langle \varphi(t), \phi(t) \rangle_X \, \mathrm{d}t \quad \text{for } \varphi, \phi \in L^2(0,T;X).$$

Lemma 1.13. *Let X be a (separable) real Hilbert space and $y^k \in L^2(0,T;X)$, $1 \le k \le \wp$, be given snapshot trajectories. Then, the operator $\mathscr{R}$ introduced in* (1.24) *is compact, nonnegative and selfadjoint.*

Proof. First we write $\mathscr{R}$ as a product of an operator and its Hilbert space adjoint. For that purpose let us define the linear operator $\mathscr{Y} : L^2(0,T;\mathbb{R}^\wp) \to X$ by

$$\mathscr{Y}\phi = \sum_{k=1}^{\wp} \int_0^T \phi^k(t) y^k(t) \, \mathrm{d}t \quad \text{for } \phi = (\phi^1, \ldots, \phi^\wp) \in L^2(0,T;\mathbb{R}^\wp). \tag{1.25}$$

Utilizing the Cauchy-Schwarz inequality [60, p. 38] and $y^k \in L^2(0,T;X)$ for $1 \le k \le \wp$ we infer that

$$\begin{aligned}
\|\mathscr{Y}\phi\|_X &\le \sum_{k=1}^{\wp} \int_0^T |\phi^k(t)| \|y^k(t)\|_X \, \mathrm{d}t \le \sum_{k=1}^{\wp} \|\phi^k\|_{L^2(0,T)} \|y^k\|_{L^2(0,T;X)} \\
&\le \Big(\sum_{k=1}^{\wp} \|\phi^k\|_{L^2(0,T)}^2 \Big)^{1/2} \Big(\sum_{k=1}^{\wp} \|y^k(t)\|_X^2 \Big)^{1/2} \\
&= C_{\mathscr{Y}} \, \|\phi\|_{L^2(0,T;\mathbb{R}^\wp)} \quad \text{for any } \phi \in L^2(0,T;\mathbb{R}^\wp),
\end{aligned}$$

where we set $C_{\mathscr{Y}} = (\sum_{k=1}^{\wp} \|y^k(t)\|_X^2)^{1/2} < \infty$. Hence, the operator $\mathscr{Y}$ is bounded. Its Hilbert space adjoint $\mathscr{Y}^\star : X \to L^2(0,T;\mathbb{R}^\wp)$ satisfies

$$\langle \mathscr{Y}^\star \psi, \phi \rangle_{L^2(0,T;\mathbb{R}^\wp)} = \langle \psi, \mathscr{Y}\phi \rangle_X \quad \text{for } \psi \in X \text{ and } \phi \in L^2(0,T;\mathbb{R}^\wp).$$

Since we derive

$$\begin{aligned}
\langle \mathscr{Y}^\star \psi, \phi \rangle_{L^2(0,T;\mathbb{R}^\wp)} &= \langle \psi, \mathscr{Y}\phi \rangle_X = \Big\langle \psi, \sum_{k=1}^{\wp} \int_0^T \phi^k(t) y^k(t) \, \mathrm{d}t \Big\rangle_X \\
&= \sum_{k=1}^{\wp} \int_0^T \langle \psi, y^k(t) \rangle_X \phi^k(t) \, \mathrm{d}t = \Big\langle \big(\langle \psi, y^k(\cdot) \rangle_X \big)_{1\le k\le \wp}, \phi \Big\rangle_{L^2(0,T;\mathbb{R}^\wp)}
\end{aligned}$$

for $\psi \in X$ and $\phi \in L^2(0,T;\mathbb{R}^\wp)$, the adjoint operator is given by

$$(\mathscr{Y}^\star\psi)(t) = \begin{pmatrix} \langle \psi, y^1(t)\rangle_X \\ \vdots \\ \langle \psi, y^\wp(t)\rangle_X \end{pmatrix} \quad \text{for } \psi \in X \text{ and } t \in [0,T] \text{ a.e.},$$

where 'a.e.' stands for 'almost everywhere'. From (1.4) and

$$\big(\mathscr{Y}\mathscr{Y}^\star\big)\psi = \mathscr{Y}\begin{pmatrix} \langle \psi, y^1(\cdot)\rangle_X \\ \vdots \\ \langle \psi, y^\wp(\cdot)\rangle_X \end{pmatrix} = \sum_{k=1}^{\wp}\int_0^T \langle \psi, y^k(t)\rangle_X y^k(t)\,\mathrm{d}t \quad \text{for } \psi \in X$$

we infer that $\mathscr{R} = \mathscr{Y}\mathscr{Y}^\star$ holds. Since the operator $\mathscr{Y}$ is bounded, its adjoint and therefore $\mathscr{R} = \mathscr{Y}\mathscr{Y}^\star$ are bounded operators. To prove that $\mathscr{R}$ is compact, we show that $\mathscr{Y}^\star$ is compact. Let $\{\chi_n\}_{n\in\mathbb{N}} \subset X$ be sequence converging weakly to an element $\chi \in X$, i.e.,

$$\lim_{n\to\infty} \langle \chi_n, \psi\rangle_X = \langle \chi, \psi\rangle_X \quad \text{for all } \psi \in X.$$

This implies that

$$\lim_{n\to\infty} (\mathscr{Y}^\star\chi_n)(t) = \begin{pmatrix} \langle \chi_n, y^1(t)\rangle_X \\ \vdots \\ \langle \chi_n, y^\wp(t)\rangle_X \end{pmatrix} = \begin{pmatrix} \langle \chi_n, y^1(t)\rangle_X \\ \vdots \\ \langle \chi_n, y^\wp(t)\rangle_X \end{pmatrix} = \big(\mathscr{Y}^\star\chi\big)(t)$$

for $t \in [0,T]$ a.e. Thus, the sequence $\{\mathscr{Y}^\star\chi_n\}_{n\in\mathbb{N}}$ converges weakly to $\mathscr{Y}^\star\chi$ in $L^2(0,T;\mathbb{R}^\wp)$. Consequently, $\mathscr{R} = \mathscr{Y}\mathscr{Y}^\star$ is compact. From

$$\begin{aligned} \langle \mathscr{R}\psi, \psi\rangle_X &= \Big\langle \sum_{k=1}^{\wp}\int_0^T \langle \psi, y^k(t)\rangle_X\, y^k(t)\,\mathrm{d}t, \psi \Big\rangle_X \\ &= \sum_{k=1}^{\wp}\int_0^T \big|\langle \psi, y^k(t)\rangle_X\big|^2\,\mathrm{d}t \geq 0 \quad \text{for all } \psi \in X \end{aligned}$$

we infer that $\mathscr{R}$ is nonnegative. Finally, we have $\mathscr{R}^\star = (\mathscr{Y}\mathscr{Y}^\star)^\star = \mathscr{R}$, i.e. the operator $\mathscr{R}$ is selfadjoint. □

Remark 1.14. It follows from the proof of Lemma 1.13 that $\mathscr{K} = \mathscr{Y}^\star\mathscr{Y} : L^2(0,T;\mathbb{R}^\wp) \to L^2(0,T;\mathbb{R}^\wp)$ is compact as well. We find that

$$\big(\mathscr{K}\phi\big)(t) = \begin{pmatrix} \sum_{k=1}^{\wp}\int_0^T \langle y^k(s), y^1(t)\rangle_X \phi^k(s)\,\mathrm{d}s \\ \vdots \\ \sum_{k=1}^{\wp}\int_0^T \langle y^k(s), y^\wp(t)\rangle_X \phi^k(s)\,\mathrm{d}s \end{pmatrix}, \ \phi \in L^2(0,T;\mathbb{R}^\wp).$$

The compactness of $\mathscr{K}$ can also be shown as follow: Notice that the kernel function

$$r_{ik}(s,t) = \langle y^k(s), y^i(t)\rangle_X, \quad (s,t) \in [0,T]\times[0,T] \text{ and } 1 \leq i,k \leq \wp,$$

belongs to $L^2(0,T)\times L^2(0,T)$. Here, we shortly write $L^2(0,T)$ for $L^2(0,T;\mathbb{R})$. Then, it follows from [78, pp. 197 and 277] that the linear integral operator $\mathscr{K}_{ik}: L^2(0,T)\to L^2(0,T)$ defined by

$$\mathscr{K}_{ik}(t)=\int_0^T r_{ik}(s,t)\phi(s)\,\mathrm{d}s,\quad \phi\in L^2(0,T),$$

is a compact operator. This implies, that the operator $\sum_{k=1}^{\wp}\mathscr{K}_{ik}$ is compact for $1\le i\le\wp$ as well. ◊

In the next theorem we formulate how the solution to $(\mathbf{P}^\ell)$ and $(\hat{\mathbf{P}}^\ell)$ can be found.

Theorem 1.15. *Let X be a separable real Hilbert space and $y^k\in L^2(0,T;X)$ are given trajectories for $1\le k\le\wp$. Suppose that the linear operator $\mathscr{R}$ is defined by* (1.24)*. Then, there exist nonnegative eigenvalues $\{\bar\lambda_i\}_{i\in\mathscr{I}}$ and associated orthonomal eigenfunctions $\{\bar\psi_i\}_{i\in\mathscr{I}}$ satisfying*

$$\mathscr{R}\bar\psi_i=\bar\lambda_i\bar\psi_i,\quad \bar\lambda_1\ge\ldots\ge\bar\lambda_d>\bar\lambda_{d+1}=\ldots=0. \tag{1.26}$$

For every $\ell\in\{1,\ldots,d\}$ the first ℓ eigenfunctions $\{\bar\psi_i\}_{i=1}^\ell$ solve $(\mathbf{P}^\ell)$ and $(\hat{\mathbf{P}}^\ell)$. Moreover, the value of the objectives evaluated at the optimal solution $\{\bar\psi_i\}_{i=1}^\ell$ satisfies

$$\sum_{k=1}^{\wp}\int_0^T\Big\|y^k(t)-\sum_{i=1}^{\ell}\langle y^k(t),\bar\psi_i\rangle_X\,\bar\psi_i\Big\|_X^2\,\mathrm{d}t=\sum_{i=\ell+1}^{d}\bar\lambda_i \tag{1.27}$$

and

$$\sum_{k=1}^{\wp}\int_0^T\sum_{i=1}^{\ell}\langle y^k(t),\bar\psi_i\rangle_X^2\,\mathrm{d}t=\sum_{i=1}^{\ell}\bar\lambda_i, \tag{1.28}$$

respectively.

Proof. The existence of sequences $\{\bar\lambda_i\}_{i\in\mathscr{I}}$ of eigenvalues and $\{\bar\psi_i\}_{i\in\mathscr{I}}$ of associated eigenfunctions satisfying (1.26) follows from Lemma 1.13, Theorem 1.5 and Theorem 1.6. Analogous to the proof of Theorem 1.8 in Section 1.2.1 one can show that $\{\bar\psi_i\}_{i=1}^\ell$ solves $(\mathbf{P}^\ell)$ as well as $(\hat{\mathbf{P}}^\ell)$ and that (1.27) respectively (1.28) are valid. □

Remark 1.16. Similar to (1.6) we have

$$\sum_{k=1}^{\wp}\int_0^T\|y^k(t)\|_X^2\,\mathrm{d}t=\sum_{i=1}^{d}\bar\lambda_i. \tag{1.29}$$

In fact,

$$\mathscr{R}\bar\psi_i=\sum_{k=1}^{\wp}\int_0^T\langle y^k(t),\bar\psi_i\rangle_X\,y^k(t)\,\mathrm{d}t\quad\text{for every } i\in\mathscr{I}.$$

Taking the inner product with $\bar\psi_i$, using (1.26) and summing over i we get

$$\sum_{i=1}^{d}\sum_{k=1}^{\wp}\int_0^T\langle y^k(t),\bar\psi_i\rangle_X^2\,\mathrm{d}t=\sum_{i=1}^{d}\langle\mathscr{R}\bar\psi_i,\bar\psi_i\rangle_X=\sum_{i=1}^{d}\bar\lambda_i.$$

Expanding each $y^k(t) \in X$ in terms of $\{\bar{\psi}_i\}_{i\in\mathscr{I}}$ for each $1 \le k \le \wp$ we have

$$y^k(t) = \sum_{i=1}^{d} \langle y^k(t), \bar{\psi}_i \rangle_X \, \bar{\psi}_i$$

and hence

$$\sum_{k=1}^{\wp} \int_0^T \|y^k(t)\|_X^2 \, \mathrm{d}t = \sum_{k=1}^{\wp} \sum_{i=1}^{d} \int_0^T \langle y^k(t), \bar{\psi}_i \rangle_X^2 \, \mathrm{d}t = \sum_{i=1}^{d} \bar{\lambda}_i,$$

which is (1.29). ◊

Remark 1.17 (Singular value decomposition). Suppose that $y^k \in L^2(0,T;X)$ holds. By Theorem 1.15 there exist nonnegative eigenvalues $\{\bar{\lambda}_i\}_{i\in\mathscr{I}}$ and associated orthonomal eigenfunctions $\{\bar{\psi}_i\}_{i\in\mathscr{I}}$ satisfying (1.26). From $\mathscr{K} = \mathscr{Y}^\star\mathscr{Y}$ it follows that there is a sequence $\{\bar{\phi}_i\}_{i\in\mathscr{I}}$ such that

$$\mathscr{K}\bar{\phi}_i = \bar{\lambda}_i \bar{\phi}_i, \quad 1\ldots,\ell.$$

We set $\mathbb{R}_0^+ = \{s \in \mathbb{R} \,|\, s \ge 0\}$ and $\bar{\sigma}_i = \bar{\lambda}_i^{1/2}$. The sequence $\{\bar{\sigma}_i, \bar{\phi}_i, \bar{\psi}_i\}_{i\in\mathscr{I}}$ in $\mathbb{R}_0^+ \times L^2(0,T;\mathbb{R}^\wp) \times X$ can be interpreted as a singular value decomposition of the mapping $\mathscr{Y} : L^2(0,T;\mathbb{R}^\wp) \to X$ introduced in (1.25). In fact, we have

$$\mathscr{Y}\bar{\phi}_i = \bar{\sigma}_i \bar{\psi}_i, \quad \mathscr{Y}^\star\bar{\psi}_i = \bar{\sigma}_i \bar{\phi}_i, \quad i \in \mathscr{I}.$$

Since $\bar{\sigma}_i > 0$ holds for $1 = 1\ldots,d$, we have $\bar{\psi}_i = \mathscr{Y}\bar{\phi}_i/\sigma_i$ for $i = 1,\ldots,d$. ◊

1.2.3 ▪ Perturbation analysis for the POD basis

The eigenvalues $\{\bar{\lambda}_i^n\}_{i\in\mathscr{I}}$ satisfying (1.5) depend on the time grid $\{t_j\}_{j=1}^n$. In this section we investigate the sum $\sum_{i=\ell+1}^{d^n} \bar{\lambda}_i^n$, the value of the cost in $(\mathbf{P}_n^\ell)$ evaluated at the solution $\{\bar{\psi}_i^n\}_{i=1}^\ell$ for $n \to \infty$. Clearly, $n \to \infty$ is equivalent with $\Delta t = T/(n-1) \to 0$.

In general the spectrum $\sigma(\mathscr{T})$ of an operator $\mathscr{T} \in \mathscr{L}(X)$ does not depend continuously on $\mathscr{T}$. This is an essential difference to the finite dimensional case. For the compact and selfadjoint operator $\mathscr{R}$ we have $\sigma(\mathscr{R}) = \{\bar{\lambda}_i\}_{i\in\mathscr{I}}$. Suppose that for $\ell \in \mathbb{N}$ we have $\bar{\lambda}_\ell > \bar{\lambda}_{\ell+1}$ so that we can seperate the spectrum as follows: $\sigma(\mathscr{R}) = \mathscr{S}_\ell \cup \mathscr{S}_\ell'$ with $\mathscr{S}_\ell = \{\bar{\lambda}_1,\ldots,\bar{\lambda}_\ell\}$ and $\mathscr{S}_\ell' = \sigma(\mathscr{R}) \setminus \mathscr{S}_\ell$. Then, $\mathscr{S}_\ell \cap \mathscr{S}_\ell' = \emptyset$. Moreover, setting $V^\ell = \operatorname{span}\{\bar{\psi}_1,\ldots,\bar{\psi}_\ell\}$ we have $X = V^\ell \oplus (V^\ell)^\perp$, where the linear space $(V^\ell)^\perp$ stands for the X-orthogonal complement of V^ℓ. Let us assume that

$$\lim_{n\to\infty} \|\mathscr{R}^n - \mathscr{R}\|_{\mathscr{L}(X)} = 0 \tag{1.30}$$

holds. Then it follows from the perturbation theory of the spectrum of linear operators [35, pp. 212-214] that the space $V_n^\ell = \operatorname{span}\{\bar{\psi}_1^n,\ldots,\bar{\psi}_\ell^n\}$ is isomorphic to V^ℓ if n is sufficiently large. Furthermore, the change of a finite set of eigenvalues of $\mathscr{R}$ is small provided $\|\mathscr{R}^n - \mathscr{R}\|_{\mathscr{L}(X)}$ is sufficiently small. Summarizing, the behavior of the spectrum is much the same as in the finite dimensional case if we can ensure (1.30). Therefore, we start this section by investigating the convergence of $\mathscr{R}^n - \mathscr{R}$ in the operator norm.

Recall that the Sobolev space $H^1(0,T;X)$ is given by

$$H^1(0,T;X) = \left\{\varphi \in L^2(0,T;X) \,\middle|\, \varphi_t \in L^2(0,T;X)\right\},$$

where φ_t denotes the weak derivative of φ. The space $H^1(0,T;X)$ is a Hilbert space with the inner product

$$\langle \varphi, \phi \rangle_{H^1(0,T;X)} = \int_0^T \langle \varphi(t), \phi(t) \rangle_X + \langle \varphi_t(t), \phi_t(t) \rangle_X \, \mathrm{d}t \text{ for } \varphi, \phi \in H^1(0,T;X)$$

and the induced norm $||\varphi||_{H^1(0,T;X)} = \langle \varphi, \varphi \rangle^{1/2}_{H^1(0,T;X)}$.

Let us choose the trapezoidal weights

$$\alpha_1^n = \frac{T}{2(n-1)}, \ \alpha_j^n = \frac{T}{n-1} \text{ for } 2 \le j \le n-1, \ \alpha_n^n = \frac{T}{2(n-1)}. \tag{1.31}$$

For this choice we observe that for every $\psi \in X$ the element $\mathscr{R}^n \psi$ is a trapezoidal approximation for $\mathscr{R}\psi$. We will make use of the following lemma.

Lemma 1.18. *Suppose that X is a (separable) real Hilbert space and that the snapshot trajectories y^k belong to $H^1(0,T;X)$ for $1 \le k \le \wp$. Then,* (1.30) *holds true.*

Proof. For an arbitrary $\psi \in X$ with $||\psi||_X = 1$ we define $F : [0,T] \to X$ by

$$F(t) = \sum_{k=1}^{\wp} \langle y^k(t), \psi \rangle_X \, y^k(t) \quad \text{for } t \in [0,T].$$

It follows that

$$\begin{aligned} \mathscr{R}\psi &= \int_0^T F(t)\,\mathrm{d}t = \sum_{j=1}^{n-1} \int_{t_j}^{t_{j+1}} F(t)\,\mathrm{d}t, \\ \mathscr{R}^n \psi &= \sum_{j=1}^{n} \alpha_j F(t_j) = \frac{\Delta t}{2} \sum_{j=1}^{n-1} \big(F(t_j) + F(t_{j+1}) \big). \end{aligned} \tag{1.32}$$

Then, we infer from $||\psi||_X = 1$ that

$$||F(t)||_X^2 \le \Big(\sum_{k=1}^{\wp} ||y^k(t)||_X^2 \Big)^2. \tag{1.33}$$

Now we show that F belongs to $H^1(0,T;X)$ and its norm is bounded independently of ψ. Recall that $y^k \in H^1(0,T;X)$ imply that $y^k \in C([0,T];X)$ holds for $1 \le k \le \wp$. Using (1.33) we have

$$||F||^2_{L^2(0,T;X)} \le \int_0^T \Big(\sum_{k=1}^{\wp} ||y^k||^2_{C([0,T];X)} \Big)^2 \mathrm{d}t \le C_1$$

with $C_1 = T(\sum_{k=1}^{\wp} ||y^k||^2_{C([0,T];X)})^2$. Moreover, $F \in H^1(0,T;X)$ with

$$F_t(t) = \sum_{k=1}^{\wp} \langle y_t^k(t), \psi \rangle_X \, y^k(t) + \langle y^k(t), \psi \rangle_X \, y_t^k(t) \quad \text{f.a.a. } t \in [0,T],$$

where 'f.a.a.' stands for 'for almost all'. Thus, we derive

$$||F_t||^2_{L^2(0,T;X)} \le 4 \int_0^T \Big(\sum_{k=1}^{\wp} ||y^k(t)||_X ||y_t^k(t)||_X \Big)^2 \mathrm{d}t \le C_2$$

with $C_2 = 4\sum_{k=1}^{\wp} \|y^k\|^2_{C([0,T];X)} \sum_{l=1}^{\wp} \|y^l_t\|^2_{L^2(0,T;X)} < \infty$. Consequently,

$$\|F\|_{H^1(0,T;X)} = \Big(\int_0^T \|F(t)\|_X^2 + \|F_t(t)\|_X^2 \,\mathrm{d}t \Big)^{1/2} \le C_3 \tag{1.34}$$

with the constant $C_3 = (C_1 + C_2)^{1/2}$, which is independent of ψ. To evaluate $\mathscr{R}^n\psi$ we notice that

$$\begin{aligned} \int_{t_j}^{t_{j+1}} F(t)\,\mathrm{d}t &= \frac{1}{2}\int_{t_j}^{t_{j+1}} \Big(F(t_j) + \int_{t_j}^{t} F_t(s)\,\mathrm{d}s \Big)\mathrm{d}t \\ &\quad + \frac{1}{2}\int_{t_j}^{t_{j+1}} \Big(F(t_{j+1}) + \int_{t_{j+1}}^{t} F_t(s)\,\mathrm{d}s \Big)\mathrm{d}t \\ &= \frac{\Delta t}{2}\big(F(t_j) + F(t_{j+1}) \big) \\ &\quad + \frac{1}{2}\int_{t_j}^{t_{j+1}} \Big(\int_{t_{j+1}}^{t} F_t(s)\,\mathrm{d}s + \int_{t_{j+1}}^{t} F_t(s)\,\mathrm{d}s \Big)\mathrm{d}t. \end{aligned} \tag{1.35}$$

Utilizing (1.32) and (1.35) we obtain

$$\begin{aligned} \big\|\mathscr{R}^n\psi - \mathscr{R}\psi\big\|_X &= \Big\| \sum_{j=1}^{n-1} \Big(\frac{\Delta t}{2}\big(F(t_j) + F(t_{j+1}) \big) - \int_{t_j}^{t_{j+1}} F(t)\,\mathrm{d}t \Big) \Big\|_X \\ &\le \frac{1}{2}\sum_{j=1}^{n-1} \Big\| \int_{t_j}^{t_{j+1}} \int_{t_j}^{t} F_t(s)\,\mathrm{d}s\mathrm{d}t \Big\|_X + \frac{1}{2}\sum_{j=1}^{n-1} \Big\| \int_{t_j}^{t_{j+1}} \int_{t_{j+1}}^{t} F_t(s)\,\mathrm{d}s\mathrm{d}t \Big\|_X . \end{aligned}$$

From the Cauchy-Schwarz inequality [60, p. 38] we deduce that

$$\begin{aligned} &\sum_{j=1}^{n-1} \Big\| \int_{t_j}^{t_{j+1}} \int_{t_j}^{t} F_t(s)\,\mathrm{d}s\mathrm{d}t \Big\|_X \le \sum_{j=1}^{n-1} \int_{t_j}^{t_{j+1}} \Big\| \int_{t_j}^{t} F_t(s)\,\mathrm{d}s \Big\|_X \mathrm{d}t \\ &\le \sqrt{\Delta t} \sum_{j=1}^{n-1} \Big(\int_{t_j}^{t_{j+1}} \Big\| \int_{t_j}^{t} F_t(s)\,\mathrm{d}s \Big\|_X^2 \mathrm{d}t \Big)^{1/2} \\ &\le \sqrt{\Delta t} \sum_{j=1}^{n-1} \Big(\int_{t_j}^{t_{j+1}} \Big(\int_{t_j}^{t} \|F_t(s)\|_X \,\mathrm{d}s \Big)^2 \mathrm{d}t \Big)^{1/2} \\ &\le \Delta t \sum_{j=1}^{n-1} \Big(\int_{t_j}^{t_{j+1}} \int_{t_j}^{t} \|F_t(s)\|_X^2 \,\mathrm{d}s\mathrm{d}t \Big)^{1/2} \le T\sqrt{\Delta t}\,\|F\|_{H^1(0,T;X)}. \end{aligned} \tag{1.36}$$

Analogously, we derive

$$\sum_{j=1}^{n-1} \Big\| \int_{t_j}^{t_{j+1}} \int_{t_{j+1}}^{t} F_t(s)\,\mathrm{d}s\mathrm{d}t \Big\|_X \le T\sqrt{\Delta t}\,\|F\|_{H^1(0,T;X)}. \tag{1.37}$$

From (1.34), (1.36) and (1.37) it follows that

$$\big\|\mathscr{R}^n\psi - \mathscr{R}\psi\big\|_X \le \frac{C_4}{\sqrt{n}},$$

where $C_4 = C_3 T^{3/2}$ is independent of n and ψ. Consequently,

$$\|\mathscr{R}^n - \mathscr{R}\|_{\mathscr{L}(X)} = \sup_{\|\psi\|_X=1} \|\mathscr{R}^n\psi - \mathscr{R}\psi\|_X \overset{n\to\infty}{\longrightarrow} 0$$

which gives the claim. □

Now we follow [40, Section 3.2]. We suppose that $y^k \in H^1(0,T;X)$ for $1 \le k \le \wp$. Thus $y^k \in C([0,T];X)$ holds, which implies that

$$\sum_{k=1}^{\wp}\sum_{j=1}^{n} \alpha_j^n \|y^k(t_j)\|_X^2 \to \sum_{k=1}^{\wp}\int_0^T \|y^k(t)\|_X^2 \,\mathrm{d}t \quad \text{as } n\to\infty. \tag{1.38}$$

Combining (1.38) with (1.8) and (1.29) we find

$$\sum_{i=1}^{d^n} \bar{\lambda}_i^n \to \sum_{i=1}^{d} \bar{\lambda}_i \quad \text{as } n\to\infty. \tag{1.39}$$

Now choose and fix

$$\ell \quad \text{such that} \quad \bar{\lambda}_\ell \neq \bar{\lambda}_{\ell+1}. \tag{1.40}$$

Then, by spectral analysis of compact operators and Lemma 1.18 it follows that

$$\bar{\lambda}_i^n \to \bar{\lambda}_i \quad \text{for } 1 \le i \le \ell \text{ as } n\to\infty. \tag{1.41}$$

Combining (1.39) and (1.41) we derive

$$\sum_{i=\ell+1}^{d^n} \bar{\lambda}_i^n \to \sum_{i=\ell+1}^{d} \bar{\lambda}_i \quad \text{as } n\to\infty.$$

Especially, if $\lambda_1 > \lambda_2 > \cdots > \lambda_\ell$ is satisfied, we conclude from (1.40) and Lemma 1.18 that $\lim_{n\to\infty} \|\bar{\psi}_i^n - \bar{\psi}_i\|_X = 0$ for $i = 1,\dots,\ell$. Summarizing the following theorem has been shown.

Theorem 1.19. *Let X be a separable real Hilbert space, the weighting parameters $\{\alpha_j^n\}_{j=1}^n$ be given by* (1.31) *and $y^k \in H^1(0,T;X)$ for $1 \le k \le \wp$. Let $\{(\bar{\psi}_i^n, \bar{\lambda}_i^n)\}_{i\in\mathscr{I}}$ and $\{(\bar{\psi}_i, \bar{\lambda}_i)\}_{i\in\mathscr{I}}$ be eigenvector-eigenvalue pairs satisfying* (1.5) *and* (1.26)*, respectively. Suppose that $\ell \in \mathbb{N}$ is fixed such that* (1.40) *holds. Then we have*

$$\lim_{n\to\infty} \left|\bar{\lambda}_i^n - \bar{\lambda}_i\right| = 0 \quad \textit{for } 1 \le i \le \ell,$$

and

$$\lim_{n\to\infty} \sum_{i=\ell+1}^{d^n} \bar{\lambda}_i^n = \sum_{i=\ell+1}^{d} \bar{\lambda}_i.$$

In particular, if $\lambda_1 > \lambda_2 > \cdots > \lambda_\ell$ holds, then we even have

$$\lim_{n\to\infty} \|\bar{\psi}_i^n - \bar{\psi}_i\|_X = 0 \qquad \textit{for } 1 \le i \le \ell.$$

Remark 1.20. Theorem 1.19 gives an answer to the two questions posed at the beginning of Section 1.2.2: The time instances $\{t_j\}_{j=1}^n$ and the associated positive weights $\{\alpha_j^n\}_{j=1}^n$

should be chosen such that $\mathscr{R}^n$ is a quadrature approximation of $\mathscr{R}$ and $\|\mathscr{R}^n - \mathscr{R}\|_{\mathscr{L}(X)}$ is small (for reasonable n). A different strategy is applied in [42], where the time instances $\{t_j\}_{j=1}^n$ are chosen by an optimization approach. Clearly, other choices for the weights $\{\alpha_j^n\}_{j=1}^n$ are also possible provided (1.30) is guaranteed. For instance, we can choose the Simpson weights. ◊

1.3 ▪ Reduced-order modelling for evolution problems

In this section we utilize the POD method to derive low-dimensional models for evolution problems. For that purpose the POD basis of rank ℓ serves as test and ansatz functions in a POD Galerkin approximation. The a-priori error of the POD Galerkin scheme is investigated. It turns out that the resulting error bounds depend on the number of POD basis functions. Let us refer, e.g., to [20, 22, 30, 38, 40, 39, 62] and [32], where POD Galerkin schemes for parabolic equations and elliptic equations are studied. Moreover, we would like to mention the recent papers [8] and [66], where improved rates of convergence results are derived.

1.3.1 ▪ The abstract evolution problem

In this subsection we introduce our abstract evolution problem which will be under consideration in Sections 1.3 and 1.4. Let V and H be real, separable Hilbert spaces and suppose that V is dense in H with compact embedding. By $\langle\cdot,\cdot\rangle_H$ and $\langle\cdot,\cdot\rangle_V$ we denote the inner products in H and V, respectively. In particular, there exists a constant $c_V > 0$ such that

$$\|\varphi\|_H \le c_V \|\varphi\|_V \quad \text{for all } \varphi \in V. \tag{1.42}$$

Let $T > 0$ the final time. For $t \in [0,T]$ we define a time-dependent symmetric bilinear form $a(t;\cdot,\cdot) : V \times V \to \mathbb{R}$ satisfying

$$|a(t;\varphi,\psi)| \le \gamma \|\varphi\|_V \|\psi\|_V \qquad \forall \varphi \in V \text{ a.e. in } [0,T], \tag{1.43a}$$

$$a(t;\varphi,\varphi) \ge \gamma_1 \|\varphi\|_V^2 - \gamma_2 \|\varphi\|_H^2 \qquad \forall \varphi \in V \text{ a.e. in } [0,T] \tag{1.43b}$$

for constants γ, $\gamma_1 > 0$ and $\gamma_2 \ge 0$ which do not depend on t. In (1.43), the abbreviation "a.e." stands for "almost everywhere". By identifying H with its dual H' it follows that $V \hookrightarrow H = H' \hookrightarrow V'$ each embedding being continuous and dense. Here, V' denotes the dual space of V. Recall that the function space (see [10, pp. 472-479] and [70, pp. 146-148], for instance)

$$W(0,T) = \left\{ \varphi \in L^2(0,T;V) \,\middle|\, \varphi_t \in L^2(0,T;V') \right\}$$

is a Hilbert space endowed with the inner product

$$\langle \varphi, \phi \rangle_{W(0,T)} = \int_0^T \langle \varphi(t), \phi(t) \rangle_V + \langle \varphi_t(t), \phi_t(t) \rangle_{V'} \, \mathrm{d}t \text{ for } \varphi, \phi \in W(0,T)$$

and the induced norm $\|\varphi\|_{W(0,T)} = \langle \varphi, \varphi \rangle_{W(0,T)}^{1/2}$. Furthermore, $W(0,T)$ is continuously embedded into the space $C([0,T];H)$. Hence, $\varphi(0)$ and $\varphi(T)$ are meaningful in H for an element $\varphi \in W(0,T)$. The integration by parts formula reads

$$\begin{aligned}\int_0^T \langle \varphi_t(t), \phi(t) \rangle_{V',V} \, \mathrm{d}t + \int_0^T \langle \phi_t(t), \varphi(t) \rangle_{V',V} \, \mathrm{d}t &= \frac{\mathrm{d}}{\mathrm{d}t} \int_0^T \langle \varphi(t), \psi(t) \rangle_H \, \mathrm{d}t \\ &= \varphi(T)\phi(T) - \varphi(0)\phi(0)\end{aligned}$$

for $\varphi, \phi \in W(0,T)$, where $\langle \cdot, \cdot \rangle_{V',V}$ stands for the dual pairing between V and its dual space V'. Moreover, we have the formula

$$\langle \varphi_t(t), \phi \rangle_{V',V} = \frac{\mathrm{d}}{\mathrm{d}t} \langle \varphi(t), \phi \rangle_H \quad \text{for } (\varphi, \phi) \in W(0,T) \times V \text{ and f.a.a. } t \in [0,T].$$

Since we will consider optimal control problems in Section 1.4, we already introduce the evolution problem with an input term. We suppose that for $N_u \in \mathbb{N}$ the input space $U = L^2(0,T;\mathbb{R}^{N_u})$ is chosen. In particular, we identify U with its dual space U'. For $u \in U$, $y_\circ \in H$ and $f \in L^2(0,T;V')$ we consider the linear evolution problem

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \langle y(t), \varphi \rangle_H + a(t; y(t), \varphi) &= \langle (f + \mathscr{B}u)(t), \varphi \rangle_{V',V} \\ & \qquad \forall \varphi \in V \text{ a.e. in } (0,T], \\ \langle y(0), \varphi \rangle_H &= \langle y_\circ, \varphi \rangle_H \qquad \forall \varphi \in H, \end{aligned} \tag{1.44}$$

where $\mathscr{B} : U \to L^2(0,T;V')$ is a continuous, linear (control or input) operator.

Remark 1.21. Notice that the techniques presented in this work can be adapted for problems, where the input space U is given by $L^2(0,T;L^2(\mathscr{D}))$ for some open and bounded domain $\mathscr{D} \subset \mathbb{R}^{\widetilde{N}_u}$ for an $\widetilde{N}_u \in \mathbb{N}$. ◊

Theorem 1.22. *For $t \in [0,T]$ let $a(t;\cdot,\cdot) : V \times V \to \mathbb{R}$ be a time-dependent symmetric bilinear form satisfying* (1.43)*. Then, for every $u \in U$, $f \in L^2(0,T;V')$ and $y_\circ \in H$ there is a unique weak solution $y \in W(0,T)$ satisfying* (1.44) *and*

$$\|y\|_{W(0,T)} \leq C \left(\|y_\circ\|_H + \|f\|_{L^2(0,T;V')} + \|u\|_U \right) \tag{1.45}$$

for a constant $C > 0$ which is independent of u, $y_\circ$ and f. If $f \in L^2(0,T;H)$, $a(t;\cdot,\cdot) = a(\cdot,\cdot)$ (independent of t) and $y_\circ \in V$ hold, we even have $y \in L^\infty(0,T;V) \cap H^1(0,T;H)$. Here, $L^\infty(0,T;V)$ stands for the Banach space of all measurable functions $\varphi : [0,T] \to V$ with $\operatorname{ess\,sup}_{t\in[0,T]} \|\varphi(t)\|_V < \infty$ (see [70, p. 143]*, for instance).*

Proof. For a proof of the existence of a unique solution we refer to [10, pp. 512-520]. The a-priori error estimate follows from standard variational techniques and energy estimates. The regularity result follows from [10, pp. 532-533] and [17, pp. 360-364]. □

Remark 1.23. We split the solution to (1.44) in one part, which depends on the fixed initial condition $y_\circ$ and right-hand f, and another part depending linearly on the input variable u. Let $\hat{y} \in W(0,T)$ be the unique solution to

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \langle \hat{y}(t), \varphi \rangle_H + a(t; \hat{y}(t), \varphi) &= \langle f(t), \varphi \rangle_{V',V} \qquad \forall \varphi \in V \text{ a.e. in } (0,T], \\ \hat{y}(0) &= y_\circ \qquad \text{in } H. \end{aligned}$$

We define the subspace

$$W_0(0,T) = \left\{ \varphi \in W(0,T) \,\middle|\, \varphi(0) = 0 \text{ in } H \right\}$$

endowed with the topology of $W(0,T)$. Let us now introduce the linear solution operator $\mathscr{S}: U \to W_0(0,T)$: for $u \in U$ the function $y = \mathscr{S}u \in W_0(0,T)$ is the unique solution to

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle y(t),\varphi\rangle_H + a(t;y(t),\varphi) = \langle(\mathscr{B}u)(t),\varphi\rangle_{V',V} \quad \forall \varphi \in V \text{ a.e. in } (0,T].$$

From $y \in W_0(0,T)$ we infer $y(0) = 0$ in H. The boundedness of $\mathscr{S}$ follows from (1.45). Now, the solution to (1.44) can be expressed as $y = \hat{y} + \mathscr{S}u$. ◊

1.3.2 ▪ The POD method for the evolution problem

Let $u \in U$, $f \in L^2(0,T;V')$ and $y_\circ \in H$ be given and $y = \hat{y} + \mathscr{S}u$. To keep the notation simple we apply only a spatial discretization with POD basis functions, but no time integration by, e.g., the implicit Euler method. Therefore, we utilize the continuous version of the POD method introduced in Section 1.2.2. In this section we distinguish two choices for X: $X = H$ and $X = V$. It turns out that the choice for X leads to different rate of convergence results. We suppose that the snapshots y^k, $k = 1,\dots,\wp$, belong to $L^2(0,T;V)$. Later, we will present different rate of convergence results for appropriate choices of the y^k's. Let us introduce the following notations:

$$\begin{aligned}
\mathscr{R}_V\psi &= \sum_{k=1}^{\wp}\int_0^T \langle\psi, y^k(t)\rangle_V\, y^k(t)\,\mathrm{d}t && \text{for } \psi \in V,\\
\mathscr{R}_H\psi &= \sum_{k=1}^{\wp}\int_0^T \langle\psi, y^k(t)\rangle_H\, y^k(t)\,\mathrm{d}t && \text{for } \psi \in H.
\end{aligned} \tag{1.46}$$

Moreover, we set $\mathscr{K}_V = \mathscr{R}_V^\star$ and $\mathscr{K}_H = \mathscr{R}_H^\star$. In Remark 1.17 we have introduced the singular value decomposition of the operator $\mathscr{Y}$ defined by (1.25). To distinguish the two choices for the Hilbert space X we denote by the sequence $\{(\sigma_i^V,\psi_i^V,\phi_i^V)\}_{i\in\mathscr{I}}^{\ell} \subset \mathbb{R}_0^+ \times V \times L^2(0,T;\mathbb{R}^\wp)$ of triples the singular value decomposition for $X = V$, i.e., we have that

$$\mathscr{R}_V\psi_i^V = \lambda_i^V\psi_i^V, \quad \mathscr{K}_V\phi_i^V = \lambda_i^V\phi_i^V, \quad \sigma_i^V = \sqrt{\lambda_i^V}, \quad i \in \mathscr{I}.$$

Furthermore, let the sequence $\{(\sigma_i^H,\psi_i^H,\phi_i^H)\}_{i\in\mathscr{I}}^{\ell} \subset \mathbb{R}_0^+ \times H \times L^2(0,T;\mathbb{R}^\wp)$ in satisfy

$$\mathscr{R}_H\psi_i^H = \lambda_i^H\psi_i^H, \quad \mathscr{K}_H\phi_i^H = \lambda_i^H\phi_i^H, \quad \sigma_i^H = \sqrt{\lambda_i^H}, \quad i \in \mathscr{I}. \tag{1.47}$$

The relationship between the singular values σ_i^H and σ_i^V is investigated in the next lemma, which is taken from [66].

Lemma 1.24. *Suppose that the snapshots $y^k \in L^2(0,T;V)$, $k = 1,\dots,\wp$. Then we have:*

1) *For all $i \in \mathscr{I}$ with $\sigma_i^H > 0$ we have $\psi_i^H \in V$.*
2) *$\sigma_i^V = 0$ for all $i > d$ with some $d \in \mathbb{N}$ if and only if $\sigma_i^H = 0$ for all $i > d$, i.e., we have $d_H = d_V$ if the rank of $\mathscr{R}_V$ is finite.*
3) *$\sigma_i^V > 0$ for all $i \in \mathscr{I}$ if and only if $\sigma_i^H > 0$ for all $i \in \mathscr{I}$.*

Proof. We argue similarly as in the proof of Lemma 3.1 in [66].

1) Let $\sigma_i^H > 0$ hold. Then, it follows that $\lambda_i^H > 0$. We infer from $y^k \in L^2(0,T;V)$ that $\mathscr{R}_H \psi \in V$ for any $\psi \in H$. Hence, we infer from (1.47) and that $\psi_i^H = \mathscr{R}_H \psi_i^H / \lambda_i^H \in V$.

2) Assume that $\sigma_i^V = 0$ for all $i > d$ with some $d \in \mathbb{N}$. Then, we deduce from (1.27) that

$$y^k(t) = \sum_{i=1}^{d} \langle y^k(t), \psi_i^V \rangle_V \, \psi_i^V \quad \text{for every } k = 1, \dots, \wp. \tag{1.48}$$

From

$$\begin{aligned} \mathscr{R}_H \psi_j^H &= \sum_{k=1}^{\wp} \int_0^T \langle \psi_j^H, y^k(t) \rangle_H \, y^k(t) \, \mathrm{d}t \\ &= \sum_{i=1}^{d} \Big(\sum_{k=1}^{\wp} \int_0^T \langle \psi_j^H, y^k(t) \rangle_H \langle y^k(t), \psi_i^V \rangle_V \, \mathrm{d}t \Big) \psi_i^V, \quad j \in \mathscr{I}, \end{aligned}$$

we conclude that that the range of $\mathscr{R}_H$ is at most d-dimensional, which implies that $\lambda_i^H = 0$ for all $i > d$. Analogously, we deduce from $\sigma_i^H = 0$ for all $i > d$ that the range of $\mathscr{R}_V$ is at most d.

3) The claim follows directly from part 2).

Thus, Lemma 1.24 is proved. □

Let us define the two POD subspaces

$$V^\ell = \operatorname{span}\{\psi_1^V, \dots, \psi_\ell^V\} \subset V, \quad H^\ell = \operatorname{span}\{\psi_1^H, \dots, \psi_\ell^H\} \subset V \subset H,$$

where $H^\ell \subset V$ follows from part 1) of Lemma 1.24. Moreover, we introduce the orthogonal projection operators $\mathscr{P}_H^\ell : V \to H^\ell \subset V$ and $\mathscr{P}_V^\ell : V \to V^\ell \subset V$ as follows:

$$\begin{aligned} v^\ell = \mathscr{P}_H^\ell \varphi \text{ for any } \varphi \in V \quad &\text{iff } v^\ell \text{ solves } \min_{w^\ell \in H^\ell} \|\varphi - w^\ell\|_V, \\ v^\ell = \mathscr{P}_V^\ell \varphi \text{ for any } \varphi \in V \quad &\text{iff } v^\ell \text{ solves } \min_{w^\ell \in V^\ell} \|\varphi - w^\ell\|_V. \end{aligned} \tag{1.49}$$

It follows from the first-order optimality conditions that $v^\ell = \mathscr{P}_H^\ell \varphi$ satisfies

$$\langle v^\ell, \psi_i^H \rangle_V = \langle \varphi, \psi_i^H \rangle_V, \quad 1 \le i \le \ell. \tag{1.50}$$

Writing $v^\ell \in H^\ell$ in the form $v^\ell = \sum_{j=1}^{\ell} \mathrm{v}_j^\ell \psi_j^H$ we derive from (1.50) that the vector $\mathrm{v}^\ell = (\mathrm{v}_1^\ell, \dots, \mathrm{v}_\ell^\ell)^\top \in \mathbb{R}^\ell$ satisfies the linear system

$$\sum_{j=1}^{\ell} \langle \psi_j^H, \psi_i^H \rangle_V \, \mathrm{v}_j^\ell = \langle \varphi, \psi_i^H \rangle_V, \quad 1 \le i \le \ell.$$

For the operator $\mathscr{P}_V^\ell$ we have the explicit representation

$$\mathscr{P}_V^\ell \varphi = \sum_{i=1}^{\ell} \langle \varphi, \psi_i^V \rangle_V \, \psi_i^V \text{ for } \varphi \in V.$$

Since the linear operators $\mathscr{P}_V^\ell$ and $\mathscr{P}_H^\ell$ are orthogonal projections, we have $\|\mathscr{P}_V^\ell\|_{\mathscr{L}(V)} = \|\mathscr{P}_H^\ell\|_{\mathscr{L}(V)} = 1$. As $\{\psi_i^V\}_{i\in\mathscr{I}}$ is a complete orthonormal basis in V, we have

$$\lim_{\ell\to\infty}\int_0^T \|w(t)-\mathscr{P}_V^\ell w(t)\|_V^2\,\mathrm{d}t = 0 \quad \text{for all } w\in L^2(0,T;V). \tag{1.51}$$

Next we review an essential result from [66, Theorem 5.2], which we will use in our a-priori error analysis for the choice $X = H$. Recall that $\psi_i^H \in V$ holds for $1 \le i \le d_H$ and the image of $\mathscr{P}_H^\ell$ belongs to V. Consequently, $\|\psi_i^H - \mathscr{P}_H^\ell\psi_i^H\|_V$ is well-defined for $1 \le i \le d_H$.

Theorem 1.25. *Suppose that $y^k \in L^2(0,T;V)$ for $1 \le k \le \wp$. Then,*

$$\sum_{k=1}^{\wp}\int_0^T \|y^k(t)-\mathscr{P}_H^\ell y^k(t)\|_V^2\,\mathrm{d}t = \sum_{i=\ell+1}^{d_H}\lambda_i^H\,\|\psi_i^H-\mathscr{P}_H^\ell\psi_i^H\|_V^2. \tag{1.52}$$

Here, d_H is the rank of the operator $\mathscr{R}_H$, which may be infinite. Moreover, $\mathscr{P}_H^\ell y^k$ converges to y^k in $L^2(0,T;V)$ as ℓ tends to ∞ for each $k\in\{1,\dots,\wp\}$.

Proof. Suppose that $1 \le \ell \le d_H$ and $1 \le \ell_\circ < \infty$ hold. Then, $\lambda_i^H > 0$ for $1 \le i \le \ell$. Let $\mathscr{I} \in \mathscr{L}(V)$ denote the identity operator. As $\mathscr{I} - \mathscr{P}_H^\ell$ is an orthonormal projection on V, we conclude $\|\mathscr{I}-\mathscr{P}_H^\ell\|_{\mathscr{L}(V)} = 1$. Furthermore, $y^k \in L^2(0,T;V)$ holds for each $k\in\{1,\dots\wp\}$. Thus, (1.51) implies that $\mathscr{P}_V^{\ell_\circ} y^k \to y^k$ in $L^2(0,T;V)$ as $\ell_\circ\to\infty$ for each k. The proof of (1.52) is essentially based on Hilbert-Schmidt theory and on the following result [66, Lemma 5.1]:

$$\begin{aligned}&\sum_{k=1}^{\wp}\int_0^T \|(\mathscr{I}-\mathscr{P}_H^\ell)\mathscr{P}_V^{\ell_\circ}y^k(t))\|_V^2\,\mathrm{d}t\\ &\quad=\sum_{i=1}^{\ell_\circ}\lambda_i^V\,\|\psi_i^V-\mathscr{P}_H^\ell\psi_i^V\|_V^2 \le \sum_{i:\lambda_i^V>0}\lambda_i^V\,\|\psi_i^V-\mathscr{P}_H^\ell\psi_i^V\|_V^2 < \infty\end{aligned} \tag{1.53}$$

for any $\ell_\circ\in\mathbb{N}$. To prove that $\mathscr{P}_H^\ell y^k$ converges to y^k in $L^2(0,T;V)$ as ℓ tends to ∞ for each $k\in\{1,\dots,\wp\}$ we observe that

$$\begin{aligned}\sum_{i=\ell+1}^{d_H}\lambda_i^H\,\|\psi_i^H-\mathscr{P}_H^\ell\psi_i^H\|_V^2 &\le \sum_{i=\ell+1}^{d_H}\lambda_i^H\,\|\mathscr{I}-\mathscr{P}_H^\ell\|_{\mathscr{L}(V)}\|\psi_i^H\|_V^2\\ &= \sum_{i=\ell+1}^{d_H}\lambda_i^H\,\|\psi_i^H\|_V^2\end{aligned}$$

By utilizing the singular value decomposition (see Remark 1.17) it is shown in [66, Theorem 5.2] that $\sum_{i=\ell+1}^{d_H}\lambda_i^H\,\|\psi_i^H\|_V^2 < \infty$ holds. Therefore,

$$\lim_{\ell_\circ\to\infty}\sum_{k=1}^{\wp}\int_0^T \|(\mathscr{I}-\mathscr{P}_H^\ell)\mathscr{P}_V^{\ell_\circ}y^k(t))\|_V^2\,\mathrm{d}t = 0$$

which gives the claim. □

We will also need the following result, which follows from the continuous embedding $V \hookrightarrow H$. For a proof we refer to [66, Proposition 5.5].

Lemma 1.26. *Let $y^k \in L^2(0,T;V)$ for each $k \in \{1,\dots,\wp\}$ and $\lambda_i^H > 0$ for all $i \in \mathscr{I}$. Then,*

$$\lim_{\ell\to\infty} \|\varphi - \mathscr{P}_H^\ell \varphi\|_V = 0 \quad \textit{for all } \varphi \in V.$$

1.3.3 ▪ The POD Galerkin approximation

After the computation of a POD basis of rank ℓ we are interested in deriving a low-dimensional approximation for the evolution problem (1.44). In the context of Section 1.2.2 we choose $\wp = 1$, $y^1 = \mathscr{S}u$ and compute a POD basis $\{\psi_i\}_{i=1}^\ell$ of rank ℓ by solving ($\mathbf{P}^\ell$) with $\psi_i = \psi_i^V$ for $X = V$ and $\psi_i = \psi_i^H$ for $X = H$. Then, we define the subspace $X^\ell = \operatorname{span}\{\psi_1,\dots,\psi_\ell\}$, i.e., $X^\ell = V^\ell$ for $X = V$ and $X^\ell = H^\ell$ for $X = H$. Now we approximate the state variable y by the Galerkin expansion

$$y^\ell(t) = \hat{y}(t) + \sum_{i=1}^{\ell} \mathrm{y}_i^\ell(t)\psi_i \in V \quad \text{a.e. in } [0,T] \tag{1.54}$$

with coefficient functions $\mathrm{y}_i^\ell : [0,T] \to \mathbb{R}$. We introduce the vector-valued coefficient function

$$\mathrm{y}^\ell = \big(\mathrm{y}_1^\ell,\dots,\mathrm{y}_\ell^\ell\big) : [0,T] \to \mathbb{R}^\ell.$$

Since $\hat{y}(0) = y_\circ$ holds, we suppose that $\mathrm{y}^\ell(0) = 0$. Then, $y^\ell(0) = y_\circ$ is valid, i.e., the POD state matches exactly the initial condition. Inserting (1.54) into (1.44) and using the test space in V^ℓ for $1 \le i \le \ell$ we obtain the following POD Galerkin scheme for (1.44): $y^\ell \in W(0,T)$ solves

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}\langle y^\ell(t),\psi\rangle_H + a(t;y^\ell(t),\psi) = \langle (f+\mathscr{B}u)(t),\psi\rangle_{V',V} \ \ \forall \psi \in X^\ell \text{ a.e.},\\ &\mathrm{y}^\ell(0) = 0. \end{aligned} \tag{1.55}$$

We call (1.55) a *low dimensional* or *reduced-order model* for (1.44).

Proposition 1.27. *Let all assumptions of Theorem* 1.22 *be satisfied and the POD basis of rank ℓ be computed as desribed at the beginning of Section* 1.3.1. *Then, there exists a unique solution $y^\ell \in H^1(0,T;V) \hookrightarrow W(0,T)$ solving* (1.55).

Proof. Choosing $\psi = \psi_i$, $1 \le i \le \ell$, and applying (1.54) we infer from (1.55) that the coefficient vector y^ℓ satisfies

$$\mathrm{M}^\ell \dot{\mathrm{y}}^\ell(t) + \mathrm{A}^\ell(t)\mathrm{y}(t) = \hat{\mathrm{F}}^\ell(t) \text{ a.e. in } [0,T], \quad \mathrm{y}^\ell(0) = 0, \tag{1.56}$$

where we have set

$$\begin{aligned} &\mathrm{M}^\ell = \big(\big(\langle\psi_i,\psi_j\rangle_H\big)\big) \in \mathbb{R}^{\ell\times\ell}, \quad \mathrm{A}^\ell(t) = \big(\big(a(t;\psi_i,\psi_j)\big)\big) \in \mathbb{R}^{\ell\times\ell},\\ &\hat{\mathrm{F}}^\ell(t) = \big(\langle (f+\mathscr{B}u)(t) - \hat{y}_t(t),\psi_i\rangle_{V',V} - a(t;\hat{y}(t),\psi_i)\big) \in \mathbb{R}^\ell \end{aligned} \tag{1.57}$$

with $\psi_i = \psi_i^V$ for $X = V$ and $\psi_i = \psi_i^H$ for $X = H$. Since (1.56) is a linear ordinary differential equation system the existence of a unique $\mathrm{y}^\ell \in H^1(0,T;\mathbb{R}^\ell)$ follows by standard arguments. □

Remark 1.28.

1) Suppose $\hat{y} \neq 0$. In contrast to [28, 71], for instance, the POD approximation does admit values $y^\ell(t)$ in X^ℓ, but $(y^\ell - \hat{y})(t) \in X^\ell$ holds. The benefit of this approach is that $y^\ell(0) = y_\circ$ – and not $y^\ell(0) = \mathscr{P}_H^\ell y_\circ$ or $y^\ell(0) = \mathscr{P}_V^\ell y_\circ$. This improves the approximation quality of the POD basis which is illustrated in our numerical tests.

2) We proceed analogously to Remark 1.23 and introduce the linear and bounded solution operator $\mathscr{S}^\ell : U \to W_0(0,T)$: for $u \in U$ the function $w^\ell = \mathscr{S}^\ell u \in W(0,T)$ satisfies $w^\ell(0) = 0$ and

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle w^\ell(t), \psi\rangle_H + a(t; w^\ell(t), \psi) = \langle (\mathscr{B}u)(t), \psi\rangle_{V',V} \quad \forall \psi \in X^\ell \text{ a.e.}$$

Then, the solution to (1.55) is given by $y^\ell = \hat{y} + \mathscr{S}^\ell u$. Analogous to the proof of (1.45) we derive that there exists a positive constant C_2 which does not depend on ℓ or u so that

$$\|\mathscr{S}^\ell u\|_{W(0,T)} \leq C \|u\|_U.$$

Thus, $\mathscr{S}^\ell$ is bounded uniformly with respect to ℓ. ◊

To investigate the convergence of the error $y - y^\ell$ we make use of the following two inequalities:

1) *Gronwall's inequality*: For $T > 0$ let $v : [0,T] \to \mathbb{R}$ be a nonnegative, differentiable function satisfying

$$v'(t) \leq \varphi(t)v(t) + \chi(t) \quad \text{for all } t \in [0,T],$$

where φ and χ are real-valued, nonnegative, integrable functions on $[0,T]$. Then

$$v(t) \leq \exp\left(\int_0^t \varphi(s)\,\mathrm{d}s\right)\left(v(0) + \int_0^t \chi(s)\,\mathrm{d}s\right) \quad \text{for all } t \in [0,T]. \tag{1.58}$$

In particular, if

$$v' \leq \varphi v \text{ in } [0,T] \quad \text{and} \quad v(0) = 0$$

hold, then $v = 0$ in $[0,T]$.

2) *Young's inequality*: For every $a, b \in \mathbb{R}$ and for every $\varepsilon > 0$ we have

$$ab \leq \frac{\varepsilon a^2}{2} + \frac{b^2}{2\varepsilon}.$$

Theorem 1.29. *Let $u \in U$ be chosen arbitrarily so that $0 \neq \mathscr{S}u \in H^1(0,T;V)$.*

1) *To compute a POD basis $\{\psi_i\}_{i=1}^\ell$ of rank ℓ we choose $\wp = 1$ and $y^1 = \mathscr{S}u$. Then, $y = \hat{y} + \mathscr{S}u$ and $y^\ell = \hat{y} + \mathscr{S}^\ell u$ satisfies the a-priori error estimate*

$$\begin{aligned} &\|y^\ell - y\|^2_{H^1(0,T;V)} \\ &\leq C_1 \cdot \begin{cases} \displaystyle\sum_{i=\ell+1}^{d_V} \lambda_i^V + \|y_t^1 - \mathscr{P}_V^\ell y_t^1\|^2_{L^2(0,T;V)} & \text{if } X = V, \\ \displaystyle\sum_{i=\ell+1}^{d_H} \lambda_i^H \|\psi_i^H - \mathscr{P}_H^\ell \psi_i^H\|_V^2 & \\ \qquad + \|y_t^1 - \mathscr{P}_H^\ell y_t^1\|^2_{L^2(0,T;V)} & \text{if } X = H, \end{cases} \end{aligned} \tag{1.59}$$

where the constant C_1 depends on the terminal time T and the constants γ, γ_1, γ_2 introduced in (1.43).

2) *If we set $\wp = 2$ and compute a POD basis of rank ℓ using the trajectories $y^1 = \mathscr{S} u$ and $y^2 = (\mathscr{S} u)_t$, it follows that*

$$\begin{aligned} &\|y^\ell - y\|^2_{H^1(0,T;V)} \\ &\quad \le C_3 \cdot \begin{cases} \sum\limits_{i=\ell+1}^{d_V} \lambda_i^V & \text{for } X = V, \\ \sum\limits_{i=\ell+1}^{d_H} \lambda_i^H \|\psi_i^H - \mathscr{P}_H^\ell \psi_i^H\|_V^2 & \text{for } X = H \end{cases} \end{aligned} \tag{1.60}$$

for a constant C_3 which depends on γ, γ_1, γ_2, and T.

3) *If $\mathscr{S}\tilde{u}$ belongs to $H^1(0,T;V)$ for every $\tilde{u} \in U$ and if $\lambda_i^H > 0$ for all $i \in \mathscr{I}$, then we have*

$$\lim_{\ell \to \infty} \|\mathscr{S} - \mathscr{S}^\ell\|_{\mathscr{L}(U, W(0,T))} = 0. \tag{1.61}$$

Proof.

1) For almost all $t \in [0,T]$ we make use of the decomposition

$$\begin{aligned} y^\ell(t) - y(t) &= \hat{y}(t) + (\mathscr{S}^\ell u)(t) - \hat{y}(t) - (\mathscr{S} u)(t) \\ &= (\mathscr{S}^\ell u)(t) - \mathscr{P}^\ell\big((\mathscr{S} u)(t)\big) + \mathscr{P}^\ell\big((\mathscr{S} u)(t)\big) - (\mathscr{S} u)(t) \\ &= \vartheta^\ell(t) + \varrho^\ell(t), \end{aligned} \tag{1.62}$$

where $\vartheta^\ell = \mathscr{S}^\ell u - \mathscr{P}^\ell(\mathscr{S} u) \in X^\ell$ and $\varrho^\ell = \mathscr{P}^\ell(\mathscr{S} u) - \mathscr{S} u$. In (1.62) we will consider the two choices $\mathscr{P}^\ell = \mathscr{P}_H^\ell$ for $X = H$ and $\mathscr{P}^\ell = \mathscr{P}_V^\ell$ for $X = V$. From $y^1 = \mathscr{S} u$ and (1.27) we infer that

$$\|\varrho^\ell\|^2_{H^1(0,T;V)} = \sum_{i=\ell+1}^{d_V} \lambda_i^V + \|y_t^1 - \mathscr{P}_V^\ell y_t^1\|^2_{L^2(0,T;V)} \tag{1.63}$$

in case of $X = V$, where d_V stands for rank of $\mathscr{R}_V$. For the choice $X = H$ we derive from $y^1 = \mathscr{S} u$ and Theorem 1.25 that

$$\|\varrho^\ell\|^2_{H^1(0,T;V)} = \sum_{i=\ell+1}^{d_H} \lambda_i^H \|\psi_i^H - \mathscr{P}_H^\ell \psi_i^H\|_V^2 + \|y_t^1 - \mathscr{P}_H^\ell y_t^1\|^2_{L^2(0,T;V)}. \tag{1.64}$$

Here, d_H denotes for rank of $\mathscr{R}_H$. Using $\vartheta_t^\ell(t) \in H$ for almost all $t \in [0,T]$, (1.44), (1.55) and (1.43a) we derive that

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t} \langle \vartheta^\ell(t), \psi \rangle_H + a(t; \vartheta^\ell(t), \psi) \\ &\quad = \langle y_t^1(t) - \mathscr{P}^\ell y_t^1(t), \psi \rangle_H + a(t; y^1(t) - \mathscr{P}^\ell y^1(t), \psi) \\ &\quad \le \|y_t^1(t) - \mathscr{P}^\ell y_t^1(t)\|_H \|\psi\|_H + \gamma \|y^1(t) - \mathscr{P}^\ell y^1(t)\|_V \|\psi\|_V \end{aligned} \tag{1.65}$$

for all $\psi \in X^\ell$ and for almost all $t \in [0,T]$. From choosing $\psi = \vartheta^\ell(t)$, (1.43b) and (1.65) we find

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\,||\vartheta^\ell(t)||_H^2 + \gamma_1\,||\vartheta^\ell(t)||_V^2 - 3\gamma_2\,||\vartheta^\ell(t)||_H^2\\ &\quad\le \frac{1}{\gamma_2}\,||y_t^1(t) - \mathscr{P}^\ell y_t^1(t)||_H^2 + \frac{\gamma^2}{\gamma_1}\,||y^1(t) - \mathscr{P}^\ell y^1(t)||_V^2.\end{aligned}$$

From (1.58) – setting $v(t) = ||\vartheta^\ell(t)||_H^2 \ge 0$,

$$\chi(t) = \frac{1}{\gamma_2}\,||y_t^1(t) - \mathscr{P}^\ell y_t^1(t)||_H^2 + \frac{\gamma^2}{\gamma_1}\,||y^1(t) - \mathscr{P}^\ell y^1(t)||_V^2 \ge 0,$$

$\varphi(t) = 3\gamma_2 > 0$ – and $\vartheta^\ell(0) = 0$ it follows that

$$||\vartheta^\ell(t)||_H^2 \le c_1\big(||y_t^1 - \mathscr{P}^\ell y_t^1||_{L^2(0,T;H)}^2 + ||y^1 - \mathscr{P}^\ell y^1||_{L^2(0,T;V)}^2\big)$$

for almost all $t \in [0,T]$ with the constants $c_1 = c_2 \exp(3\gamma_2 T)$ and $c_2 = \max(1/\gamma_2, \gamma^2/\gamma_1)$, so that

$$\begin{aligned}||\vartheta^\ell||_{L^2(0,T;V)}^2 &\le c_3\Big(||\vartheta^\ell||_{L^2(0,T;H)}^2 + ||y_t^1 - \mathscr{P}^\ell y_t^1||_{L^2(0,T;H)}^2\Big)\\ &\quad + c_3\,||y^1 - \mathscr{P}^\ell y^1||_{L^2(0,T;V)}^2\\ &\le c_4\Big(||y_t^1 - \mathscr{P}^\ell y_t^1||_{L^2(0,T;H)}^2 + ||\varrho^\ell(t)||_{L^2(0,T;V)}^2\Big)\end{aligned} \tag{1.66}$$

with $c_3 = \max(3\gamma_2, c_2)/\gamma_1$ and $c_4 = c_3(1 + c_1 T)$. We conclude from (1.43a), (1.59) and (1.66) that

$$\begin{aligned}&||\vartheta_t^\ell||_{L^2(0,T;(V^\ell)')}\\ &= \sup\left\{\int_0^T \langle\vartheta_t^\ell(t), \psi(t)\rangle_{V',V}\,\middle|\,||\psi||_{L^2(0,T;V)} = 1,\ \psi(t) \in V^\ell\right\}\\ &\le \gamma\,||\vartheta^\ell||_{L^2(0,T;V)} + ||y_t^1 - \mathscr{P}^\ell y_t^1||_{L^2(0,T;H)}\\ &\le c_5\Big(||y_t^1 - \mathscr{P}^\ell y_t^1||_{L^2(0,T;H)} + ||y^1 - \mathscr{P}^\ell y^1||_{L^2(0,T;V)}\Big)\end{aligned} \tag{1.67}$$

with $c_5 = 1 + c_4\gamma$. By assumption we have $\vartheta_t^\ell \in L^2(0,T;V)$. Then, by the Riesz theorem [60, p. 43] we have

$$||\vartheta_t^\ell||_{L^2(0,T;(V^\ell)')} = ||\vartheta_t^\ell||_{L^2(0,T;V^\ell)} = ||\vartheta_t^\ell||_{L^2(0,T;V)}$$

so that (1.62)-(1.64), (1.66) and (1.67) imply (1.59).

2) The claim follows directly from

$$||y_t^1 - \mathscr{P}^\ell y_t^1||_{L^2(0,T;V)}^2 = ||y^2 - \mathscr{P}^\ell y^2||_{L^2(0,T;V)}^2$$

(1.27) and Theorem 1.25.

3) Recall that the space $H^1(0,T;V)$ is continuously embedded in $W(0,T)$. Therefore, there is a constant $c_W > 0$ satisfying

$$\|\varphi\|_{W(0,T)} \le c_W \|\varphi\|_{H^1(0,T;V)} \quad \text{for all } \varphi \in H^1(0,T;V).$$

Using $\mathcal{S}\tilde{u} \in H^1(0,T;V)$ for any $\tilde{u} \in U$, Remark 1.9 and applying the arguments as in part 1) we infer that there exists a constant c_6 which is independent of ℓ satisfying

$$\begin{aligned}
&\left\|\mathcal{S} - \mathcal{S}^\ell\right\|_{\mathcal{L}(U,W(0,T))} \\
&\quad = \sup_{\|\tilde{u}\|_U=1} \left\|(\mathcal{S}-\mathcal{S}^\ell)\tilde{u}\right\|_{W(0,T)} \le c_W \sup_{\|\tilde{u}\|_U=1} \left\|(\mathcal{S}-\mathcal{S}^\ell)\tilde{u}\right\|_{H^1(0,T;V)} \\
&\quad \le c_6 \sup_{\|\tilde{u}\|_U=1} \int_0^T \|\tilde{y}(t)-\mathcal{P}^\ell\tilde{y}(t)\|_V^2 + \|\tilde{y}_t(t)-\mathcal{P}^\ell\tilde{y}_t(t)\|_V^2 \,\mathrm{d}t
\end{aligned}$$

with $\tilde{y} = \mathcal{S}\tilde{u}$. By assumption, the elements $\tilde{y}(t)$ and $\tilde{y}_t(t)$ belong to the space $L^2(0,T;V)$. Therefore, the claim follows for $X = V$ from (1.51) and for $X = H$ from Lemma 1.26.

Thus, Theorem 1.29 is proved. □

Remark 1.30.

1) Note that the a-priori error estimates (1.59) and (1.60) depend an the arbitrarily chosen, but fixed control $u \in U$, which is also utilized to compute the POD basis. Moreover, these a-priori estimates do not involve errors by the POD discretization of the initial condition $y_\circ$ – in contrast to the error analysis presented in [28, 38, 40, 62, 71], for instance. Further, let us mention that the a-priori error analysis holds for $T < \infty$.

2) From (1.61) we infer

$$\left\|\hat{y} + \mathcal{S}^\ell\tilde{u} - \hat{y} - \mathcal{S}\tilde{u}\right\|_{W(0,T)} \le \left\|\mathcal{S}-\mathcal{S}^\ell\right\|_{\mathcal{L}(U,W(0,T))} \|\tilde{u}\|_U \xrightarrow{\ell\to\infty} 0$$

for any $\tilde{u} \in U$.

3) For the numerical realization we have to utilize also a time integration method like, e.g., the implicit Euler or the Crank-Nicolson method. We refer the reader to [38, 40, 39], where different time discretization schemes are considered. Moreover, in [45, 62] also a finite element discretization of the ansatz space V is incorporated in the a-priori error analysis. ◊

Example 1.31. Accurate approximation results are achieved if the subspace spanned by the snapshots is (approximatively) of low dimension. Let $T > 0$, $\Omega = (0,2) \subset \mathbb{R}$ and $Q = (0,T) \times \Omega$. We set $f(t,x) = e^{-t}(\pi^2 - 1)\sin(\pi x)$ for $(t,x) \in Q$ and $y_\circ(x) = \sin(\pi x)$ for $x \in \Omega$. Let $H = L^2(\Omega)$, $V = H_0^1(\Omega)$ and

$$a(t;\varphi,\phi) = \int_\Omega \varphi'(x)\phi'(x)\,\mathrm{d}x \quad \text{for } \varphi,\phi \in V,$$

i.e., the bilinear form a is independent of t. Finally, we choose $u = 0$. Then, the exact solution to (1.44) is given by $y(t,x) = e^{-t}\sin(\pi x)$. Thus, the snapshot space $\mathcal{V}$ is the

one-dimensional space $\{\alpha\psi \mid \alpha \in \mathbb{R}\}$ with $\psi(x) = \sin(\pi x)$. Choosing the space $X = H$, this implies that all eigenvalues of the operator $\mathscr{R}_H$ introduced in (1.46) except of the first one are zero and $\psi_1 = \psi \in V$ is the single POD element corresponding to a nontrivial eigenvalue of $\mathscr{R}_H$. Further, the reduced order model of the rank-1 POD-Galerkin ansatz

$$\begin{aligned} \dot{\mathrm{y}}^1(t) + \|\psi_1'\|_H^2 \, \mathrm{y}^1(t) &= \langle f(t), \psi_1 \rangle_H \quad \text{for } t \in (0, T], \\ \mathrm{y}^1(0) &= \langle y_\circ, \psi_1 \rangle_H \end{aligned}$$

has the solution $\mathrm{y}^1(t) = e^{-t}$, so both the projection

$$\big(\mathscr{P}^1 y\big)(t, x) = \langle y(t), \psi_1 \rangle_X \psi_1(x), \quad (t, x) \in \overline{Q},$$

of the state y on the POD-Galerkin space and the reduced-order solution $y^1(t) = \mathrm{y}^1(t)\psi_1$ coincide with the exact solution y. In the latter case, this is due to the fact that the data functions f and $y_\circ$ as well as all time derivative snapshots $\dot{y}(t)$ are already elements of $\operatorname{span}(\psi_1)$, so no projection error occurs here, cp. the a priori error bounds given in (1.60). In the case $X = V$, we get the same results with $\psi_1(x) = \sin(\pi x)/\sqrt{1+\pi^2}$ and $\mathrm{y}^1(t) = \sqrt{1+\pi^2} e^{-t}$. ∎

Utilizing the techniques as in the proof of Theorem 6.5 in [66] one can derive an a-priori error bound without including the time derivatives into the snapshot subspace. In the next proposition we formulate the a-priori error estimate.

Proposition 1.32. *Let $y_\circ \in V$ and $u \in U$ be chosen arbitrarily so that $\mathscr{S} u \neq 0$. To compute a POD basis $\{\psi_i\}_{i=1}^{\ell}$ of rank ℓ we choose $\wp = 1$ and $y^1 = \mathscr{S} u$. Then, $y = \hat{y} + \mathscr{S} u$ and $y^\ell = \hat{y} + \mathscr{S}^\ell u$ satisfies the a-priori error estimate*

$$\|y^\ell - y\|_{L^2(0,T;V)}^2 \leq C \cdot \begin{cases} \displaystyle\sum_{i=\ell+1}^{d_V} \lambda_i^V \, \|\psi_i^V - \mathscr{P}_{H,V^\ell}^{\ell} \psi_i^V\|_V^2 & \text{if } X = V, \\ \displaystyle\sum_{i=\ell+1}^{d_H} \lambda_i^H \, \|\psi_i^H\|_V^2 & \text{if } X = H, \end{cases} \tag{1.68}$$

where the constant C depends on the terminal time T and the constants γ, γ_1, γ_2 introduced in (1.43)*. Moreover, $\mathscr{P}_{H,V^\ell}^{\ell} : H \to V^\ell$ is the H-orthogonal projection given as follows:*

$$v^\ell = \mathscr{P}_{H,V^\ell}^{\ell} \varphi \text{ for any } \varphi \in H \text{ iff } v^\ell \text{ solves } \min_{w^\ell \in V^\ell} \|\varphi - w^\ell\|_H.$$

In particular, we have $y^\ell \to y$ in $L^2(0, T; V)$ as $\ell \to \infty$.

1.4 ▪ The linear-quadratic optimal control problem

In this section we apply a POD Galerkin approximation to linear-quadratic optimal control problems. Linear-quadratic problems are interesting in several respects: In particular, they occur in each level of a sequential quadratic programming (SQP) methods; see, e.g., [52]. In contrast to methods of balanced truncation type, the POD method is somehow lacking a reliable a-priori error analysis. Unless its snapshots are generating a sufficiently rich state space, it is not a-priorly clear how far the optimal solution of the POD problem is from the exact one. On the other hand, the POD method is a universal tool that is applicable also to problems with time-dependent coefficients or to nonlinear equations.

By generating snapshots from the real (large) model, a space is constructed that contains the main and relevant physical properties of the state system. This, and its ease of use makes POD very competitive in practical use, despite of certain heuristic.

Here we prove convergence and derive a-priori error estimates for the optimal control problem. The error estimates rely on the (unrealistic) assumption that the POD basis is computed from the (exact) optimal solution. However, these estimates are utilized to develop an a-posteriori error analysis for the POD Galerkin appproximation of the optimal control problem. Using a perturbation method [16] we deduce how far the suboptimal control, computed by the POD Galerkin approximation, is from the (unknown) exact one. This idea turned out to be very efficient in our numerical examples. Thus, we are able to compensate for the lack of an a-priori error analysis for the POD method.

1.4.1 ▪ Problem formulation

In this section we introduce our optimal control problem, which is an constrained optimization problem in a Hilbert space. The objective is a quadratic function. The evolution problem (1.44) serves as an equality constraint. Moreover, bilateral control bounds lead to inequality constraints in the minimization. For the readers' convenience we recall (1.44) here. Let $U = L^2(0,T;\mathbb{R}^{N_u})$ denote the control space with $N_u \in \mathbb{N}$. For $u \in U$, $y_\circ \in H$ and $f \in L^2(0,T;V')$ we consider the state equation

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\langle y(t),\varphi\rangle_H + a(t;y(t),\varphi) &= \langle (f+\mathscr{B}u)(t),\varphi\rangle_{V',V} \\ &\qquad \forall \varphi \in V \text{ a.e. in } (0,T], \\ \langle y(0),\varphi\rangle_H &= \langle y_\circ,\varphi\rangle_H \qquad \forall \varphi \in H, \end{aligned} \tag{1.69}$$

where $\mathscr{B} : U \to L^2(0,T;V')$ is a continuous, linear operator. Due to Theorem 1.22 there exists a unique solution $y \in W(0,T)$ to (1.69).

We introduce the Hilbert space

$$X = W(0,T) \times U$$

endowed with the natural product topology, i.e., with the inner product

$$\langle x,\tilde{x}\rangle_X = \langle y,\tilde{y}\rangle_{W(0,T)} + \langle u,\tilde{u}\rangle_U \quad \text{for } x=(y,u),\ \tilde{x}=(\tilde{y},\tilde{u}) \in X$$

and the norm $\|x\|_X = (\|y\|^2_{W(0,T)} + \|u\|^2_U)^{1/2}$ for $x=(y,u)\in X$.

Assumption 1. *For $t \in [0,T]$ let $a(t;\cdot,\cdot) : V \times V \to \mathbb{R}$ be a time-dependent symmetric bilinear form satisfying* (1.43)*. Moreover, $f \in L^2(0,T;V')$, $y_\circ \in H$ and $\mathscr{B} \in \mathscr{L}(U, L^2(0,T;V'))$ holds.*

In Remark 1.23 we have introduced the particular solution $\hat{y} \in W(0,T)$ as well as the linear, bounded solution operator $\mathscr{S}$. Then, the solution to (1.69) can be expressed as $y = \hat{y} + \mathscr{S}u$. By X_{ad} we denote the closed, convex and bounded set of admissible solutions for the optimization problem as

$$X_{\mathsf{ad}} = \big\{(\hat{y}+\mathscr{S}u,u) \in X \,\big|\, u_a \le u \le u_b \text{ in } \mathbb{R}^{N_u} \text{ a.e. in } [0,T]\big\},$$

where $u_a = (u_{a,1},\dots,u_{a,N_u}), u_b = (u_{b,1},\dots,u_{b,N_u}) \in U$ satisfy $u_{a,i} \le u_{b,i}$ for $1 \le i \le N_u$ a.e. in $[0,T]$. Since $u_{a,i} \le u_{b,i}$ holds for $1 \le i \le N_u$, we infer from Theorem 1.22 that the set X_{ad} is nonempty.

The quadratic objective $J : X \to \mathbb{R}$ is given by

$$J(x) = \frac{\sigma_Q}{2} \int_0^T \|y(t) - y_Q(t)\|_H^2 \, \mathrm{d}t + \frac{\sigma_\Omega}{2} \|y(T) - y_\Omega\|_H^2 + \frac{\sigma}{2} \|u\|_U^2 \tag{1.70}$$

for $x = (y, u) \in X$, where $(y_Q, y_\Omega) \in L^2(0, T; H) \times H$ are given desired states. Furthermore, σ_Q, $\sigma_\Omega \geq 0$ and $\sigma > 0$. Of course, more general cost functionals can be treated analogously.

Now the quadratic programming problem is given by

$$\min J(x) \quad \text{subject to (s.t.)} \quad x \in X_{\mathsf{ad}}. \tag{P}$$

From $x = (y, u) \in X_{\mathsf{ad}}$ we infer that $y = \hat{y} + \mathscr{S} u$ holds. Hence, y is a dependent variable. We call u the *control* and y the *state*. In this way, (**P**) becomes an *optimal control problem*. Utilizing the relationship $y = \hat{y} + \mathscr{S} u$ we define a so-called *reduced cost functional* $\hat{J} : U \to \mathbb{R}$ by

$$\hat{J}(u) = J(\hat{y} + \mathscr{S} u, u) \quad \text{for } u \in U.$$

Moreover, the set of admissible controls is given as

$$U_{\mathsf{ad}} = \left\{ u \in U \,\middle|\, u_a \leq u \leq u_b \text{ in } \mathbb{R}^{N_u} \text{ a.e. in } [0, T] \right\},$$

which is convex, closed and bounded in U. Then, we consider the reduced optimal control problem:

$$\min \hat{J}(u) \quad \text{s.t.} \quad u \in U_{\mathsf{ad}}. \tag{$\hat{\mathbf{P}}$}$$

Clearly, if $\bar{u}$ is the optimal solution to ($\hat{\mathbf{P}}$), then $\bar{x} = (\hat{y} + \mathscr{S} \bar{u}, \bar{u})$ is the optimal solution to (**P**). On the other hand, if $\bar{x} = (\bar{y}, \bar{u})$ is the solution to (**P**), then $\bar{u}$ solves ($\hat{\mathbf{P}}$).

Example 1.33. We introduce an example for (**P**) and discuss the presented theory for this application. Let $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, be an open and bounded domain with Lipschitz-continuous boundary $\Gamma = \partial\Omega$. For $T > 0$ we set $Q = (0, T) \times \Omega$ and $\Sigma = (0, T) \times \Gamma$. We choose $H = L^2(\Omega)$ and $V = H_0^1(\Omega)$ endowed with the usual inner products

$$\langle \varphi, \psi \rangle_H = \int_\Omega \varphi \psi \, \mathrm{d}\boldsymbol{x}, \quad \langle \varphi, \psi \rangle_V = \int_\Omega \varphi \psi + \nabla \varphi \cdot \nabla \psi \, \mathrm{d}\boldsymbol{x}$$

and their induced norms, respectively. Let $\chi_i \in H$, $1 \leq i \leq m$, denote given control shape functions. Then, for given control $u \in U$, initial condition $y_\circ \in H$ and inhomogeneity $f \in L^2(0, T; H)$ we consider the linear heat equation

$$\begin{aligned} y_t(t, \boldsymbol{x}) - \Delta y(t, \boldsymbol{x}) &= f(t, \boldsymbol{x}) + \sum_{i=1}^{N_u} u_i(t) \chi_i(\boldsymbol{x}), && \text{a.e. in } Q, \\ y(t, \boldsymbol{x}) &= 0, && \text{a.e. in } \Sigma, \\ y(0, \boldsymbol{x}) &= y_\circ(\boldsymbol{x}), && \text{a.e. in } \Omega. \end{aligned} \tag{1.71}$$

We introduce the time-independent, symmetric bilinear form

$$a(\varphi, \psi) = \int_\Omega \nabla \varphi \cdot \nabla \psi \, \mathrm{d}\boldsymbol{x} \quad \text{for } \varphi, \psi \in V$$

and the bounded, linear operator $\mathscr{B} : U \to L^2(0,T;H) \hookrightarrow L^2(0,T;V')$ as

$$(\mathscr{B}u)(t,x) = \sum_{i=1}^{m} u_i(t)\chi_i(x) \quad \text{for } (t,x) \in Q \text{ a.e. and } u \in U.$$

Hence, we have $\gamma = \gamma_1 = \gamma_2 = 1$ in (1.43). It follows that the weak formulation of (1.71) can be expressed in the form (1.44). Moreover, the unique weak solution to (1.71) belongs to the space $L^\infty(0,T;V)$ provided $y_\circ \in V$ holds. ∎

1.4.2 ▪ Existence of a unique optimal solution

We suppose the following hypothesis for the objective.

Assumption 2. *In* (1.70) *the desired states* (y_Q, y_Ω) *belong to* $L^2(0,T;H) \times H$. *Furthermore,* $\sigma_Q, \sigma_\Omega \geq 0$ *and* $\sigma > 0$ *are satisfied.*

Let us review the following result for quadratic optimization problems in Hilbert spaces; see [70, pp. 50-51].

Theorem 1.34. *Suppose that* $\mathscr{U}$ *and* $\mathscr{H}$ *are given Hilbert spaces with norms* $\|\cdot\|_{\mathscr{U}}$ *and* $\|\cdot\|_{\mathscr{H}}$, *respectively. Furthermore, let* $\mathscr{U}_{\mathsf{ad}} \subset \mathscr{U}$ *be non-empty, bounded, closed, convex and* $z_d \in \mathscr{H}$, $\varkappa \geq 0$. *The mapping* $\mathscr{G} : \mathscr{U} \to \mathscr{H}$ *is assumed to be a linear and continuous operator. Then there exists an optimal control* $\bar{u}$ *solving*

$$\min_{u \in \mathscr{U}_{\mathsf{ad}}} \mathscr{J}(u) := \frac{1}{2}\|\mathscr{G}u - z_d\|_{\mathscr{H}}^2 + \frac{\varkappa}{2}\|u\|_{\mathscr{U}}^2. \tag{1.72}$$

If $\varkappa > 0$ *holds or if* $\mathscr{G}$ *is injective, then* $\bar{u}$ *is uniquely determined.*

Remark 1.35. In the proof of Theorem 1.34 it is only used that $\mathscr{J}$ is continuous and convex. Therefore, the existence of an optimal control follows for general convex, continuous cost functionals $\mathscr{J} : \mathscr{U} \to \mathbb{R}$ with a Hilbert space $\mathscr{U}$. ◊

Next we can use Theorem 1.34 to obtain an existence result for the optimal control problem ($\hat{\mathbf{P}}$), which imply the existence of an optimal solution to (**P**).

Theorem 1.36. *Let Assumptions* 1 *and* 2 *be valid. Moreover, let the bilateral control constraints* $u_a, u_b \in U$ *satisfy* $u_a \leq u_b$ *componentwise in* $\mathbb{R}^{N_u}$ *a.e. in* $[0,T]$. *Then,* ($\hat{\mathbf{P}}$) *has a unique optimal solution* $\bar{u}$.

Proof. Let us choose the Hilbert spaces $\mathscr{H} = L^2(0,T;H) \times H$ and $\mathscr{U} = U$. Moreover, $\mathscr{E} : W(0,T) \to L^2(0,T;H)$ is the canonical embedding operator, which is linear and bounded. We define the operator $\mathscr{E}_2 : W(0,T) \to H$ by $\mathscr{E}_2\varphi = \varphi(T)$ for $\varphi \in W(0,T)$. Since $W(0,T)$ is continuously embedded into $C([0,T];H)$, the linear operator $\mathscr{E}_2$ is continuous. Finally, we set

$$\mathscr{G} = \begin{pmatrix} \sqrt{\sigma_Q}\,\mathscr{E}_1\mathscr{S} \\ \sqrt{\sigma_\Omega}\,\mathscr{E}_2\mathscr{S} \end{pmatrix} \in \mathscr{L}(\mathscr{U},\mathscr{H}), \quad z_d = \begin{pmatrix} \sqrt{\sigma_Q}(y_Q - \hat{y}) \\ \sqrt{\sigma_\Omega}\big(y_\Omega - \hat{y}(T)\big) \end{pmatrix} \in \mathscr{H} \tag{1.73}$$

and $\mathscr{U}_{\mathsf{ad}} = U_{\mathsf{ad}}$. Then, ($\hat{\mathbf{P}}$) and (1.72) coincide. Consequently, the claim follows from Theorem 1.34 and $\sigma > 0$. □

Next we consider the case that $u_a = -\infty$ or/and $u_b = +\infty$. In this case U_{ad} is not bounded. However, we have the following result [70, p. 52].

Theorem 1.37. *Let Assumptions 1 and 2 be satisfied. If $u_a = -\infty$ or/and $u_b = +\infty$, problem ($\hat{\mathbf{P}}$) admits a unique solution.*

Proof. We utilize the setting of the proof of Theorem 1.36. By assumption there exists an element $u_0 \in U_{\mathsf{ad}}$. For $u \in U$ with $\|u\|_U^2 > 2\hat{J}(u_0)/\sigma$ we have

$$\hat{J}(u) = \mathcal{J}(u) = \frac{1}{2}\|\mathcal{G}u - z_d\|_{\mathcal{H}}^2 + \frac{\sigma}{2}\|u\|_U^2 \geq \frac{\sigma}{2}\|u\|_U^2 > \hat{J}(u_0).$$

Thus, the minimization of $\hat{J}$ over U_{ad} is equivalent with the minimization of $\hat{J}$ over the bounded, convex and closed set

$$U_{\mathsf{ad}} \cap \left\{ u \in U \,\middle|\, \|u\|_U^2 \leq \frac{2\hat{J}(u_0)}{\sigma} \right\}.$$

Now the claim follows from Theorem 1.34. □

1.4.3 ▪ First-order necessary optimality conditions

In (1.72) we have introduced the quadratic programming problem

$$\min_{u \in \mathcal{U}_{\mathsf{ad}}} \mathcal{J}(u) = \frac{1}{2}\|\mathcal{G}u - z_d\|_{\mathcal{H}}^2 + \frac{\sigma}{2}\|u\|_{\mathcal{U}}^2. \tag{1.74}$$

Existence of a unique solution has been investigated in Section 1.4.2. In this section we characterize the solution to (1.74) by first-order optimality conditions, which are essential to prove convergence and rate of convergence results for the POD approximations in Section 1.4.4. To derive first-order conditions we require the notion of derivatives in function spaces. Therefore, we recall the following definition [70, pp. 56-57].

Definition 1.38. *Suppose that $\mathcal{B}_1$ and $\mathcal{B}_2$ are real Banach spaces, $\mathcal{U} \subset \mathcal{B}_1$ be an open subset and $\mathcal{F} : \mathcal{U} \supset \mathcal{B}_1 \to \mathcal{B}_2$ a given mapping. The of $\mathcal{F}$ at a point $u \in \mathcal{U}$ in the direction $h \in \mathcal{B}_2$ is defined by*

$$D\mathcal{F}(u;h) := \lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon}\big(\mathcal{F}(u + \varepsilon h) - \mathcal{F}(u)\big)$$

provided the limit exists in $\mathcal{B}_2$. Suppose that the directional derivative exists for all $h \in \mathcal{B}_1$ and there is a linear, continuous operator $\mathcal{T} : \mathcal{U} \to \mathcal{B}_2$ satisfying

$$D\mathcal{F}(u;h) = \mathcal{T}h \quad \text{for all } h \in \mathcal{U}.$$

Then, $\mathcal{F}$ is said to be -differentiable *at u and $\mathcal{T}$ is the* Gâteaux derivative *of $\mathcal{F}$ at u. We write $\mathcal{T} = \mathcal{F}'(u)$.*

Remark 1.39. Let $\mathcal{H}$ be a real Hilbert space and $\mathcal{F} : \mathcal{H} \to \mathbb{R}$ be *Gâteaux-differentiable* at $u \in \mathcal{H}$. Then, its Gâteaux derivative $\mathcal{F}'(u)$ at u belongs to $\mathcal{H}' = \mathcal{L}(\mathcal{H}, \mathbb{R})$. Due to Riesz theorem [60, p. 43] there exists a unique element $\nabla\mathcal{F}(u) \in \mathcal{H}$ satisfying

$$\langle \nabla\mathcal{F}(u), v \rangle_{\mathcal{H}} = \langle \mathcal{F}'(u), v \rangle_{\mathcal{H}',\mathcal{H}} \quad \text{for all } v \in \mathcal{H}.$$

We call $\nabla \mathscr{F}(u)$ the *(Gâteaux) gradient* of $\mathscr{F}$ at u. $\Diamond$

Theorem 1.40. *Let $\mathscr{U}$ be a real Hilbert space and $\mathscr{U}_{\mathsf{ad}}$ be convex subset. Suppose that $\bar{u} \in \mathscr{U}_{\mathsf{ad}}$ is a solution to* (1.74)

$$\min_{u \in \mathscr{U}_{\mathsf{ad}}} \mathscr{J}(u).$$

Then the following variational inequality holds

$$\langle \nabla \mathscr{J}(\bar{u}), u - \bar{u} \rangle_{\mathscr{U}} \geq 0 \quad \text{for all } u \in \mathscr{U}_{\mathsf{ad}}, \tag{1.75}$$

where the gradient of $\mathscr{J}$ is given by

$$\nabla \mathscr{J}(u) = \mathscr{G}^{\star}(\mathscr{G}u - z_d) + \sigma u \quad \text{for } u \in \mathscr{U}.$$

If $\bar{u} \in \mathscr{U}_{\mathsf{ad}}$ solves (1.75)*, then $\bar{u}$ is a solution to* (1.74)*.*

Proof. Since $\mathscr{J}$ is Gâteaux-differentiable and convex in $\mathscr{U}$, the result follows directly from [70, pp. 63-63]. □

Inequality (1.75) is a first-order necessary and sufficient condition for (1.74), which can be expressed as

$$\langle \mathscr{G}\bar{u} - z_d, \mathscr{G}u - \mathscr{G}\bar{u} \rangle_{\mathscr{H}} + \langle \sigma \bar{u}, u - \bar{u} \rangle_{\mathscr{U}} \geq 0 \quad \text{for all } u \in \mathscr{U}_{\mathsf{ad}}. \tag{1.76}$$

Next we study (1.76) for ($\hat{\mathbf{P}}$). Utilizing the setting from (1.73) we obtain

$$\begin{aligned}
&\langle \mathscr{G}\bar{u} - z_d, \mathscr{G}(u - \bar{u}) \rangle_{\mathscr{H}} \\
&\quad = \sigma_Q \langle \mathscr{S}\bar{u} - (y_Q - \hat{y}), \mathscr{S}(u - \bar{u}) \rangle_{L^2(0,T;H)} \\
&\qquad + \sigma_\Omega \langle (\mathscr{S}\bar{u})(T) - (y_\Omega - \hat{y}(T)), (\mathscr{S}(u - \bar{u}))(T) \rangle_H \\
&\quad = \sigma_Q \langle \mathscr{S}\bar{u}, \mathscr{S}(u - \bar{u}) \rangle_{L^2(0,T;H)} + \sigma_\Omega \langle (\mathscr{S}\bar{u})(T), (\mathscr{S}(u - \bar{u}))(T) \rangle_H \\
&\qquad - \sigma_Q \langle y_Q - \hat{y}, \mathscr{S}(u - \bar{u}) \rangle_{L^2(0,T;H)} - \sigma_\Omega \langle y_\Omega - \hat{y}(T), (\mathscr{S}(u - \bar{u}))(T) \rangle_H.
\end{aligned}$$

Let us define the two linear, bounded operators $\Theta : W_0(0,T) \to W_0(0,T)'$ and $\Xi : L^2(0,T;H) \times H \to W_0(0,T)'$ by

$$\begin{aligned}
\langle \Theta\varphi, \phi \rangle_{W_0(0,T)', W_0(0,T)} &= \int_0^T \langle \sigma_Q \varphi(t), \phi(t) \rangle_H \, \mathrm{d}t + \langle \sigma_\Omega \varphi(T), \phi(T) \rangle_H, \\
\langle \Xi z, \phi \rangle_{W_0(0,T)', W_0(0,T)} &= \int_0^T \langle \sigma_Q z_Q(t), \phi(t) \rangle_H \, \mathrm{d}t + \langle \sigma_\Omega z_\Omega, \phi(T) \rangle_H
\end{aligned} \tag{1.77}$$

for $\varphi, \phi \in W_0(0,T)$ and $z = (z_Q, z_\Omega) \in L^2(0,T;H) \times H$. Then, we find

$$\begin{aligned}
&\langle \mathscr{G}\bar{u} - z_d, \mathscr{G}(u - \bar{u}) \rangle_{\mathscr{H}} \\
&\quad = \langle \Theta(\mathscr{S}\bar{u}) - \Xi(y_Q - \hat{y}, y_\Omega - \hat{y}(T)), \mathscr{S}(u - \bar{u}) \rangle_{W_0(0,T)', W_0(0,T)} \\
&\quad = \langle \mathscr{S}'\Theta\mathscr{S}\bar{u}, u - \bar{u} \rangle_U - \langle \mathscr{S}'\Xi(y_Q - \hat{y}, y_\Omega - \hat{y}(T)), u - \bar{u} \rangle_U.
\end{aligned} \tag{1.78}$$

Let us define the linear $\mathscr{A} : U \to W(0,T)$ as follows: for given $u \in U$ the function $p = \mathscr{A}u \in W(0,T)$ is the unique solution to

$$\begin{aligned}
-\frac{\mathrm{d}}{\mathrm{d}t} \langle p(t), \varphi \rangle_H + a(t; p(t), \varphi) &= -\sigma_Q \langle (\mathscr{S}u)(t), \varphi \rangle_H \quad \forall \varphi \in V \text{ a.e.}, \\
p(T) &= -\sigma_\Omega (\mathscr{S}u)(T) \qquad \text{in } H.
\end{aligned} \tag{1.79}$$

It follows from (1.43) and $\mathscr{S}u \in W(0,T)$ that the operator $\mathscr{A}$ is well-defined and bounded.

Lemma 1.41. *Let Assumption* 1 *be satisfied and* $u, v \in U$. *We set* $y = \mathscr{S}u \in W_0(0,T)$, $w = \mathscr{S}v \in W_0(0,T)$, *and* $p = \mathscr{A}v \in W(0,T)$. *Then,*

$$\int_0^T \langle (\mathscr{B}u)(t), p(t)\rangle_{V',V}\,\mathrm{d}t = -\int_0^T \sigma_Q \langle w(t), y(t)\rangle_H\,\mathrm{d}t - \sigma_\Omega \langle w(T), y(T)\rangle_H.$$

Proof. We derive from $y = \mathscr{S}u$, $p = \mathscr{A}u$, $y \in W_0(0,T)$ and integration by parts

$$\begin{aligned}\int_0^T \langle (\mathscr{B}u)(t), p(t)\rangle_{V',V}\,\mathrm{d}t &= \int_0^T \langle y_t(t), p(t)\rangle_{V',V} + a(t; y(t), p(t))\,\mathrm{d}t\\ &= \int_0^T -\langle p_t(t), y(t)\rangle_{V',V} + a(t; p(t), y(t))\,\mathrm{d}t + \langle p(T), y(T)\rangle_H\\ &= -\int_0^T \sigma_Q \langle w(t), y(t)\rangle_H\,\mathrm{d}t - \sigma_\Omega \langle w(T), y(T)\rangle_H\end{aligned}$$

which is the claim. □

We define $\hat{p} \in W(0,T)$ as the unique solution to

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\langle \hat{p}(t), \varphi\rangle_H + a(t; \hat{p}(t), \varphi) &= \sigma_Q \langle y_Q(t) - \hat{y}(t), \varphi\rangle_H \quad && \forall \varphi \in V \text{ a.e.},\\ p(T) &= \sigma_\Omega (y_\Omega - \hat{y}(T)) && \text{in } H. \end{aligned} \tag{1.80}$$

Then, for every $u \in U$ the function $p = \hat{p} + \mathscr{A}u$ is the unique solution to

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\langle p(t), \varphi\rangle_H + a(t; p(t), \varphi) &= \sigma_Q \langle y_Q(t) - y(t), \varphi\rangle_H \quad && \forall \varphi \in V \text{ a.e.},\\ p(T) &= \sigma_\Omega (y_\Omega - y(T)) && \text{in } H \end{aligned}$$

with $y = \hat{y} + \mathscr{S}u$. Moreover, we have the following result.

Lemma 1.42. *Let Assumption* 1 *be satisfied. Then,* $\mathscr{B}'\mathscr{A} = -\mathscr{S}'\Theta\mathscr{S} \in \mathscr{L}(U)$, *where linear and bounded operator* Θ *has been defined in* (1.77). *Moreover,* $\mathscr{B}'\hat{p} = \mathscr{S}'\Xi(y_Q - \hat{y}, y_\Omega - \hat{y}(T))$, *where* $\hat{p}$ *is the solution to* (1.80).

Proof. Let $u, v \in \mathscr{U}$ be chosen arbitrarily. We set $y = \mathscr{S}u \in W_0(0,T)$ and $w = \mathscr{S}v \in W_0(0,T)$. Recall that we identify U with its dual space U'. From the integration by parts formula and Lemma 1.41 we infer that

$$\begin{aligned}\langle \mathscr{S}'\Theta\mathscr{S}v, u\rangle_U &= \langle \Theta\mathscr{S}v, \mathscr{S}u\rangle_{W_0(0,T)',W_0(0,T)} = \langle \Theta w, y\rangle_{W_0(0,T)',W_0(0,T)}\\ &= \int_0^T \sigma_Q \langle w(t), y(t)\rangle_H\,\mathrm{d}t + \sigma_\Omega \langle w(T), y(T)\rangle_H\\ &= -\langle \mathscr{B}u, p\rangle_{L^2(0,T;V'),L^2(0,T;V)} = -\langle u, \mathscr{B}'p\rangle_U = -\langle \mathscr{B}'\mathscr{A}v, u\rangle_U.\end{aligned}$$

Since $u, v \in U$ are chosen arbitrarily, we have $\mathscr{B}'\mathscr{A} = \mathscr{S}'\Theta\mathscr{S}$. Further, we find

$$\begin{aligned}
&\langle \mathscr{S}'\Xi(y_Q - \hat{y}, y_\Omega - \hat{y}(T)), u\rangle_U = \langle \Xi(y_Q - \hat{y}), y_\Omega - \hat{y}(T)), \mathscr{S}u\rangle_{W_0(0,T)', W_0(0,T)} \\
&= \int_0^T \sigma_Q \langle y_Q(t) - \hat{y}(t), y(t)\rangle_H \,\mathrm{d}t + \sigma_\Omega \langle y_\Omega - \hat{y}(T), y(T)\rangle_H \\
&= \int_0^T -\langle \hat{p}_t(t), y(t)\rangle_H + a(t; \hat{p}(t), y(t))\,\mathrm{d}t + \langle \hat{p}(T), y(T)\rangle_H \\
&= \int_0^T \langle y_t(t), \hat{p}(t)\rangle_H + a(t; y(t), \hat{p}(t))\,\mathrm{d}t = \int_0^T \langle (\mathscr{B}u)(t), \hat{p}(t)\rangle_{V',V}\,\mathrm{d}t \\
&= \langle \mathscr{B}'\hat{p}, u\rangle_U.
\end{aligned}$$

which gives the claim. □

We infer from (1.78) and Lemma 1.42 that

$$\langle \mathscr{G}\bar{u} - z_d, \mathscr{G}(u - \bar{u})\rangle_{\mathscr{H}} = -\langle \mathscr{B}'(\hat{p} + \mathscr{A}\bar{u}), u - \bar{u}\rangle_U.$$

This implies the following variational inequality for ($\hat{\mathbf{P}}$)

$$\begin{aligned}
&\langle \mathscr{G}\bar{u} - z_d, \mathscr{G}u - \mathscr{G}\bar{u}\rangle_{\mathscr{H}} + \sigma\,\langle \bar{u}, u - \bar{u}\rangle_{\mathscr{U}} \\
&\quad = \langle \sigma\bar{u} - \mathscr{B}'(\hat{p} + \mathscr{A}\bar{u}), u - \bar{u}\rangle_U \geq 0 \quad \text{for all } u \in U_{\mathsf{ad}}.
\end{aligned}$$

Summarizing we have proved the following result.

Theorem 1.43. *Suppose that Assumptions* 1 *and* 2 *hold. Then,* $(\bar{y}, \bar{u})$ *is a solution to* (**P**) *if and only if* $(\bar{y}, \bar{u})$ *satisfy together with the adjoint variable* $\bar{p}$ *the first-order optimality system*

$$\bar{y} = \hat{y} + \mathscr{S}\bar{u}, \quad \bar{p} = \hat{p} + \mathscr{A}\bar{u}, \quad u_a \leq \bar{u} \leq u_b \tag{1.81a}$$

$$\langle \sigma\bar{u} - \mathscr{B}'\bar{p}, u - \bar{u}\rangle_U \geq 0 \quad \textit{for all } u \in U_{\mathsf{ad}}. \tag{1.81b}$$

Remark 1.44. By using a Lagrangian framework it follows from Theorem 1.43 and [70] that the variational inequality (1.81b) is equivalent to the existence of two functions $\bar{\mu}_a, \bar{\mu}_b \in U$ satisfying $\bar{\mu}_a, \bar{\mu}_b \geq 0$,

$$\sigma\bar{u} - \mathscr{B}'\bar{p} + \bar{\mu}_b - \bar{\mu}_a = 0$$

and the complementarity condition

$$\bar{\mu}_a(t)^\top(u_a(t) - \bar{u}(t)) = \bar{\mu}_b(t)^\top(\bar{u}(t) - u_b(t)) = 0 \quad \text{f.a.a. } t \in [0, T].$$

Thus, (1.81) is equivalent to the system

$$\begin{aligned}
&\bar{y} = \hat{y} + \mathscr{S}\bar{u}, \quad \bar{p} = \hat{p} + \mathscr{A}\bar{u}, \quad \sigma\bar{u} - \mathscr{B}'\bar{p} + \bar{\mu}_b - \bar{\mu}_a = 0, \\
&u_a \leq \bar{u} \leq u_b, \quad 0 \leq \bar{\mu}_a, \quad 0 \leq \bar{\mu}_b, \\
&\bar{\mu}_a(t)^\top(u_a(t) - \bar{u}(t)) = \bar{\mu}_b(t)^\top(\bar{u}(t) - u_b(t)) = 0 \text{ a.e. in } [0, T].
\end{aligned} \tag{1.82}$$

Utilizing a complementarity function it can be shown that (1.82) is equivalent with

$$\begin{aligned}
&\bar{y} = \hat{y} + \mathscr{S}\bar{u}, \quad \bar{p} = \hat{p} + \mathscr{A}\bar{u}, \quad \sigma\bar{u} - \mathscr{B}'\bar{p} + \bar{\mu}_b - \bar{\mu}_a = 0, \quad u_a \leq \bar{u} \leq u_b, \\
&\bar{\mu}_a = \max\big(0,\, \bar{\mu}_a + \eta(\bar{u} - u_a)\big),\ \bar{\mu}_b = \max\big(0,\, \bar{\mu}_b + \eta(\bar{u} - u_b)\big),
\end{aligned} \tag{1.83}$$

where $\eta > 0$ is an arbitrary real number. The max-and min-operations are interpreted componentwise in the pointwise everywhere sense. $\Diamond$

The gradient $\nabla\hat{J} : U \to U$ of the reduced cost functional $\hat{J}$ is given by

$$\nabla J(u) = \sigma u - \mathscr{B}^\star p, \quad u \in U,$$

where $p = \hat{p} + \mathscr{A} u$ holds true; see, e.g., [26]. Thus, a first-order sufficient optimality condition for $(\hat{\mathbf{P}})$ is given by the variational inequality

$$\langle \sigma \bar{u} - \mathscr{B}' \bar{p}, u - \bar{u} \rangle_U \geq 0 \quad \text{for all } u \in U_{\mathsf{ad}}, \tag{1.84}$$

with $\bar{p} = \hat{p} + \mathscr{A} \bar{u}$.

Problem $(\hat{\mathbf{P}})$ can be solved numerically by a primal-dual active set strategy with the choice $\eta = \sigma$. In this case the method is equivalent to a locally superlinearly convergent semi-smooth Newton algorithm applied to (1.83); see [24, 26, 72]. In Algorithm 1.1 we formulate the method in the context of our application. In Section 1.5 we compare Algorithm 1.1 with the Banach fixed point iteration as well as with the projected gradient method [36, 52].

ALGORITHM 1.1. **(Primal-dual active set strategy).**

Require: Starting value (u^0, λ^0) and maximal iteration number k_{max}.

1: Set $k = 0$. For $i = 1, \ldots, m$ determine the active sets

$$\mathscr{A}_{ai}^k = \left\{ t \in [0,T] \,\middle|\, \sigma u_i^k + \lambda_i^k < u_{ai} \text{ a.e.} \right\},$$
$$\mathscr{A}_{bi}^k = \left\{ t \in [0,T] \,\middle|\, \sigma u_i^k + \lambda_i^k > u_{bi} \text{ a.e.} \right\}$$

and the inactive set $\mathscr{I}_i^k = [0,T] \backslash \mathscr{A}_i^k$ with $\mathscr{A}_i^k = \mathscr{A}_{ai}^k \cup \mathscr{A}_{bi}^k$.

2: **repeat**

3: Compute the solution (y, p, u) to the optimality system

$$y = \hat{y} + \mathscr{S} u, \quad p = \hat{p} + \mathscr{A} u, \quad u_i = \begin{cases} u_{ai} & \text{on } \mathscr{A}_{ia}^k, \\ u_{bi} & \text{on } \mathscr{A}_{ib}^k, \\ (\mathscr{B}' p)_i / \sigma & \text{on } \mathscr{I}_i^k, \end{cases} \quad (1 \leq i \leq m)$$

4: Set $(y^{k+1}, u^{k+1}, p^{k+1}) = (y, u, p)$, $\lambda^{k+1} = \mathscr{B}' p^{k+1} - \sigma u^{k+1}$ and $k = k + 1$.

5: Compute the active and inactive sets according to step 1.

6: **until** $(\mathscr{A}_{ai}^k = \mathscr{A}_{ai}^{k-1}$ **and** $\mathscr{A}_{bi}^k = \mathscr{A}_{bi}^{k-1})$ **or** $k = k_{\mathsf{max}}$.

1.4.4 ▪ The POD Galerkin approximation for $(\hat{\mathbf{P}})$

In this subsection we introduce the POD Galerkin schemes for the variational inequality (1.84) using a POD Galerkin approximation for the state and dual variables. Moreover, we study the convergence of the POD discretizations, where we make use of the analysis in [28, 38, 40, 39, 66, 71]. For a general introduction we also refer the reader to the survey paper [27].

In Section 1.3.3 we have introduced a POD Galerkin scheme for the state equation (1.69). Suppose that $\{\psi_i\}_{i=1}^\ell$ be a POD basis of rank ℓ computed from $(\mathbf{P}^\ell)$ with $\psi_i = \psi_i^V$ in case of $X = V$ and $\psi_i = \psi_i^H$ in case of $X = H$. We set $X^\ell = \operatorname{span}\{\psi_1, \ldots, \psi_\ell\} \subset V$.

Let the linear and bounded projection operator $\mathscr{P}^\ell$ denote $\mathscr{P}_V^\ell$ for $X = V$ and $\mathscr{P}_H^\ell$ for $X = H$; see (1.49).

Recall the POD Galerkin ansatz (1.54) for the state variable. Analogously, we approximate the adjoint variable $p = \hat{p} + \mathscr{A}u$ by the Galerkin expansion

$$p^\ell(t) = \hat{p}(t) + \sum_{i=1}^{\ell} \mathrm{p}_i^\ell(t)\psi_i \in V \quad \text{for } t \in [0,T] \tag{1.85}$$

with coefficient functions $\mathrm{p}_i^\ell : [0,T] \to \mathbb{R}$ and with $\hat{p}$ from (1.80). Let the vector-valued coefficient function given by

$$\mathrm{p}^\ell = \big(\mathrm{p}_1^\ell, \dots, \mathrm{p}_\ell^\ell\big) : [0,T] \to \mathbb{R}^\ell$$

If we assume that $\mathrm{p}^\ell(T) = -\sigma_\Omega \mathrm{y}^\ell(T)$ holds, then we infer from $\hat{p}(T) = \sigma_\Omega(y_\Omega - \hat{y}(T))$ and (1.85) that

$$p^\ell(T) = \hat{p}(T) - \sigma_\Omega \sum_{i=1}^{\ell} \mathrm{y}_i^\ell(t)\psi_i = \sigma_\Omega\big(y_\Omega - y^\ell(T)\big).$$

This motivates the following POD scheme for the approximation of $p = \hat{p} + \mathscr{A}u$ is given as follows: $p^\ell \in W(0,T)$ satisfies

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\left\langle p^\ell(t), \psi\right\rangle_H + a(t; p^\ell(t), \psi) &= \sigma_Q \left\langle (y_Q - y^\ell)(t), \psi\right\rangle_H \quad \forall \psi \in X^\ell \text{ a.e.}, \\ \mathrm{p}^\ell(T) &= -\sigma_\Omega \mathrm{y}^\ell(T). \end{aligned} \tag{1.86}$$

It follows by similar arguments as for (1.55) that there is a unique solution $p^\ell \in W(0,T)$.

Remark 1.45. Recall that we have introduced the linear and bounded solution operator $\mathscr{S}^\ell : U \to W(0,T)$ as an approximation for the state solution operator $\mathscr{S}$; see Remark 1.28-2). Analogously, we define an approximation of the adjoint solution operator $\mathscr{A}$ as follows: Let $\mathscr{A}^\ell : U \to W(0,T)$ denote the solution operator to

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\left\langle w^\ell(t), \psi\right\rangle_H + a(t; w^\ell(t), \psi) &= -\sigma_Q \left\langle (\mathscr{S}^\ell u)(t), \psi\right\rangle_H \qquad \forall \psi \in X^\ell \text{ a.e.}, \\ w^\ell(T) &= -\sigma_\Omega (\mathscr{S}^\ell u)(T). \end{aligned}$$

Then $p^\ell = \hat{p} + \mathscr{A}^\ell u$ is the unique solution to (1.86). ◊

Lemma 1.46. *Let Assumption* 1 *on page 32 be satisfied and $u, v \in U$. We set $y^\ell = \mathscr{S}^\ell u \in W_0(0,T)$, $w^\ell = \mathscr{S}^\ell v \in W_0(0,T)$, and $p^\ell = \mathscr{A}^\ell v \in W(0,T)$. Then,*

$$\int_0^T \left\langle (\mathscr{B}u)(t), p^\ell(t)\right\rangle_{V',V} \mathrm{d}t = -\int_0^T \sigma_Q \left\langle w^\ell(t), y^\ell(t)\right\rangle_H \mathrm{d}t - \sigma_\Omega \left\langle w^\ell(T), y^\ell(T)\right\rangle_H.$$

Moreover, $\mathscr{B}'\mathscr{A}^\ell = -(\mathscr{S}^\ell)'\Theta\mathscr{S}^\ell \in \mathscr{L}(U)$, where linear and bounded operator Θ has been defined in (1.77).

Proof. Since the POD basis for the state and adjoint coincide, the claim follows by the same arguments used to prove Lemmas 1.41 and 1.42. □

Theorem 1.47. *Suppose that Assumptions* 1 *and* 2 *hold. Let $u \in U$ be arbitrarily given so that $\mathscr{S}u, \mathscr{A}u \in H^1(0,T;V)\setminus\{0\}$.*

1) *To compute a POD basis $\{\psi_i\}_{i=1}^{\ell}$ of rank ℓ we choose $\wp = 4$, $y^1 = \mathscr{S}u$, $y^2 = (\mathscr{S}u)_t$, $y^3 = \mathscr{A}u$ and $y^4 = (\mathscr{A}u)_t$. Then, $p = \hat{p} + \mathscr{A}u$ and $p^\ell = \hat{p} + \mathscr{A}^\ell u$ satisfies the a-priori error estimate*

$$\|p^\ell - p\|^2_{H^1(0,T;V)} \le \begin{cases} C \sum\limits_{i=\ell+1}^{d_V} \lambda_i^V & \text{if } X = V, \\ C \sum\limits_{i=\ell+1}^{d_H} \lambda_i^H \|\psi_i^H - \mathscr{P}_H^\ell \psi_i^H\|_V^2 & \text{if } X = H \end{cases} \tag{1.87}$$

for a constant C which depends on γ, γ_1, γ_2, T, σ_Ω and σ_Q.

2) *If $\mathscr{S}\tilde{u}$ and $\mathscr{A}\tilde{u}$ belong to $H^1(0,T;V)$ for every $\tilde{u} \in U$ and if $\lambda_i^H > 0$ for all $i \in \mathscr{I}$, then we have*

$$\lim_{\ell\to\infty} \|\mathscr{A} - \mathscr{A}^\ell\|_{\mathscr{L}(U,W(0,T))} = 0. \tag{1.88}$$

Proof. Analogous to (1.62) we have $p^\ell(t) - p(t) = \theta^\ell(t) + \rho^\ell(t)$ for almost all $t \in [0,T]$ with $\theta^\ell = \mathscr{A}^\ell u - \mathscr{P}^\ell(\mathscr{A}u)$ and $\rho^\ell = \mathscr{P}^\ell(\mathscr{A}u) - \mathscr{A}u$. Here, $\mathscr{P}^\ell = \mathscr{P}_V^\ell$ for $X = V$ and $\mathscr{P}^\ell = \mathscr{P}_H^\ell$ for $X = H$. Now, the proof of the claims follows by similar arguments as the proofs of Theorem 1.29, Proposition 4.7 in [28], Proposition 4.6 in [71] and Theorem 6.3 in [66]. To estimate the terminal term $\theta^\ell(T)$ we use observe that

$$\begin{aligned} \|\theta^\ell(T)\|_H &= \left\|\mathscr{P}^\ell\big((\mathscr{A}u)(T)\big) - (\mathscr{A}^\ell u)(T)\right\|_H \\ &\le \sigma_\Omega\Big(\left\|\mathscr{P}^\ell\big((\mathscr{S}u)(T)\big) - (\mathscr{S}u)(T)\right\|_H + \left\|(\mathscr{S}u)(T) - (\mathscr{S}^\ell u)(T)\right\|_H\Big) \\ &\le \sigma_\Omega\Big(\left\|\mathscr{P}^\ell(\mathscr{S}u) - \mathscr{S}u\right\|_{C([0,T];H)} + \left\|\mathscr{S}u - \mathscr{S}^\ell u\right\|_{C([0,T];H)}\Big) \\ &\le \sigma_\Omega c_E\Big(\left\|\mathscr{P}^\ell(\mathscr{S}u) - \mathscr{S}u\right\|_{H^1(0,T;V)} + \left\|\mathscr{S}u - \mathscr{S}^\ell u\right\|_{H^1(0,T;V)}\Big) \end{aligned}$$

with an embedding constant c_E. The first term on the right-hand side can be handled by (1.27), the second term is estimated in Theorem 1.29. Finally, (1.88) follows from (1.61) and the fact that the operator $\mathscr{S}^\ell$ is bounded uniformly with respect to ℓ. □

Remark 1.48.

1) The inclusion of adjoint information into the snapshot ensemble improves the approximation quality also for nonlinear problems; see [15].

2) Analogous to Remark 1.30-2) the a-priori estimate (1.87) holds for an arbitrarily chosen, but fixed control $u \in U$. Furthermore, (1.88) implies that

$$\lim_{\ell\to\infty} \|\hat{p} + \mathscr{A}^\ell\tilde{u} - \hat{p} - \mathscr{A}\tilde{u}\|_{W(0,T)} = 0$$

for any $\tilde{u} \in U$.

3) We can also extend the results in Proposition 1.32 for the adjoint equation and get an a-priori error estimate choosing $\wp = 2$, $y^1 = \mathscr{S} u$ and $y^2 = \mathscr{A} u$. $\Diamond$

The POD Galerkin approximation for $(\hat{\mathbf{P}})$ is as follows:

$$\min \hat{J}^\ell(u) \quad \text{s.t.} \quad u \in U_{\mathsf{ad}}, \tag{$\hat{\mathbf{P}}^\ell$}$$

where the cost is defined by $\hat{J}^\ell(u) = J(\hat{y} + \mathscr{S}^\ell u, u)$ for $u \in U$. Let $\bar{u}^\ell$ be the solution to $(\hat{\mathbf{P}}^\ell)$. Then, a first-order sufficient optimality condition is given by the variational inequality

$$\langle \sigma \bar{u}^\ell - \mathscr{B}' \bar{p}^\ell, u - \bar{u}^\ell \rangle_U \geq 0 \quad \text{for all } u \in U_{\mathsf{ad}}, \tag{1.89}$$

where $\bar{p}^\ell = \hat{p}^\ell + \mathscr{A}^\ell \bar{u}^\ell$ holds.

Theorem 1.49. *Suppose that Assumptions* 1 *and* 2 *hold. Let $u \in U$ be arbitrarily given so that $\mathscr{S} u$, $\mathscr{A} u \in H^1(0,T;V) \setminus \{0\}$.*

1) *To compute a POD basis $\{\psi_i\}_{i=1}^\ell$ of rank ℓ we choose $\wp = 4$, $y^1 = \mathscr{S} u$, $y^2 = (\mathscr{S} u)_t$, $y^3 = \mathscr{A} u$ and $y^4 = (\mathscr{A} u)_t$. Then, the optimal solution $\bar{u}$ to $(\hat{\mathbf{P}})$ and the associated POD suboptimal solution $\bar{u}^\ell$ to $(\hat{\mathbf{P}}^\ell)$ satisfy*

$$\lim_{\ell \to \infty} \left\| \bar{u}^\ell - \bar{u} \right\|_U = 0 \tag{1.90}$$

for $X = V$ and $X = H$.

2) *If an optimal POD basis of rank is computed by choosing $\wp = 4$, $y^1 = \mathscr{S} \bar{u}$, $y^2 = (\mathscr{S} \bar{u})_t$, $y^3 = \mathscr{A} \bar{u}$ and $y^4 = (\mathscr{A} \bar{u})_t$, then we have*

$$\left\| \bar{u}^\ell - \bar{u} \right\|_U \leq \begin{cases} \dfrac{C}{\sigma} \displaystyle\sum_{i=\ell+1}^{d_V} \lambda_i^V & \text{if } X = V, \\ \dfrac{C}{\sigma} \displaystyle\sum_{i=\ell+1}^{d_H} \lambda_i^H \| \psi_i^H - \mathscr{P}_H^\ell \psi_i^H \|_V^2 & \text{if } X = H, \end{cases} \tag{1.91}$$

where the constant C depends on γ, γ_1, γ_2, T, σ_Ω, σ_Q and the norm $\|\mathscr{B}'\|_{\mathscr{L}(L^2(0,T;V),U)}$.

Proof. Choosing $u = \bar{u}^\ell$ in (1.84) and $u = \bar{u}$ in (1.89) we get the variational inequality

$$0 \leq \langle \sigma(\bar{u} - \bar{u}^\ell) - \mathscr{B}'(\bar{p} - \bar{p}^\ell), \bar{u}^\ell - \bar{u} \rangle_U. \tag{1.92}$$

Utilizing Lemma 1.46 and $\langle \Theta \varphi, \varphi \rangle_{W_0(0,T)', W_0(0,T)} \geq 0$ for all $\varphi \in W_0(0,T)$ we infer from (1.92) that

$$\begin{aligned} 0 &\leq \langle \mathscr{B}' \mathscr{A}^\ell \bar{u}^\ell - \mathscr{B}' \mathscr{A} \bar{u}, \bar{u}^\ell - \bar{u} \rangle_U - \sigma \|\bar{u} - \bar{u}^\ell\|_U^2 \\ &= \langle \mathscr{B}' \mathscr{A}^\ell (\bar{u}^\ell - \bar{u}) + \mathscr{B}' (\mathscr{A}^\ell - \mathscr{A}) \bar{u}, \bar{u}^\ell - \bar{u} \rangle_U - \sigma \|\bar{u} - \bar{u}^\ell\|_U^2 \\ &\leq \langle \Theta \mathscr{S}^\ell (\bar{u} - \bar{u}^\ell), \mathscr{S}^\ell (\bar{u}^\ell - \bar{u}) \rangle_U + \|\mathscr{B}' (\mathscr{A}^\ell - \mathscr{A}) \bar{u}\|_U \|\bar{u}^\ell - \bar{u}\|_U - \sigma \|\bar{u} - \bar{u}^\ell\|_U^2 \\ &\leq \|\mathscr{B}' (\mathscr{A}^\ell - \mathscr{A}) \bar{u}\|_U \|\bar{u}^\ell - \bar{u}\|_U - \sigma \|\bar{u} - \bar{u}^\ell\|_U^2. \end{aligned}$$

Consequently,

$$\|\bar{u}-\bar{u}^\ell\|_U \le \frac{1}{\sigma}\|\mathscr{B}'(\mathscr{A}^\ell-\mathscr{A})\bar{u}\|_U.$$

Now (1.90) and (1.91) follow from (1.88) and (1.87), respectively. □

In Algorithm 1.2 we formulate a discrete version of the primal-dual active set method (see Algorithm 1.1) which is utilized to solve ($\hat{\mathbf{P}}^\ell$) in Section 1.5.

ALGORITHM 1.2. **(POD discretized primal-dual active set strategy).**

Require: POD basis $\{\psi_i\}_{i=1}^\ell$, starting value $(u^{\ell 0}, \lambda^{\ell 0})$, maximal iteration number $k_{\max}$.

1: Set $k=0$, determine the active sets

$$\mathscr{A}_{ai}^{\ell k} = \{t\in[0,T] \,|\, \sigma u_i^{k\ell} + \lambda_i^{k\ell} < u_{ai} \text{ a.e.}\},$$
$$\mathscr{A}_{bi}^{\ell k} = \{t\in[0,T] \,|\, \sigma u_i^{k\ell} + \lambda_i^{k\ell}(t) > u_{bi}(t)\}$$

and the inactive sets $\mathscr{I}_i^{\ell k} = [0,T]\setminus \mathscr{A}_i^{\ell k}$ with $\mathscr{A}_i^{\ell k} = \mathscr{A}_{ai}^{\ell k}\cup\mathscr{A}_{bi}^{\ell k}$.

2: **repeat**

3: Determine the solution (y^ℓ, u^ℓ, p^ℓ) to the optimality system

$$y^\ell = \hat{y} + \mathscr{S}^\ell u^\ell, \quad p^\ell = \hat{p} + \mathscr{A}^\ell u^\ell, \quad u^\ell = \begin{cases} u_a & \text{on } \mathscr{A}_a^{k\ell}, \\ u_b & \text{on } \mathscr{A}_b^{k\ell}, \\ \mathscr{B}' p^\ell/\sigma & \text{on } \mathscr{I}^{k\ell}. \end{cases}$$

4: Set $(y^{\ell,k+1}, u^{\ell,k+1}, p^{\ell,k+1}) = (y^\ell, u^\ell, p^\ell)$, $\lambda^{\ell,k+1} = \mathscr{B}' p^{\ell,k+1} - \sigma u^{\ell,k+1}$ and $k = k+1$.

5: Compute the active and inactive sets according to step 1.

6: **until** ($\mathscr{A}_a^{\ell k} = \mathscr{A}_a^{\ell,k-1}$ **and** $\mathscr{A}_b^{\ell k} = \mathscr{A}_b^{\ell,k-1}$) **or** $k = k_{\max}$.

1.4.5 ▪ POD a-posteriori error analysis

In [71] a POD a-posteriori error estimates are presented which can be applied to our optimal control problem as well. Based on a perturbation method [16] it is deduced how far the suboptimal control $\bar{u}^\ell$ is from the (unknown) exact optimal control $\bar{u}$. Thus, our goal is to estimate the norm $\|\bar{u}-\bar{u}^\ell\|_U$ without the knowledge of the optimal solution $\bar{u}$. In general, $\bar{u}^\ell \neq \bar{u}$ holds, so that $\bar{u}^\ell$ does not satisfy the variational inequality (1.84). However, there exists a function $\zeta^\ell \in U$ such that

$$\langle \sigma\bar{u}^\ell - \mathscr{B}'\tilde{p}^\ell + \zeta^\ell, u-\bar{u}^\ell\rangle_U \ge 0 \quad \forall v\in U_{\text{ad}}, \tag{1.93}$$

with $\tilde{p}^\ell = \hat{p} + \mathscr{A}\bar{u}^\ell$. Therefore, $\bar{u}^\ell$ satisfies the optimality condition of the perturbed parabolic optimal control problem

$$\min_{u\in U_{\text{ad}}} \tilde{J}(u) = J(\hat{y}+\mathscr{S}u, u) + \langle \zeta^\ell, u\rangle_U$$

with "perturbation" ζ^ℓ. The smaller ζ^ℓ is, the closer $\bar{u}^\ell$ is to $\bar{u}$. Next we estimate $\|\bar{u}-\bar{u}^\ell\|_U$ in terms of $\|\zeta^\ell\|_U$. By Lemma 1.42 we have

$$\mathscr{B}'(\bar{p}-\tilde{p}^\ell) = \mathscr{B}'\mathscr{A}(\bar{u}-\bar{u}^\ell) = -\mathscr{S}'\Theta\mathscr{S}(\bar{u}-\bar{u}^\ell) = \mathscr{S}'\Theta(\tilde{y}^\ell-\bar{y}) \tag{1.94}$$

with $\tilde{y}^\ell = \hat{y} + \mathscr{S}\bar{u}^\ell$. Choosing $u = \bar{u}^\ell$ in (1.84), $u = \bar{u}$ in (1.93) and using (1.94) we obtain

$$\begin{aligned}
0 &\le \langle -\sigma(\bar{u} - \bar{u}^\ell) + \mathscr{B}'(\bar{p} - \tilde{p}^\ell) + \zeta^\ell, \bar{u} - \bar{u}^\ell \rangle_U \\
&= -\sigma \|\bar{u} - \bar{u}^\ell\|_U^2 + \langle \mathscr{S}'\Theta(\tilde{y}^\ell - \bar{y}), \bar{u} - \bar{u}^\ell \rangle_U + \langle \zeta^\ell, \bar{u} - \bar{u}^\ell \rangle_U \\
&= -\sigma \|\bar{u} - u_p\|_U^2 - \langle \Theta(\bar{y} - \tilde{y}^\ell), \bar{y} - \tilde{y}^\ell \rangle_{W_0(0,T)', W_0(0,T)} + \langle \zeta^\ell, \bar{u} - \bar{u}^\ell \rangle_U \\
&= -\sigma \|\bar{u} - \bar{u}^\ell\|_U^2 + \langle \zeta^\ell, \bar{u}^\ell - \bar{u}^\ell \rangle_U \le -\sigma \|\bar{u} - \bar{u}^\ell\|_U^2 + \|\zeta^\ell\|_U \|\bar{u} - \bar{u}^\ell\|_U.
\end{aligned}$$

Hence, we get the a-posteriori error estimation

$$\|\bar{u} - \bar{u}^\ell\|_U \le \frac{1}{\sigma} \|\zeta^\ell\|_U.$$

Theorem 1.50. *Suppose that Assumptions* 1 *and* 2 *hold. Let $u \in U$ be arbitrarily given so that $\mathscr{S}u$, $\mathscr{A}u \in H^1(0,T;V) \setminus \{0\}$. To compute a POD basis $\{\psi_i\}_{i=1}^\ell$ of rank ℓ we choose $\wp = 4$, $y^1 = \mathscr{S}u$, $y^2 = (\mathscr{S}u)_t$, $y^3 = \mathscr{A}u$ and $y^4 = (\mathscr{A}u)_t$. Define the function $\zeta^\ell \in U$ by*

$$\zeta_i^\ell(t) = \begin{cases} -\min(0, \xi_i^\ell(t)) & \text{a.e. in } \mathscr{A}_{ai}^\ell = \{t \in [0,T] \,|\, \bar{u}_i^\ell(t) = u_{ai}(t)\}, \\ -\max(0, \xi_i^\ell(t)) & \text{a.e. in } \mathscr{A}_{bi}^\ell = \{t \in [0,T] \,|\, \bar{u}_i^\ell(t) = u_{bi}(t)\}, \\ -\xi_i^\ell(t) & \text{a.e. in } [0,T] \setminus (\mathscr{A}_{ai}^\ell \cup \mathscr{A}_{bi}^\ell), \end{cases}$$

where $\xi^\ell = \sigma\bar{u}^\ell - \mathscr{B}'(\hat{p} + \mathscr{A}\bar{u}^\ell)$ in U. Then, the a-posteriori error estimate

$$\|\bar{u} - \bar{u}^\ell\|_U \le \frac{1}{\sigma} \|\zeta^\ell\|_U. \tag{1.95}$$

In particular, $\lim_{\ell \to \infty} \|\zeta^\ell\|_U = 0$.

Proof. Estimate (1.95) has already be shown. We proceed by constructing the function ζ^ℓ. Here we adapt the lines of the proof of Proposition 3.2 in [71] to our optimal control problem. Suppose that we know $\bar{u}^\ell$ and $\tilde{p}^\ell = \hat{p} + \mathscr{A}\bar{u}^\ell$. The goal is to determine $\zeta^\ell \in U$ satisfying (1.93). We distinguish three different cases.

- Case $\bar{u}_i^\ell(t) = u_{ai}(t)$ for fixed $t \in [0,T]$ and $i \in \{1, \dots, N_u\}$: Then, $u_i(t) - \bar{u}_i^\ell(t) = u_i(t) - u_{ai}(t) \ge 0$ for all $u \in U_{ad}$. Hence, $\zeta_i^\ell(t)$ has to satisfy

$$\left(\sigma\bar{u}^\ell - \mathscr{B}'\tilde{p}^\ell\right)_i(t) + \zeta_i^\ell(t) \ge 0. \tag{1.96}$$

  Setting $\zeta_i^\ell(t) = -\min(0, (\sigma\bar{u}^\ell - \mathscr{B}'\tilde{p}^\ell)_i(t))$ the value $\zeta_i^\ell(t)$ satisfies (1.96).

- Case $\bar{u}_i^\ell(t) = u_{bi}(t)$ for fixed $t \in [0,T]$ and $i \in \{1, \dots, N_u\}$: Now, $u_i(t) - \bar{u}_i^\ell(t) = u(t) - u_{bi}(t) \le 0$ for all $u \in U_{ad}$. Analogously to the first case we define $\zeta_i^\ell(t) = -\max(0, (\sigma\bar{u}^\ell - \mathscr{B}'\tilde{p}^\ell)_i(t))$ to ensure (1.96).

- Case $u_{ai}(t) < \bar{u}_i^\ell(t) < u_{bi}(t)$ for fixed $t \in [0,T]$ and $i \in \{1, \dots, N_u\}$: Consequently, $(\sigma\bar{u}^\ell - \mathscr{B}'\tilde{p}^\ell)_i(t) + \zeta_i^\ell(t) = 0$ holds so that $\zeta_i^\ell(t) = -(\sigma\bar{u}^\ell - \mathscr{B}'\tilde{p}^\ell)_i(t)$ guarantees (1.96).

It remains to prove that ζ^ℓ tends to zero for $\ell \to \infty$. Here we follow adapt the proof of Theorem 4.11 in [71]. By Theorem 1.49-1), the sequence $\{\bar{u}^\ell\}_{\ell\in\mathbb{N}}$ converges to $\bar{u}$ in U. Since the linear operator $\mathscr{B}'\mathscr{A}$ is bounded and $\tilde{p}^\ell = \hat{p} + \mathscr{A}\bar{u}^\ell$ holds, $\{\mathscr{B}'\tilde{p}^\ell\}_{\ell\in\mathbb{N}}$ tends to $\mathscr{B}'\bar{p} = \mathscr{B}'\mathscr{A}\bar{u}$ as well. Hence, there exist subsequences $\{\bar{u}^{\ell_k}\}_{k\in\mathbb{N}}$ and $\{\mathscr{B}'\tilde{p}^{\ell_k}\}_{k\in\mathbb{N}}$ satisfying

$$\lim_{k\to\infty} \bar{u}_i^{\ell_k}(t) = \bar{u}_i(t) \quad \text{and} \quad \lim_{k\to\infty} (\mathscr{B}'\tilde{p}^{\ell_k})_i(t) = (\mathscr{B}'\bar{p})_i(t)$$

f.a.a. $t \in [0,T]$ and for $1 \le i \le N_u$. Next we consider the active and inactive sets for $\bar{u}$.

- Let $t \in \mathscr{J}_i = \{t \in [0,T] \,|\, u_{ai}(t) < \bar{u}_i(t) < u_{bi}(t)\}$ for $i \in \{1,\dots,N_u\}$. For $k_\circ = k_\circ(t) \in \mathbb{N}$ sufficiently large, $\bar{u}_i^{\ell_k}(t) \in (u_{ai}(t), u_{bi}(t))$ for all $k \ge k_\circ$ and f.a.a. $t \in \mathscr{J}_i$. Thus, $(\sigma\bar{u}^{\ell_k} - \mathscr{B}'\tilde{p}^{\ell_k})_i(t) = 0$ for all $k \ge k_\circ(t)$ in $\mathscr{J}_i$ a.e. This implies

$$\zeta_i^{\ell_k}(t) = 0 \quad \forall k \ge k_\circ \text{ and f.a.a. } t \in \mathscr{J}_i. \tag{1.97}$$

- Suppose that $t \in \mathscr{A}_{ai} = \{t \in [0,T] \,|\, u_{ai}(t) = \bar{u}_i(t)\}$ for $i \in \{1,\dots,N_u\}$. From $(\sigma\bar{u}_i - \mathscr{B}'\bar{p})_i(t) \ge 0$ in $\mathscr{A}_{ai}$ a.e. we deduce

$$\lim_{k\to\infty} \zeta_i^{\ell_k}(t) = -\min(0, (\sigma\bar{u}^{\ell_k} - \mathscr{B}'\tilde{p}^{\ell_k})_i(t)) = 0 \quad \text{f.a.a. } t \in \mathscr{A}_{ai}.$$

- Suppose that $t \in \mathscr{A}_{bi} = \{t \in [0,T] \,|\, u_{bi}(t) = \bar{u}_i(t)\}$. Analogously to the second case we find

$$\lim_{k\to\infty} \zeta_i^{\ell_k}(t) = 0 \quad \text{f.a.a. } t \in \mathscr{A}_{bi}. \tag{1.98}$$

Combining (1.97)-(1.98) we conclude that $\lim_{k\to\infty} \zeta_i^{\ell_k} = 0$ a.e. in $[0,T]$ and for $1 \le i \le N_u$. Utilizing the dominated convergence theorem [60, p. 24] we have

$$\lim_{k\to\infty} \left\|\zeta^{\ell_k}\right\|_U = 0.$$

Since all subsequences contain a subsequence converging to zero, the claim follows from a standard argument. □

Remark 1.51.

1) Theorem 1.50 shows that $\|\zeta^\ell\|_U$ tends to zero as ℓ goes to infinity. Thus, $\|\zeta^\ell\|_U$ is smaller than any tolerance $\epsilon > 0$ provided that ℓ is taken sufficiently large. Motivated by this result we set up Algorithm 1.3. Note that the approximation quality of the POD Galerkin scheme is improved by only increasing the number of POD basis elements: A rank-ℓ POD basis can be extended to a rank-$(\ell+1)$ POD basis by adding a new eigenfunction and keeping all the old ones. Especially, the system matrices and projected data functions can be extended by the new element, they do not have to be reconstructed in all components. Another approach is to update the POD basis in the optimization process; see, e.g., [1, 3, 41].

2) We infer from Proposition 1.32 and Remark 1.48-3) that Theorem 1.50 holds still true if we take $\wp = 2$, $y^1 = \mathscr{S}u$ and $y^2 = \mathscr{A}u$.

3) In [68] POD a-posteriori error estimates are tested numerically for a linear-quadratic optimal control problem. It turns out that in certain cases a change of the POD basis is required in order to improve the approximation quality of the POD scheme; see also [41, 74], for instance.

4) Let us refer to [33], where POD a-posteriori error estimates are combined with an sequential quadratic programming method in order to solve a nonlinear PDE constrained optimal control problem. Furthermore, the presented analysis for linear-quadratic problems can be extended to semilinear optimal control problems by a second-order analysis; see in [34]. ◊

ALGORITHM 1.3. **(POD reduced-order method with a-posteriori estimator).**

Require: Initial control $u^{0\ell} \in U$, initial number ℓ for the POD ansatz functions, a maximal number $\ell_{\max} > \ell$ of POD ansatz functions, and a stopping tolerance $\epsilon > 0$.

Determine $\hat{y}$, $\hat{p}$, $y^1 = \mathcal{S} u^{0\ell}$, $y^2 = \mathcal{A} u^{0\ell}$.
Compute a POD basis $\{\psi_i\}_{i=1}^{\ell_{\max}}$ choosing y^1 and y^2. Set $\ell = 1$.
repeat
  Establish the POD Galerkin discretization using $\{\psi_i\}_{i=1}^{\ell}$.
  Call Algorithm 1.2 to compute suboptimal control $\bar{u}^\ell$.
  Determine ζ^ℓ according to Theorem 1.47 and compute $\epsilon_{\text{ape}} = \|\zeta^\ell\|_U / \sigma$.
  if $\epsilon_{\text{ape}} < \epsilon$ **or** $\ell = \ell_{\max}$ **then**
    Return ℓ and suboptimal control $\bar{u}^\ell$ and STOP.
  end if
  Set $\ell = \ell + 1$.
until $\ell > \ell_{\max}$

1.5 ▪ Numerical experiments

In this section we present numerical test examples to illustrate our theoretical findings. The programs are written in MATLAB utilizing the PARTIAL DIFFERENTIAL EQUATION TOOLBOX for the computation of the finite element (FE) discretization. For the temporal integration the implicit Euler method is applied based on the equidistant time grid $t_j = (j-1)\Delta t$, $j = 1, \ldots, n$ and $\Delta t = T/(n-1)$.

Run 1 (POD for the heat equation). Let us apply the setting of Example 1.33. We choose the final time $T = 3$, the spatial domain $\Omega = (0,2) \subset \mathbb{R}$, the Hilbert spaces $H = L^2(\Omega)$, $V = H_0^1(\Omega)$, the source term $f(t,x) = t^3 - x^2$ for $(t,x) \in Q$ and the discontinuous initial value $y_\circ(x) = \chi_{(0.5,1.0)} - \chi_{(1,1.5)}$ for $x \in \Omega$, where, e.g., $\chi_{(0.5,1)}$ denotes the characteristic function on the subdomain $(0.5,1) \subset \Omega$, $\chi_{(0.5,1)}(x) = 1$ for $x \in (0.5,1)$ and $\chi_{(0.5,1)}(x) = 0$ otherwise. We consider a discretization of the controlled linear heat equation

$$\begin{aligned} y_t(t,x) - \Delta y(t,x) &= f(t,x) + \sum_{i=1}^{m} u_i(t)\chi_i(x), && \text{a.e. in } Q, \\ y(t,x) &= 0, && \text{a.e. in } \Sigma, \\ y(0,x) &= y_\circ(x), && \text{a.e. in } \Omega. \end{aligned} \tag{1.99}$$

To obtain an accurate approximation of the exact solution we choose $n = 4000$ so that $\Delta t \approx 7.5 \cdot 10^{-4}$ holds. For the FE discretization we choose $m = 500$ spatial grid points

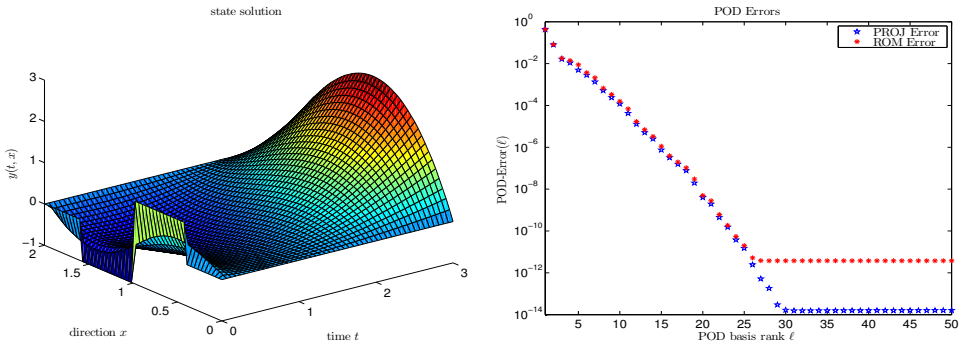

Figure 1.1. *Run 1: The FE solution y^h (left) and the residuals corresponding to the POD basis rank ℓ (right).*

and the equidistant mesh size $h = 2/(m+1) \approx 4 \cdot 10^{-3}$. Thus, the FE error – measured in the H-norm – is of the order 10^{-4}. In the left graphic of Figure 1.1, the FE solution y^h to the state equation (1.71) is visualized. To compute a POD basis $\{\psi_i\}_{i=1}^{\ell}$ of rank ℓ we utilize the multiple discrete snapshots $y_j^1 = y^h(t_j)$ for $1 \le j \le n$ as well $y_1^2 = 0$ and $y_j^2 = (y^h(t_j) - y^h(t_{j-1})/\Delta t$, $j = 2, \dots, n$, i.e., we include the temporal difference quotients. We choose $X = H$ and utilize the (stable) singular value decomposition to determine the POD basis of rank ℓ; compare Remark 1.12. We address this issue in a more detail in Run 4. Since the snapshots are FE functions, the POD basis elements are also FE functions. In the right plot of Figure 1.1, the projection and reduced-order error given by

$$\text{PROJ Error}(\ell) = \left(\sum_{j=1}^{n} \alpha_j \left\| y^h(t_j) - \sum_{i=1}^{\ell} \langle y^h(t_j), \psi_i \rangle_H \psi_i \right\|_H^2 \right)^{1/2},$$

$$\text{ROM Error}(\ell) = \left(\sum_{j=1}^{n} \alpha_j \left\| y^h(t_j) - y^\ell(t_j) \right\|_H^2 \right)^{1/2}$$

are plotted for different POD basis ranks ℓ. The chosen trapezoidal weights α_j have been introduced in (1.31). We observe that both errors decay rapidly and coincide until the accuracy 10^{-12}, which is already significant smaller than the FE discretisation error. This numerical results reflects the a-priori error estimates of Theorem 1.29.

Run 2 (POD for a convection dominated parabolic problem). To present a more challenging situation, we study a convection-reaction-diffusion equation with a source term which is close to be singular: Let T, Ω, $y_\circ$, H and V be given as in Run 1. The time-independent bilinear form a corresponding to

$$\begin{aligned} y_t(t,x) - \eta_2 y_{xx}(t,x) & \\ + \eta_1 y_x(t,x) + \eta_0 y(t,x) = f(t,x) + (\mathscr{B}u)(t,x), & \quad \text{a.e. in } Q, \\ y(t,x) = 0, & \quad \text{a.e. in } \Sigma, \\ y(0,x) = y_\circ(x), & \quad \text{a.e. in } \Omega. \end{aligned} \tag{1.100}$$

is given by

$$a(\phi, \varphi) = \eta_2 \langle \phi', \varphi' \rangle_H + \eta_1 \langle \phi', \varphi \rangle_H + \eta_0 \langle \phi, \varphi \rangle_H \quad \text{for } \varphi, \phi \in V.$$

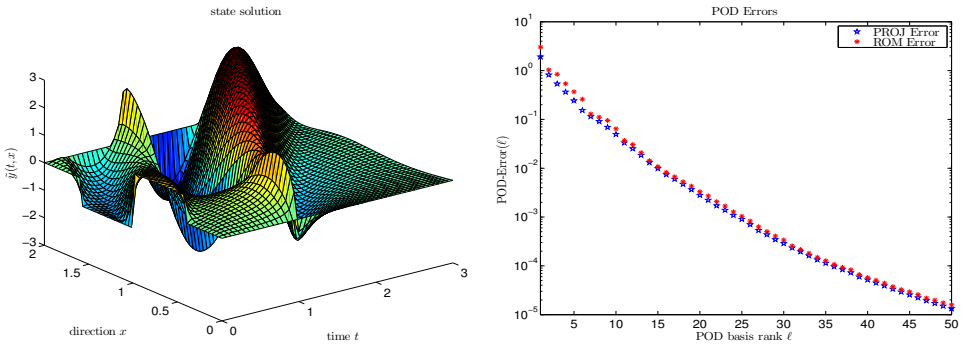

Figure 1.2. *Run 2: The FE solution y^h (left) and the residuals corresponding the POD basis rank ℓ (right).*

We choose the diffusivity $\eta_2 = 0.025$, the velocity $\eta = 1.0$ that determines the speed in which the initial profile $y_\circ$ is shifted to the boundary and the reaction rate $\eta_0 = -0.001$. Finally, $f(t,x) = \mathbb{P}(\frac{1}{1-t})\cos(\pi x)$ for $(t,x) \in Q$, where $(\mathbb{P}z)(t) = \min(+l, \max(-l, z(t)))$ restricts the image of z on a bounded interval. In this situation, the state solution y develops a jump at $t = 1$ for $l \to \infty$; see the left plot of Figure 1.2. The right plot of Figure 1.2 demonstrates that in this case, the decay of the reconstruction residuals and the decay of the errors are much slower. The manifold dynamics of the state solution require an inconvenient large number of POD basis elements. Since the supports of these ansatz functions in general cover the whole domain Ω, the corresponding system matrices M^ℓ and A^ℓ of the reduced model (compare (1.57)) are not sparse in contrast to the matrices arising in the finite element Galerkin framework, so the model order reduction cannot be provided efficiently for this example if a good accuracy of the solution function y^ℓ is required.

Run 3 (True and exact approximation error). Let us consider the setting of Run 1 again. The exact solution to (1.71) does not possess a representation by elementary functions. Hence, the presented reconstruction and reduction errors actually are the residuals with respect to a high-order finite element solution y^h. To compute an approximation y of the exact solution y_{ex} we apply a Crank-Nicolson method (with Rannacher smoothing [57]) ensuring $\|y - y_{ex}\|_{L^2(0,T;H)} = \mathcal{O}(\Delta t^2 + h^2) \approx 10^{-5}$. In the context of model reduction, such a state is sometimes called the "true" solution. To compute the FE state y^h we apply the implicit Euler method. In the left plot of Figure 1.3 we compare the true solution with the associated POD approximation for different values $n = Nt \in \{64, 128, 256, ..., 8192\}$ of the time integration and for the spatial mesh size $h = 4 \cdot 10^{-3}$. For the norm we apply a discrete $L^2(0,T;H)$-norm as in Run 1. Let us mention that we compute for every n a corresponding FE solution y^h. We observe that the residuals ignore the errors arising by the application of time and space discretization schemes for the full-order model. The errors decay below the discretization error 10^{-5}. If these discretization errors are taken into account, the residuals stagnate at the level of the full-order model accuracy instead of decaying to zero; see the right plot of Figure 3. Due to the implicit Euler method we have $\|y^h - y_{ex}\|_{L^2(0,T;H)} = \mathcal{O}(\Delta t + h^2)$ with $h = 4 \cdot 10^{-3}$. In particular, from $n \in \{64, 128, 256, ..., 8192\}$ it follows that $\Delta t > 3 \cdot 10^{-4} > h^2 = 1.6 \cdot 10^{-5}$. Therefore, the spatial error is dominated by the time error for all values of n. We can observe that the exact residuals do not decay below a limit of the order Δt. One can observe that for fixed POD basis rank ℓ, the residuals with respect to the true solution increase if the high-order accuracy is improved by enlarging n, since the reduced order model has to approximate a

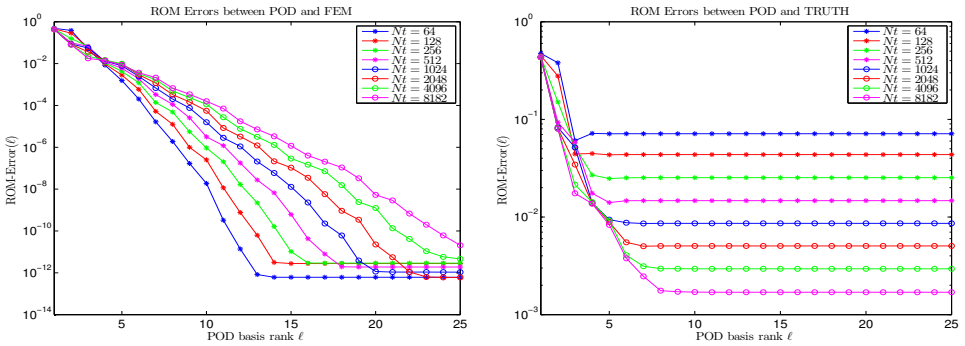

Figure 1.3. *Run 3: The ROM errors with respect to the true solution (left) and the exact one (right).*

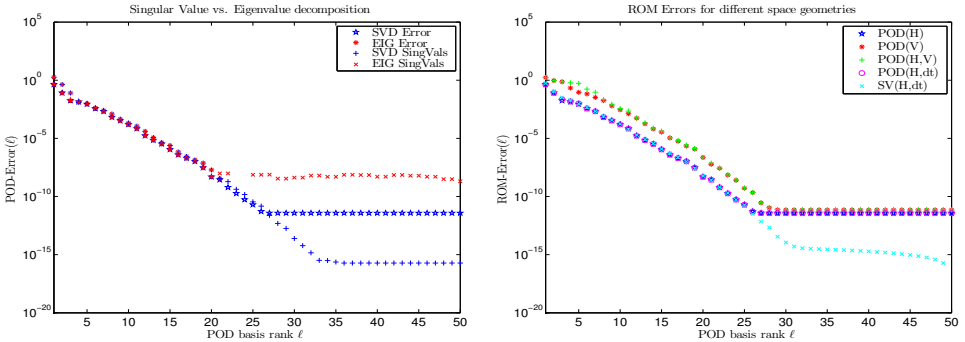

Figure 1.4. *Run 4: Singular values σ_i using the SVD (SVD Vals) or the eigenvalue decomposition (EIG Vals) and the associated ROM errors (SVD error and EIG Error, respectively) (left); ROM errors for different the choices for X, the error norm and the snapshot ensembles (right).*

more complex system in this case, where the residuals with respect to the exact solution decrease due to the lower limit of stagnation $\Delta t = 3/(n-1)$.

Run 4 (Different strategies for the POD basis computation). Let $Y \in \mathbb{R}^{m\times n}$ denote the matrix of snapshots in the discrete setting, $W = (\langle \varphi_i, \varphi_j \rangle_X) \in \mathbb{R}^{m\times m}$ be the (sparse) spatial weight matrix arising from the finite element basis $\{\varphi_i\}_{i=1}^m$ and $D = \Delta t\, \mathrm{diag}(\frac{1}{2}, 1, ..., 1, \frac{1}{2}) \in \mathbb{R}^{n\times n}$ be the trapezoidal time integration matrix fitting to implicit Euler discretization. As it is stated in Remark 1.11, the POD basis $\{\psi_i\}_{i=1}^\ell$ of rank ℓ can be determined by providing an eigenvalue decomposition of the matrix $\hat{Y}\hat{Y}^\top = W^{1/2} Y D Y^\top W^{1/2} \in \mathbb{R}^{m\times m}$, one of $\hat{Y}^\top \hat{Y} = D^{1/2} Y^\top W Y D^{1/2} \in \mathbb{R}^{n\times n}$, or a singular value decomposition of $\hat{Y} = W^{1/2} Y D^{1/2} \in \mathbb{R}^{m\times n}$. Since $n \gg m$ in Runs 1-3, the first variant is the cheepest one from a computational point of view. In case of multiple space dimensions or if a second-order time integration scheme such as some Crank-Nicolson technique is applied, the situation is converse. On the other hand, a singular value decomposition is more accurate than an eigenvalue decomposition if the POD elements corresponding to eigenvalues/singular values which are close to zero are taken into account: Since $\lambda_i = \sigma_i^2$ holds for all eigenvalues λ_i and singular values σ_i, the singular values are able to decay to machine precision, where the eigenvalues stagnate significantly above. This is illustrated in the left graphic of Figure 1.4. Indeed, for $\ell > 20$ the EIG-ROM system matrices become singular due to the numerical errors in the eigenfunctions and the reduced order system is ill-posed in this case

while the SVD-ROM model remains stable. In the right plot of Figure 1.4 POD elements are constructed with respect to different scalar products and the resulting ROM errors are compared: $\|\cdot\|_H$-residuals for $X = H$ (denoted by POD(H)), $\|\cdot\|_V$-residuals for $X = V$ (denoted by POD(V)), $\|\cdot\|_V$-residuals for $X = H$ (denoted by POD(H,V)), which also works quite well, the consideration of time derivatives in the snapshot sample (denoted by POD(H,dt)) which allows to apply the a priori error estimate given in (1.60) and the corresponding sums of singular values (denoted by SV(H,dt)) corresponding to the unused eigenfunctions in the latter case which indeed nearly coincide with the ROM errors. In many applications, the quality of the reduced order model does not vary significantly if the weights matrix W refers to the space $X = H$ or $X = V$ and if time derivatives of the used snapshots are taken into account or not. Especially, the ROM residual decays with the same order as the sum over the remaining singular values, $\|y - y^\ell\|_{W(0,T)} \sim \sum_{i=\ell+1}^{\infty} \sigma_i$ independent of the chosen geometrical framework.

Run 5 (Iterative methods for the optimal control problem). In this numerical test we consider solution techniques for the linear-quadratic optimal control problem (**P**). We define the weights $\sigma_Q = 1$, $\sigma_\Omega = 0$, the desired state $y_Q(t,x) = t(1-(x-1)^2)$ for $(t,x) \in Q$, the desired final state $y_\Omega = 0$ (which is redundant due to $\sigma_\Omega = 0$, of course), the upper and lower bounds $u_a = 0.25$, $u_b = 0.75$, the control operator $(\mathcal{B}u)(t,x) = u_1(t)\chi_{\Omega_1}(x) + \cdots + u_{10}(t)\chi_{\Omega_{10}}(x)$, where $\{\Omega_i \,|\, i = 1,...,10\}$ is a uniform partition of Ω (especially, $(\mathcal{B}^\star p)_i(t) = \int_\Omega \chi_i(x)p(t,x)\,dx$ holds) and initial control $u_\circ(t) \equiv 1$.

1) *Banach fixed point method*: The first-order necessary and sufficient optimality conditions (1.81) can be refurmaleted as the equivalent fixpoint problem

$$u = \mathbb{P}\Big(\frac{1}{\sigma}\Big(\mathcal{B}'\mathcal{A}u - \mathcal{B}'\hat{p}\Big)\Big) =: F(u),$$

where $\mathbb{P}(u) = \min(\max(u,u_a),u_b)$ is the orthogonal projection on the set of admissible points U_{ad}. The optimal control $\bar{u} \in U$ can therefore be determined by the Banach fixpoint iteration $u_{k+1} = F(u_k)$ $(k > 0)$ with arbitrary initialization $u_0 \in U_{\text{ad}}$ provided that F is a contraction. Since $\mathbb{P}$ is Lipschitz-continuous with respect to the Lipschitz constant 1, we get

$$\|F(u) - F(v)\|_U \le \frac{\|\mathcal{B}'\mathcal{A}\|_{\mathcal{L}(U)}}{\sigma}\|u - v\|_U \quad \text{for all } u,v \in U,$$

so the contraction of F is guaranteed if the regularization parameter σ is sufficiently large. Except of matrix multiplications, each iteration step requires the forward solving of the state equation for $\tilde{y}(u) = \mathcal{S}u$ and the backwards solving of the adjoint equation for $\tilde{p}(\tilde{y}) = \mathcal{A}u$. As it can be observed in the left plot of Figure 1.5, the iteration indeed does not converge if σ is smaller than some critical calue $\sigma_\circ \approx 0.02$. Furthermore, the convergence speed of the iteration loop tends to zero for $\sigma \downarrow \sigma_\circ$. We therefore can make use of this method if the control term $\|u\|_U^2/2$ in the objective functional J models a control cost such as the required energy and hence shall be small. On the other hand, if we just penalize the objective functional to enforce the strict convexity property and are interested in the case $\sigma \to 0$ (the resulting controls usually are of "bang-bang"-type in this case, i.e. $u(t) \in \{u_a, u_b\}$ for almost all $t \in [0,T]$), we shall apply some other optimization technique.

2) *Projected gradient method*: A suitable steepest descent method for the control-constrained optimization problem is the projected gradient algorithm; see, [36], for instance.

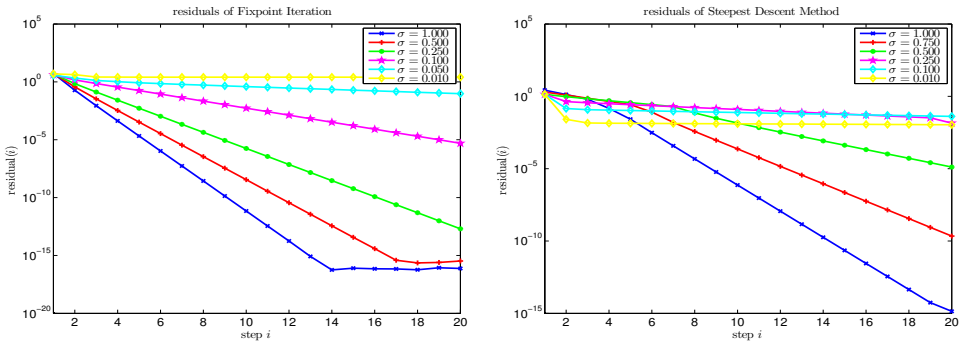

Figure 1.5. *Residuals of the Banach fixpoint iteration (left) and the projected gradient method (right) for different regularization parameters σ.*

Here, the next iteration point is given by the formula $u_{k+1} = \mathbb{P}(u_k + s_k d_k)$, where $d_k = -\nabla J(u_k) = -\sigma u_k + \mathcal{B}'(\mathcal{A} u_k + \hat{p})$ is the direction of the steepest descent of J in the current iteration point u_k and $s_k > 0$ is chosen by Algorithm 1.4. This procedure is globally convergent. However, as before, the convergence speed becomes extremely slow for $\sigma \to 0$. In addition, if the step size condition $\hat{J}(u_k + s^{(j)} d_k) \leq \hat{J}(u_k) - c/s^{(j)} \|d_k\|_U$ is just fulfilled for very small step sizes $s^{(j)}$, many evaluations of the reduced objective functional are required to test whether $\hat{J}(u_k + s^{(j)} d_k) \leq \hat{J}(u_k) - c s^{(j)} \|d_k\|_U$ is satisfied. Here, each evaluation requires to solve the state equation. Therefore, also the single iteration steps may become quite expensive. The right plot of Figure 1.5 demonstrates that also the projected gradient method cannot deal with small regularizations. In contrast to the Banach iteration, the residuals decay for arbitrarily small values of σ, but the numerical effort explodes if σ tends to zero.

3) *Primal-dual active set strategy*: This method – see Algorithm 1.1 for the infinite-dimensional case and Algorithm 1.2 for the POD discretization – solves the state and the adjoint equation simultaneously within the implicit linear scheme

$$u_{k+1}(t) = \chi_{\mathcal{A}_a^k}(t) u_a + \chi_{\mathcal{A}_b^k}(t) u_b(t) + \chi_{\mathcal{I}^k}(t) \frac{1}{\sigma} (\mathcal{B}' \mathcal{A} u_{k+1})(t)$$

f.a.a. $t \in [0, T]$. Since this technique is equivalent to a semismooth Newton procedure [24] locally superlinear convergence rates are provided. Further, the algorithm is able to deal with smaller regularizations than the other two methods presented: Reasonable computation times are provided for all $\sigma > \sigma_\circ \approx 0.0002$, see Figure 1.6. For parameters below this critical value, the bang-bang-control u oscillates between u_a and u_b at the boundary grid points of the active sets. Notice that both the critical $\sigma_\circ$ and the error between the exact solution and the suboptimal final iteration depend on the number of discretization points. The numerical effort of the simultaneous solving operations in each iteration step is significantly larger than the single solution since the initial condition for the state and the final condition for the adjoint state prevent to iteratively solve n times a system of dimension $2m$; instead, all time and space values $(y(t_i, x_j), p(t_i, x_j))$ are determined by solving a linear system of the dimension $2nm$. Here, the model order reduction techniques come into play which will lead to formidable calculation time reductions (or even make an

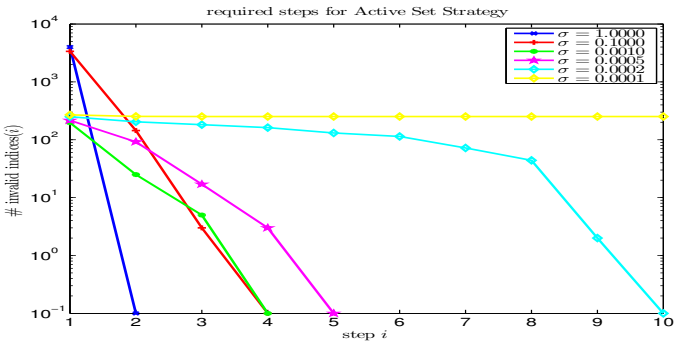

Figure 1.6. *Run 5: The number of grid points, where the previous and the actual active sets differ for different regularization parameters σ.*

execution of the primal-dual active set strategy just possible). In the following, we will make use of this optimization procedure.

ALGORITHM 1.4. **(Backtracking strategy).**

Require: Maximal number $j_{\max}$ of iterations and parameter $c \in (0,1)$.

1: Set $s^{(0)} = 1$ and $j = 1$.
2: **while** $\hat{J}(u_k + s^{(j)} d_k) > \hat{J}(u_k) - c/s^{(j)} \|d_k\|_U$ **and** $j < j_{\max}$ **do**
3: Set $s^{(j+1)} = s^{(j)}/2$ and $j = j+1$.
4: **end while**
5: **return** $s_k = s^{(j)}$

Run 6 (Different Galerkin expansions). In this run we compare the *modified* POD Galerkin expansions (1.54) for the state variable and (1.85) for the dual variable with the *standard* Galerkin approximations:

$$y^\ell(t) = \sum_{i=1}^{\ell} y_i^\ell(t)\psi_i, \quad p^\ell(t) = \sum_{i=1}^{\ell} p_i^\ell(t)\psi_i \quad \text{for } t \in [0,T]. \tag{1.101}$$

We choose the same setting as in Run 5. Let $\sigma = 0.1$. In Figures 1.7 and 1.8 we plot the optimal FE solution components $(\bar{y}^h, \bar{u}^h, \bar{p}^h, \bar{\lambda}^h)$ obtained by using the primal-dual active set strategy. We observe that the support of the multiplier $\mathcal{B}\bar{\lambda}^h$ coincides with the active set for the control variable $\mathcal{B}\bar{u}^h$. Further, the relation $\bar{u}^h = \mathbb{P}(\mathcal{B}'\bar{p}^h/\sigma)$ can be observed. As it is stated in Remark 1.28-1) the advantage of the modified Galerkin ansatz is that the ROM errors do not include the projection of the initial value on the POD space. Figure 1.9 illustrates the impact of homogenization, where we not only plot the ROM errors, but also the a-posteriori error estimate for different ℓ; compare Section 1.4.5. First we see that the ROM errors and the a-posteriori error estimate nearly coincide in all scenarios. In the left plot of Figure 1.9 the POD basis is computed from snapshots of the state equation taking the control guess $u_0 \equiv 1$. One observes that the dynamics of the corresponding homogeneous snapshots in the modified ansatz are not sufficient to decrease the control error below a level of 10^{-3} while the standard Galerkin ansatz, exploiting also the dynamics of the initial value and the inhomogeneity, induces

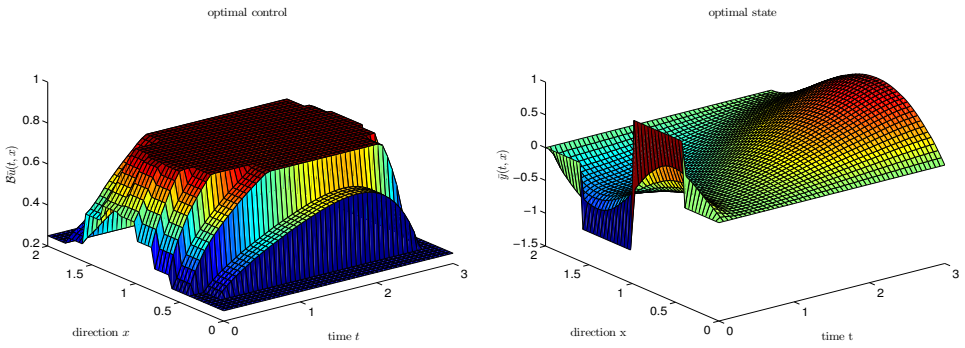

Figure 1.7. *Run 6: The optimal FE control $\mathcal{B}\bar{u}^h$ (left) and the optimal FE state $\bar{y}^h$ (right).*

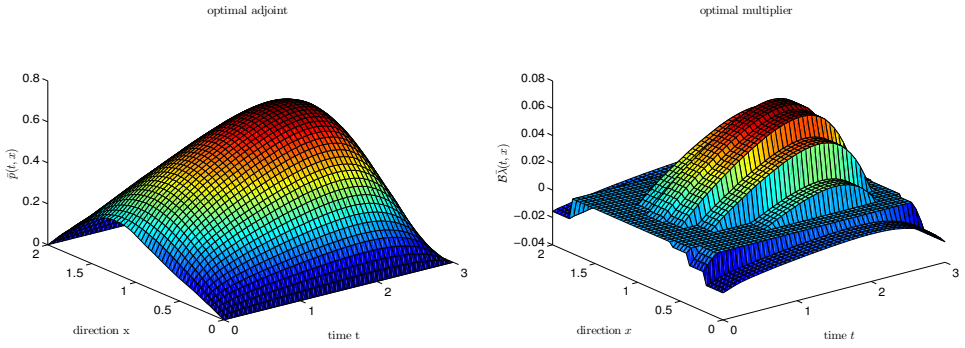

Figure 1.8. *Run 6: The optimal FE adjoint state $\bar{p}^h$ and the optimal FE Lagrange multiplier $\mathcal{B}\bar{\lambda}^h$.*

a higher dimensional POD space and leads to an error order below 10^{-6}. In the right plot of Figure 1.9 the optimal FE control $\bar{u}^h$ creates the snapshots. Here, the modified Galerkin ansatz pays: The approximation error in the standard Galerkin ansatz is dominated by the projection error of the initial value $y_\circ$ on the POD Galerkin ansatz space. This example also shows that good approximations of the reduced order model are just guaranteed in the case that the snapshots which build up the POD basis include the dynamics of the optimal state solution; otherwise, enlarging the POD basis rank does not necessarily improve the accuracy of the results. Algorithm 1.3 proposes a solution for this problem: Here, basis updates are provided if the a posteriori error estimator presented in Theorem 1.50 indicate that the control error does not decay in the current POD model. Figure 1.9 shows that these error bounds are sharp. Indeed, if the algorithm is initialized with the control guess $u_\circ \equiv 1$ and a single basis update is provided, i.e., a new POD basis is calculated with respect to the achieved suboptimal POD control $u_1^{\ell_{\max}}$. This new POD basis coincides with the POD basis associated with the best (but usually unknown) control guess $\bar{u}^h$. Thus, the resulting error decay by enlarging ℓ is the same one as in the right graphic of Figure 1.9. In Figure 1.10 the first POD basis functions are presented for the modified and standard Galerkin expansions. Consequently, the reconstruction of the initial condition $y_\circ$ with the standard Galerkin ansatz works quite well as it is demonstrated in Figure 1.11 – especially, due to the shape of the POD basis functions, no oscillations at the jump points occur as can be observed by trigonometric Fourier approximations, for instance. For the modified POD Galerkin ansatz it is neither required nor possible to build up the initial value $y_\circ$ accurately. But this is not needed, because the initial condi-

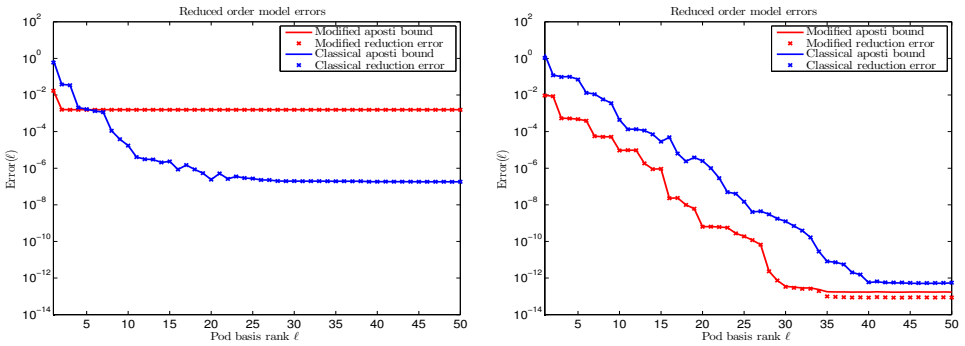

Figure 1.9. *Run 6: The ROM errors for the standard and the modified POD ansatz for initial control guesses $u_0 = 1$ (left) and $u_0 = \bar{u}^h$ (right).*

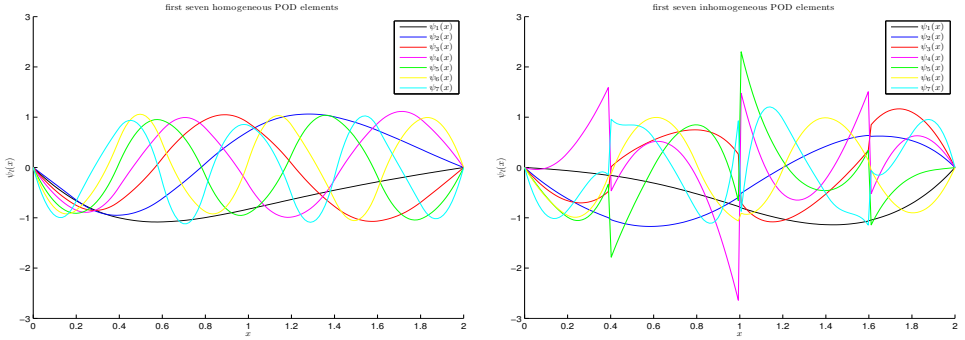

Figure 1.10. *Run 6: The first POD basis elements for the modified (left) and the standard (right) Galerkin expansion.*

tion is explicitly included in the initial condition; see (1.54). If the model data is perturbed by noise, the improvement of homogenization is even significantly stronger. For the following simulation, we add random data onto the initial value $y_\circ$. The controls gained by the modified model then reach the optimal precision 10^{-13} with 29 POD basis functions, where even 50 basis elements are not sufficient in the standard ansatz to decrease the error below a level of 10^{-11}, see Figure 1.12. We observe that the noise in the initial value is inherited to the POD basis elements of the modified Galerkin ansatz; despite of this perturbation, their shape does not differ much from those of the POD basis for the unperturbed initial conditions standard Galerkin ansatz. This is different for the standard POD Galerkin ansatz; compare Figures 1.10 and 1.13.

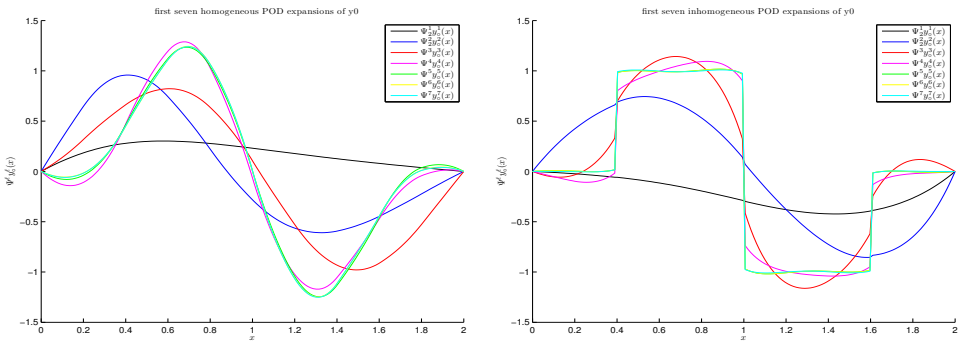

Figure 1.11. *Run 6: The reconstruction error* $\Psi^\ell y_\circ = \sum_{i=1}^{\ell} \langle y_\circ, \psi_i \rangle_H \, \psi_i$ *for the initial condition* $y_\circ$ *for the modified (left) and the standard (right) POD Galerkin expansions.*

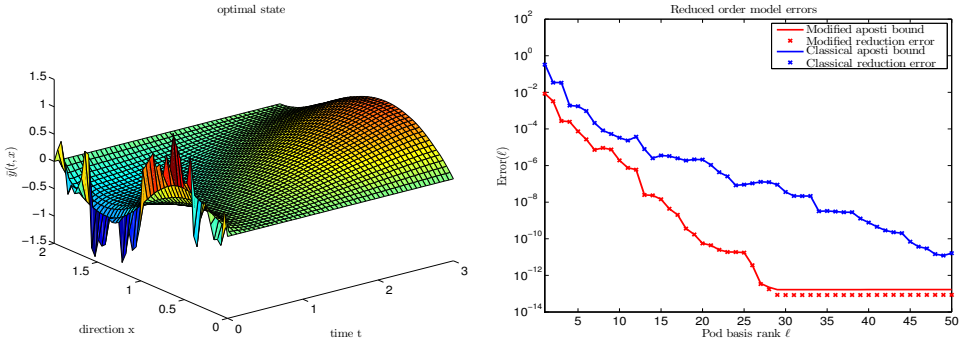

Figure 1.12. *Run 6: The optimal state solution for perturbed initial data (left) and the ROM errors for the two POD ansatzes (right).*

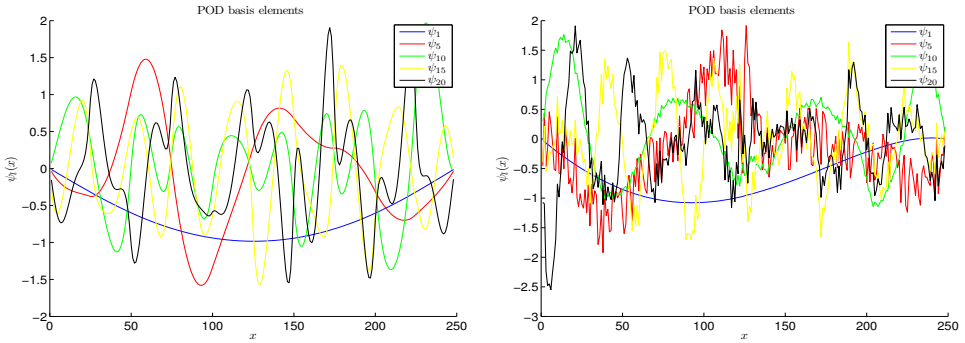

Figure 1.13. *Run 6: The first POD basis elements for the modified (left) and the standard (right) Galerkin expansion in case of the perturbed initial condition.*

Bibliography

[1] K. AFANASIEV AND M. HINZE, *Adaptive control of a wake flow using proper orthogonal decomposition*, Lecture Notes in Pure and Applied Mathematics, 216 (2001), pp. 317–332. (Cited on pp. 2, 45)

[2] A.C. ANTOULAS, *Approximation of Large-Scale Dynamical Systems*, Advances in Design and Control, Society for Industrial and Applied Mathematics, Philadelphia, 2005. (Cited on pp. 1, 10)

[3] E. ARIAN, M. FAHL, AND E.W. SACHS, *Trust-region proper orthogonal decomposition for flow control*, Tech. Report 2000-25, ICASE, 2000. (Cited on pp. 2, 45)

[4] P. ASTRID, S. WEILAND, K. WILLCOX, AND T. BACKX, *Missing point estimation in models described by proper orthogonal decomposition*, IEEE Trans. Automat. Contr., 53 (2008), pp. 2237–2251. (Cited on p. 2)

[5] J.A. ATWELL, J.T. BORGGAARD, AND B.B. KING, *Reduced-order controllers for Burgers' equation with a nonlinear observer*, International Journal of Applied Mathematics and Computer Science, 11 (2001), pp. 1311–1330. (Cited on p. 2)

[6] H.T. BANKS, M.L. JOYNER, B. WINCHESKY, AND W.P. WINFREE, *Nondestructive evaluation using a reduced-order computational methodology*, Inverse Problems, 16 (2000), pp. 1–17. (Cited on p. 2)

[7] S.C. BRENNER AND L.R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Texts in Applied Mathematics, Springer, New York, Berlin, Paris, 2008. (Cited on p. 1)

[8] D. CHAPELLE, A. GARIAH, AND J. SAINT-MARIE, *Galerkin approximation with proper orthogonal decomposition: new error estimates and illustrative examples*, Mathematical Modelling and Numerical Analysis, 46 (2012), pp. 731–757. (Cited on p. 21)

[9] A. CHATTERJEE, *An introduction to the proper orthogonal decomposition*, Current Science, 78 (2000), pp. 539–575. (Cited on p. 2)

[10] R. DAUTRAY AND J.-L. LIONS, *Mathematical Analysis and Numerical Methods for Science and Technology. Volume 5: Evolution Problems I*, Springer, Berlin, Heidelberg, New York, 2000. (Cited on pp. 21, 22)

[11] K. DECKELNICK AND M. HINZE, *Error estimates in space and time for tracking-type control of the instationary Stokes system*, International Series of Numerical Mathematics, 143 (2002), pp. 87–103. (Cited on p. 2)

[12] ———, *Semidiscretization and error estimates for distributed control of the instationary Navier-Stokes equations*, Numerische Mathematik, 97 (2004), pp. 297–320. (Cited on p. 2)

[13] L. DEDE, *Reduced basis method and a posteriori error estimation for parametrized linear-quadratic optimal control problems*, SIAM Journal on Scientific Computing, 32 (2010), pp. 997–1019. (Cited on p. 2)

[14] M. Dihlmann and B. Haasdonk, *Certified nonlinear parameter optimization with reduced basis surrogate models*, PAMM, 13 (2013), pp. 3–6. (Cited on p. 1)

[15] F. Diwoky and S. Volkwein, *Nonlinear boundary control for the heat equation utilizing proper orthogonal decomposition*, in Fast Solution of Discretized Optimization Problems, K.-H. Hoffmann, R.H.W. Hoppe, and V. Schulz, eds., vol. 138, International Series of Numerical Mathematics, 2001, pp. 73–87. (Cited on p. 41)

[16] A. L. Dontchev, W. W. Hager, A. B. Poore, and B. Yang, *Optimality, stability, and convergence in nonlinear control*, Applied Mathematics and Optimization, 31 (1995), pp. 297–326. (Cited on pp. 32, 43)

[17] L.C. Evans, *Partial Differential Equations*, Graduate studies in mathematics, American Mathematical Society, Providence (R.I.), 2008. (Cited on p. 22)

[18] M. Grepl and M. Kärcher, *A posteriori error estimation for reduced order solutions of parametrized parabolic optimal control problems*, Mathematical Modelling and Numerical Analysis, 48 (2014), pp. 1615–1638. (Cited on p. 2)

[19] M.A. Grepl, Y. Maday, N.C. Nguyen, and A.T. Patera, *Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations*, Mathematical Modelling and Numerical Analysis, 41 (2007), pp. 575–605. (Cited on p. 1)

[20] M. Gubisch and S. Volkwein, *Pod a-posteriori error analysis for optimal control problems with mixed control-state constraints*, Computational Optimization and Applications, 58 (2014), pp. 619–644. (Cited on pp. 2, 21)

[21] M. Heinkenschloss, D.C. Sorensen, and K. Sun, *Balanced truncation model reduction for a class of descriptor systems with application to the Oseen equations*, SIAM Journal on Scientific Computing, 30 (2008), pp. 1038–1063. (Cited on p. 1)

[22] S. Herkt, M. Hinze, and R. Pinnau, *Convergence analysis of Galerkin POD for linear second order evolution equations*, Electronic Transactions on Numerical Analysis, 40 (2013), pp. 321–337. (Cited on p. 21)

[23] C. Himpe and M. Ohlberger, *Cross-gramian-based combined state and parameter reduction for large-scale control systems*, Mathematical Problems in Engineering, 2014 (2014), pp. 1–13. (Cited on p. 1)

[24] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM Journal on Optimization, 13 (2003), pp. 865–888. (Cited on pp. 39, 51)

[25] M. Hinze, *A variational discretization concept in control constrained optimization: the linear-quadratic case*, Computational Optimization and Applications, 30 (2005), pp. 45–61. (Cited on p. 2)

[26] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE Constraints*, Mathematical Modelling: Theory and Applications, Springer Netherlands, 2009. (Cited on p. 39)

[27] M. Hinze and S. Volkwein, *Proper orthogonal decomposition surrogate models for nonlinear dynamical systems: error estimates and suboptimal control*, Lecture Notes in Computational Science and Engineering, 45 (2005), pp. 261–306. (Cited on pp. 2, 39)

[28] ———, *Error estimates for abstract linear-quadratic optimal control problems using proper orthogonal decomposition*, Computational Optimization and Applications, 39 (2008), pp. 319–345. (Cited on pp. 27, 30, 39, 41)

[29] P. HOLMES, J.L. LUMLEY, G. BERKOOZ, AND C.W. ROWLEY, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Cambridge Monographs on Mechanics, Cambridge University Press, 2012. (Cited on pp. 2, 9)

[30] D. HÖMBERG AND S. VOLKWEIN, *Control of laser surface hardening by a reduced-order approach using proper orthogonal decomposition*, Mathematical and Computer Modelling, 38 (2003), pp. 1003–1028. (Cited on p. 21)

[31] K. ITO AND S.S. RAVINDRAN, *A reduced basis method for control problems governed by PDEs*, in Control and Estimation of Distributed Parameter Systems, W. Desch, F. Kappel, and K. Kunisch, eds., vol. 126, Birkhäuser-Verlag, Basel, 1998, pp. 153–168. Proceedings of the International Conference in Vorau, 1996. (Cited on p. 1)

[32] M. KAHLBACHER AND S. VOLKWEIN, *Galerkin proper orthogonal decomposition methods for parameter dependent elliptic systems*, Discussiones Mathematicae: Differential Inclusions, Control and Optimization, 27 (2007), pp. 95–117. (Cited on p. 21)

[33] ———, *POD a-posteriori error based inexact SQP method for bilinear elliptic optimal control problems*, Mathematical Modelling and Numerical Analysis, 46 (2012), pp. 491–511. (Cited on pp. 2, 46)

[34] E. KAMMANN, F. TRÖLTZSCH, AND S. VOLKWEIN, *A method of a-posteriori error estimation with application to proper orthogonal decomposition*, Mathematical Modelling and Numerical Analysis, 47 (2013), pp. 555–581. (Cited on pp. 2, 46)

[35] T. KATO, *Perturbation Theory for Linear Operators*, Grundlehren der mathematischen Wissenschaften, Springer-Verlag, Berlin, Heidelberg, New York, 1980. (Cited on p. 17)

[36] C.T. KELLEY, *Iterative Methods for Optimization*, Frontiers in Applied Mathematics, SIAM, Philadelphia, 1999. (Cited on pp. 39, 50)

[37] K. KUNISCH AND S. VOLKWEIN, *Control of Burgers' equation by a reduced order approach using proper orthogonal decomposition*, Journal on Optimization Theory and Applications, 102 (1999), pp. 345–371. (Cited on p. 10)

[38] ———, *Galerkin proper orthogonal decomposition methods for parabolic problems*, Numerische Mathematik, 90 (2001), pp. 117–148. (Cited on pp. 2, 21, 30, 39)

[39] ———, *Crank-Nicolson Galerkin proper orthogonal decomposition approximations for a general equation in fluid dynamics*, in Proceedings of the 18th GAMM Seminar on Multigrid and related methods for optimization problems, 2002, pp. 97–114. (Cited on pp. 3, 21, 30, 39)

[40] ———, *Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics*, SIAM Journal on Numerical Analysis, 40 (2002), pp. 492–515. (Cited on pp. 2, 3, 20, 21, 30, 39)

[41] ———, *Proper orthogonal decomposition for optimality systems*, Mathematical Modelling and Numerical Analysis, 42 (2002), pp. 1–23. (Cited on pp. 45, 46)

[42] ———, *Optimal snapshot location for computing POD basis functions*, Mathematical Modelling and Numerical Analysis, 44 (2010), pp. 509–529. (Cited on p. 21)

[43] K. KUNISCH, S. VOLKWEIN, AND L. XIE, *HJB-POD based feedback design for the optimal control of evolution problems*, SIAM Journal on Applied Dynamical Systems, 3 (2004), pp. 701–722. (Cited on p. 2)

[44] S. LALL, J.E. MARSDEN, AND S. GLAVASKI, *A subspace approach to balanced truncation for model reduction of nonlinear control systems*, International Journal on Robust Nonlinear Control, 12 (2002), pp. 519–535. (Cited on p. 2)

[45] O. Lass and S. Volkwein, *POD Galerkin schemes for nonlinear elliptic-parabolic systems*, SIAM Journal on Scientific Computing, 35 (2013), pp. A1271–A1298. (Cited on p. 30)

[46] F. Leibfritz and S. Volkwein, *Reduced order output feedback control design for PDE systems using proper orthogonal decomposition and nonlinear semidefinite programming*, Linear Algebra and Its Applications, 415 (2006), pp. 542–575. (Cited on p. 2)

[47] H.V. Ly and H.T. Tran, *Modeling and control of physical processes using proper orthogonal decomposition*, Mathematical and Computer Modeling, 33 (2001), pp. 223–236. (Cited on p. 2)

[48] V. Mehrmann and T. Stykel, *Balanced truncation model reduction for large-scale systems in descriptor form*, Lecture Notes in Computational Science and Engineering, 45 (2005). (Cited on p. 1)

[49] F. Negri, G. Rozza, A. Manzoni, and A. Quateroni, *Reduced basis method for parametrized elliptic optimal control problems*, SIAM Journal on Scientific Computing, 35 (2013), pp. A2316–A2340. (Cited on pp. 1, 2)

[50] B. Noack, K. Afanasiev, M. Morzynski, G. Tadmor, and F. Thiele, *A hierarchy of low-dimensional models for the transient and post-transient cylinder wake*, Journal of Fluid Mechanics, 497 (2003), pp. 335–363. (Cited on p. 2)

[51] B. Noble and J.W. Daniel, *Applied Linear Algebra*, Prentice Hall, Upper Saddle River, 3rd ed. ed., 1998. (Cited on p. 10)

[52] J. Nocedal and S.J. Wright, *Numerical Optimization*, Springer Series in Operation Research, Springer, Berlin, 2006. (Cited on pp. 31, 39)

[53] M. Ohlberger and M. Schaefer, *Error control based model reduction for parameter optimization of elliptic homogenization problems*, in 1st IFAC Workshop on Control of Systems Governed by Partial Differential Equations, CPDE 2013; Paris; France; 25 September 2013 through 27 September 2013; Code 103235, Yann Le Gorrec, ed., vol. 1, International Federation of Automatic Control (IFAC), 2013, pp. 251–256. (Cited on p. 1)

[54] A.T. Patera and G. Rozza, *Reduced Basis Approximation and A Posteriori Error Estimation for Parametrized Partial Differential Equations*, MIT-Pappalardo Graduate Monographs in Mechanical Engineering, (C) MIT, Massachusetts Institute of Technology, Cambridge, MA, US, 2007. (Cited on p. 1)

[55] R. Pinnau, *Model reduction via proper orthogonal decomposition*, in Model Order Reduction: Theory, Research Aspects and Applications, W.H.A. Schilder, H.A. van der Vorst, and J. Rommes, eds., vol. 13 of Mathematics in Industry, Springer, Berlin. Heidelberg, 2008, pp. 95–109. (Cited on p. 2)

[56] O. Pironneau, *Calibration of options on a reduced basis*, Journal of Computational and Applied Mathematics, 232 (20022009), pp. 139–147. (Cited on p. 2)

[57] R. Rannacher, *Finite element solution of diffusion problems with irregular data*, Numerische Mathematik, 43 (1984), pp. 309–327. (Cited on p. 48)

[58] M. Rathinam and L. Petzold, *Dynamic iteration using reduced order models: a method for simulation of large scale modular systems*, SIAM Journal on Numerical Analysis, 40 (2002), pp. 1446–1474. (Cited on p. 2)

[59] S.S. Ravindran, *Reduced-order adaptive controllers for fluid flows using POD*, SIAM Journal on Scientific Computing, 15 (2000), pp. 457–478. (Cited on p. 2)

[60] M. Reed and B. Simon, *Methods of Modern Mathematical Physics I: Functional Analysis*, Academic Press, New York, 1980. (Cited on pp. 3, 5, 6, 14, 19, 29, 35, 45)

[61] C.W. ROWLEY, *Model reduction for fluids, using balanced proper orthogonal decomposition*, International Journal of Bifurcation and Chaos, 15 (2005), pp. 997–1013. (Cited on p. 2)

[62] E.W. SACHS AND M. SCHU, *A-priori error estimates for reduced order models in finance*, Mathematical Modelling and Numerical Analysis, 47 (2013), pp. 449–469. (Cited on pp. 2, 21, 30)

[63] E.W. SACHS AND S. VOLKWEIN, *POD Galerkin approximations in PDE-constrained optimization*, GAMM-Mitteilungen, 33 (2010), pp. 194–208. (Cited on p. 2)

[64] W.H.A. SCHILDERS, H.A. VAN DER VORST, AND J. ROMMES, *Model Order Reduction: Theory, Research Aspects and Applications*, Mathematics in Industry, Springer, 2008. (Cited on p. 1)

[65] M. SCHU, *Adaptive Trust-Region POD Methods and their Application in Finance*, PhD thesis, University of Trier, 2013. (Cited on p. 2)

[66] J.R. SINGLER, *New POD expressions, error bounds, and asymptotic results for reduced order models of parabolic PDEs*, SIAM Journal on Numerical Analysis, 52 (2014), pp. 852–876. (Cited on pp. 21, 23, 25, 26, 31, 39, 41)

[67] L. SIROVICH, *Turbulence and the dynamics of coherent structures. Parts I-II*, Quarterly of Applied Mathematics, XVL (1987), pp. 561–590. (Cited on pp. 2, 11)

[68] A. STUDINGER AND S. VOLKWEIN, *Numerical analysis of POD a-posteriori error estimation for optimal control*, International Series of Numerical Mathematics, 164 (2013), pp. 137–158. (Cited on pp. 2, 46)

[69] T. TONN, K. URBAN, AND S. VOLKWEIN, *Comparison of the reduced-basis and POD a-posteriori error estimators for an elliptic linear quadratic optimal control problem*, Mathematical and Computer Modelling of Dynamical Systems, 17 (2011), pp. 355–369. (Cited on p. 2)

[70] F. TRÖLTZSCH, *Optimal Control of Partial Differential Equations. Theory, Methods and Applications*, vol. 112, American Mathematical Society, Providence, Rhode Island, 2010. (Cited on pp. 3, 5, 14, 21, 22, 34, 35, 36, 38)

[71] F. TRÖLTZSCH AND S. VOLKWEIN, *POD a-posteriori error estimates for linear-quadratic optimal control problems*, Computational Optimization and Applications, 44 (2009), pp. 83–115. (Cited on pp. 2, 27, 30, 39, 41, 43, 44, 45)

[72] M. ULBRICH, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics, Philadelphia, 2011. (Cited on p. 39)

[73] S. VOLKWEIN, *Optimal control of a phase-field model using proper orthogonal decomposition*, Zeitschrift für Angewandte Mathematik und Mechanik, 81 (2001), pp. 83–97. (Cited on pp. 3, 9)

[74] ———, *Optimality system POD and a-posteriori error analysis for linear-quadratic problems*, Control and Cybernetics, 40 (2011), pp. 1109–1125. (Cited on pp. 2, 46)

[75] ———, *Proper Orthogonal Decomposition: Theory and Reduced-Order Modelling*. http://www.math.uni-konstanz.de/numerik/personen/volkwein/teaching/scripts.php, 2013. Lecture Notes, University of Konstanz. (Cited on p. 2)

[76] G. VOSSEN AND S. VOLKWEIN, *Model reduction techniques with a-posteriori error analysis for linear-quadratic optimal control problems*, Numerical Algebra, Control and Optimization, 2 (2012), pp. 465–485. (Cited on p. 2)

[77] K. WILLCOX AND J. PERAIRE, *Balanced model reduction via the proper orthogonal decomposition*, American Institute of Aeronautics and Astronautics, 40 (2002), pp. 2323–2330. (Cited on p. 2)

[78] K. YOSIDA, *Functional Analysis*, vol. 123 of Classics in Mathematics, Reprint of the 1980 edition, Springer-Verlag, Berlin, Heidelberg, 1995. (Cited on p. 16)

[79] K. ZHOU, J.C. DOYLE, AND K. GLOVER, *Robust and Optimal Control*, Prentice-Hall, Upper Saddle River N.J., 1996. (Cited on p. 1)