Proper Orthogonal Decomposition: Applications in Optimization and Control

S. Volkwein

joined work with M. Kahlbacher, K. Kunisch, and F. Tröltzsch

S. Volkwein, Karl-Franzens-Universität Graz, Institut für Mathematik und Wissenschaftliches Rechnen, Heinrichstrasse 36, A-8010 Graz, Austria
E-mail address: stefan.volkwein@uni-graz.at

Key words and phrases. Proper orthogonal decomposition, singular value decomposition, model reduction, error estimates, linear-quadratic regulator problems, balanced truncation, suboptimal control, a-posteriori analysis, closed-loop control

Contents

CHAPTER 1

Balanced Truncation and the POD Method

This chapter is devoted to introduce two methods of model reduction: the *balanced truncation method* and the *POD method*. The balanced truncation approach has been successfully applied in approximating the input-output behavior of linear systems, and a-posteriori error bounds can be easily computed, in particular, with respect linear-quadratic optimal control problems. Proper orthogonal decomposition (POD) is a powerful technique for model reduction of non–linear systems. It is based on a Galerkin type discretization with basis elements created from the dynamical system itself. The POD method is highly problem specific. There are first results available in the proof of a-posteriori errors [**41**]. Both methods coincide for a specific choice of snapshots, which are needed to compute a POD basis.

The chapter is organized in the following manner: In Section 1 we recall the balanced truncation method for linear, time-invariant systems. The POD method is studied in Section 2, where we discuss different application areas:

- The POD method and singular value decomposition (Section 2.1),
- POD for dynamical systems (Section 2.2 and Section 2.3),
- POD for parabolic partial differential equations (Section 2.4), and
- POD for elliptic, parameter-dependent partial differential equations (Section 2.5).

Finally, in Section 3 we discuss the relationship between the balanced truncation method and the POD method.

1. The balanced truncation method

In this section we recall the balanced truncation method. For the presentation we follow parts of the book [**46**].

1.1. Linear time-invariant dynamical systems. Let us consider the linear time-invariant system

$$\dot{x}(t) = Ax(t) + Bu(t) \text{ for } t \in (0,\infty) \quad \text{and} \quad x(0) = x_0, \tag{1.1a}$$

$$y(t) = Cx(t) \qquad \text{for } t \in [0,\infty), \tag{1.1b}$$

where $x(t) \in \mathbb{R}^{m_x}$ is called the system state, $x_0 \in \mathbb{R}^{m_x}$ is the initial condition of the system, $u(t) \in \mathbb{R}^{m_u}$ is said to be the system input and $y(t) \in \mathbb{R}^{m_y}$ is called the system output. The matrices A, B and C are assumed to have appropriate sizes.

It is helpful to analyze the linear system (1.1) through the Laplace transform.

Definition 1.1. *Let $f(t)$ be a time-varying vector. Then its* Laplace transform *is defined by*

$$\mathcal{L}[f](s) = \int_0^\infty e^{-st} f(t)\,\mathrm{d}t \quad \textit{for } s \in \mathbb{R}. \tag{1.2}$$

The Laplace transform is defined for those values of s, for which (1.2) *converges.*

The Laplace transforms of $u(t)$ and $y(t)$ are given by

$$\mathcal{L}[u](s) = \int_0^\infty e^{-st} u(t)\,\mathrm{d}t \quad \text{and} \quad \mathcal{L}[y](s) = \int_0^\infty e^{-st} y(t)\,\mathrm{d}t = C\mathcal{L}[x](s),$$

where we have used (1.1b). Note that

$$\begin{aligned}\mathcal{L}[\dot{x}](s) &= \int_0^\infty e^{-st}\dot{x}(t)\,\mathrm{d}t = -\int_0^\infty (-s)e^{-st}x(t)\,\mathrm{d}t + \left(e^{-st}x(t)\right)\Big|_{s=0}^{s=\infty} \\ &= s\mathcal{L}[x](s) - x_0.\end{aligned}$$

Therefore, the Laplace transform of the dynamical system (1.1a) yields

$$s\mathcal{L}[x](s) - x(0) = A\mathcal{L}[x](s) + B\mathcal{L}[u](s),$$

which gives

$$\mathcal{L}[x](s) = (sI - A)^{-1}x(0) + (sI - A)^{-1}B\mathcal{L}[u](s).$$

Thus,

$$\mathcal{L}[y](s) = C\mathcal{L}[x](s) = C(sI - A)^{-1}x(0) + C(sI - A)^{-1}B\mathcal{L}[u](s). \tag{1.3}$$

For $x(0) = 0$ the expression (1.3) reduces to

$$\mathcal{L}[y](s) = G(s)\mathcal{L}[u](s), \tag{1.4}$$

where

$$G(s) = C(sI - A)^{-1}B \tag{1.5}$$

is called the *transfer matrix* of the system.

Given the initial state x_0 and the input $u(t)$, the dynamical system response $x(t)$ and $y(t)$ for $t \in [0, T]$ satisfy

$$x(t) = e^{tA}x_0 + \int_0^t e^{(t-s)A}Bu(s)\,\mathrm{d}s \quad \text{and} \quad y(t) = Cx(t).$$

If $u(t) = 0$ holds for all $t \in [0, T]$, we infer that

$$x(t) = e^{(t-t_1)A}x(t_1)$$

for any t_1, $t \in [0, T]$. The matrix $e^{(t-t_1)A}$ acts as a transformation from one state to another. Therefore, $\Phi(t, t_1) = e^{(t-t_1)A}$ is often called the *state transition matrix*.

1.2. Controllability and observability. Next we turn to the essential properties *controllability* and em observability.

Definition 1.2. *The dynamical system* (1.1a) *or the pair* (A, B) *are called* controllable *if for any* $x_0 \in \mathbb{R}^{m_x}$, $t \in [0, T]$ *and final state* $x_T \in \mathbb{R}^{m_x}$ *there exists a (piecewise continuous) input* u *such that the solution to* (1.1a) *satisfies* $x(T) = x_T$. *Otherwise,* (A, B) *is said to be* uncontrollable.

Controllability can be verified as stated in the next theorem. For a proof we refer to [**46**, Theorem 3.1].

Theorem 1.3. *The following claims are equivalent:*

1) (A, B) *are controllable.*
2) *The matrix*
$$W_c(t) = \int_0^t e^{sA}BB^Te^{sA^T}\,\mathrm{d}s$$
is positive definite for every $t > 0$.
3) *The controllability matrix*
$$\mathcal{C} = \begin{bmatrix} B & AB & A^2B & \dots & A^{m_x-1}B \end{bmatrix} \in \mathbb{R}^{m_x \times (m_x m_u)}$$
has full row rank.
4) *The matrix* $[A - \lambda I | B] \in \mathbb{R}^{m_x \times (m_x + m_u)}$ *possesses full row rank for all* $\lambda \in \mathbb{C}$.
5) *Let* λ *be an eigenvalue of* A *with associated left eigenvalue* $v \neq 0$, *i.e.,* $v^HA = \lambda v^H$. *Then,* $v^HB \neq 0$.
6) *The eigenvalues of the matrix* $A + BF$ *can be freely assigned with the restriction that complex eigenvalues are in conjugate pairs) by a suitable choice of the matrix* $F \in \mathbb{R}^{m_u \times m_x}$.

Let us recall the definition of a *stable system.*

DEFINITION 1.4. 1) *The unforced system $\dot{x}(t) = Ax(t)$ is called* stable*, if the eigenvalues of A are in the open left half plane, i.e., $\Re e\lambda < 0$ for every eigenvalue λ. A matrix with this property is said to be* stable *or* Hurwitz*.*

2) *The dynamical system* (1.1a) *or (A, B) are called* stabilizable *if there exists a state-feedback $u(t) = -Kx(t)$ so that $A - BK$ is stable.*

The next result, which is proved in [**46**, Theorem 3.2], is a consequence of Theorem 1.3.

THEOREM 1.5. *The following claims are equivalent:*

1) *(A, B) are stabilizable.*
2) *The matrix $[A - \lambda I \;\; B] \in \mathbb{R}^{m_x \times (m_x + m_u)}$ has full row rank for all $\lambda \in \mathbb{C}$ with a negative real part, i.e., $\Re e\lambda < 0$.*
3) *For all $\lambda \in \mathbb{C}$ and $v \in \mathbb{R}^n \setminus \{0\}$ satisfying $v^T A = \lambda v^T$ and $\Re e\lambda \geq 0$ we have $v^T B \neq 0$.*
4) *There exists a matrix $F \in \mathbb{R}^{m_u \times m_x}$ such that $A + BF$ is Hurwitz.*

Let us now consider the notions of *observability.*

DEFINITION 1.6. *The dynamical system* (1.1) *or (A, C) are called* observable *if for any $t_1 \in (0, T]$, the initial condition $x_0 \in \mathbb{R}^{m_x}$ can be determined from the time history of the input $u(t)$ and the output $y(t)$ in the interval $[0, t_1] \subset [0, T]$. Otherwise, the system or (A, C) is said to be* unobservable*.*

For a proof of the next theorem we refer the reader to [**46**, Theorem 3.3].

THEOREM 1.7. *The following claims are equivalent:*

1) *(A, C) is observable.*
2) *The matrix*
$$W_o(t) = \int_0^t e^{sA^T} C^T C e^{sA} \, \mathrm{d}s$$
is positive definite for every $t > 0$.

(3) *The observability matrix*
$$\mathcal{O} = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{m_x - 1} \end{pmatrix} \in \mathbb{R}^{(m_x m_y) \times m_x}$$
has full column rank.

4) *The matrix*
$$\begin{pmatrix} A - \lambda I \\ C \end{pmatrix}$$
has full column rank for all $\lambda \in \mathbb{C}$.

5) *Let λ be an eigenvalue of A and $v \neq 0$ the associated right eigenvector of A, i.e., $Av = \lambda v$. Then, $Cv \neq 0$.*
6) *The eigenvalues of the matrix $A + LC$ can be freely assigned (with the restriction that complex eigenvalues are in conjugate pairs) by a suitable chosen matrix $L \in \mathbb{R}^{m_x \times m_y}$.*

DEFINITION 1.8. *We say that the system (or the pair) (C, A) is* detectable*, if there exists a matrix $L \in \mathbb{R}^{m_x \times m_y}$ such that $A + LC$ is stable.*

We have the next characterizations:

THEOREM 1.9. *The following claims are equivalent:*

1) *The pair (C, A) is detectable.*

2) *The matrix*
$$\begin{pmatrix} A-\lambda I \\ C \end{pmatrix}$$
possesses full column rank for all $\lambda \in \mathbb{C}$ *with non-negative real part.*
3) *For all* $\lambda \in \mathbb{C}$ *and* $v \in \mathbb{R}^n \setminus \{0\}$ *with* $Av = \lambda v$ *and* $\Re e(\lambda) \geq 0$ *we have* $Cv \neq 0$.
4)
5) *The pair* (A^T, C^T) *is stabilizable.*

Often the next definitions of modal controllability and observability are used.

DEFINITION 1.10. *Let* $\lambda \in \mathbb{C}$ *be an eigenvalue of* A. *Then we call* λ *a* mode of the system. *Moreover,* λ *is called* controllable *(*observable*) if* $V^H B \neq 0$ *(*$v^H C \neq 0$*) for all left (right) eigenvectors of* A *associated with the eigenvalue* λ, *i.e.,* $v^H A = \lambda v^H$ *(*$Av = \lambda v$*) and* $v \neq 0 \in \mathbb{C}^n$. *Otherwise,* λ *is said to be* uncontrollable *(*unobservable*).*

EXAMPLE 1.11 (see [**46**, pp. 52-53]). We consider the system
$$\left(\begin{array}{c|c} A & C \\ \hline C & D \end{array}\right) = \left(\begin{array}{cccc|c} \lambda_1 & 1 & 0 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 & 1 \\ 0 & 0 & \lambda_1 & 0 & \alpha \\ 0 & 0 & 0 & \lambda_2 & 1 \\ \hline 1 & 0 & 0 & \beta & 0 \end{array}\right)$$
ith $\lambda_1 \neq \lambda_2$. Then, the mode λ_1 is not controllable if $\alpha = 0$. In fact, $e_3 = (0,0,1,0)^T \in \mathbb{R}^4$ is a left eigenvalue of A, i.e., $e_3^T A = \lambda_1 e_3$. However, $e_3^T B = \alpha = 0$ holds. Thus, λ_1 is uncontrollable for $\alpha = 0$.
Notice that $e_4 = (0,0,0,1) \in \mathbb{R}^4$ is a right eigenvector of A with the associated eigenvalue λ_2, i.e., $Ae_4 = \lambda_2 e_4$. However, for $\beta = 0$ we find that $Ce_4 = \beta = 0$. Thus, λ_2 is not observable.
If $\lambda_1 = \lambda_2$ holds, the vectors e_3 and e_4 are left eigenvectors of A associated with λ_1. Hence, $v^T A = \lambda_1 v$ for $v = e_3 - \alpha e_4$, but $v^T B = \alpha - \alpha = 0$ for any $\alpha \in \mathbb{R}$. Hence, λ_1 is uncontrollable for any α. ◊

1.3. State-space realizations for transfer matrices. Let $G(s)$ be a real-rational transfer matrix. Then, we call a state-space modal (A, B, C, D) satisfying
$$G(s) = \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$$
a *realization of* $G(s)$.

DEFINITION 1.12. *We call a state-space realization* (A, B, C, D) *of* $G(s)$ minimal *if* A *has the smallest dimension.*

We have the next characterization of a minimal realization [**46**, pp. 68-69].

THEOREM 1.13. *A state-space realization* (A, B, C, D) *of* $G(s)$ *is minimal if and only if* (A, B) *is controllable and* (C, A) *is observable.*

Minimal realizations have the following property [**46**, p. 69].

THEOREM 1.14. *Let* (A_1, B_1, C_1, D) *and* (A_2, B_2, C_2, D) *be two minimal realizations of a real rational transfer matrix* $G(s)$. *Moreover, suppose that* $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{O}_1$, *and* $\mathcal{O}_2$ *are the corresponding controllability and observability matrices, respectively. Then there exists a unique non-singular matrix* T *such that*
$$A_2 = TA_1T^{-1}, \quad B_2 = TB_1, \quad C_2 = C_1T^{-1}.$$
Furthermore, T *is given by*
$$T = (\mathcal{O}_2^T\mathcal{O}_2)^{-1}\mathcal{O}_2\mathcal{O}_1 \quad or \quad T^{-1} = \mathcal{C}_1\mathcal{C}_2^T(\mathcal{C}_2\mathcal{C}_2^T)^{-1}.$$

The balanced realization method is a numerically reliable method to eliminate uncontrollable and/or unobservable states.

1.4. Lyapunov equations. To investigate the stability, controllability and Observability of the linear system one can often utilize the Lyapunov theory. For that purpose consider for $X \in \mathbb{R}^{m_x \times m_x}$ the matrix equation

$$A^T X + XA + Q = 0 \tag{1.6}$$

with given A, $Q \in \mathbb{R}^{m_x \times m_x}$. It is proved in [**46**, Chapter 2] that (1.6) has a unique solution if and only if

$$\lambda_i(A) + \overline{\lambda}_j(A) \neq 0 \quad \text{for all } i,\, j \in \{1, \ldots, m_x\} \text{ with } i \neq j,$$

where $\overline{\lambda}_j(A)$ denotes the complex conjugate of $\lambda_j(A)$. Moreover, the matrices A and X are related as stated in the next lemma.

LEMMA 1.15. *Let A be stable. Then, it follows:*

1) $X = \int_0^\infty e^{A^T s} Q e^{As}\, \mathrm{d}s$.
2) $X > 0$ *if* $Q \succ 0$, *and* $X \succeq 0$ *if* $Q \succeq 0$.
3) *if* $Q \succeq 0$, *then* (A, A) *is observable if and only if* $X \succ 0$ *holds.*

We conclude from Lemma 1.15-part 3) that for a given stable A the pair (C, A) is observable provided the solution L_o of

$$A^T L_o + L_o A + C^T C = 0 \tag{1.7}$$

is positive definite. The solution matrix $L_0 \in \mathbb{R}^{m_x \times m_x}$ is salled *observability Gramian*. Similarly, a pair (A, B) is controllable if and only if the solution L_c of

$$AL_c + L_c A^T + BB^T = 0 \tag{1.8}$$

is positive definite. The matrix $L_c \in \mathbb{R}^{m_x \times m_x}$ is the *controllability Gramian*.

If we have computed a solution to (1.6) we can say if A is stable or not. This is formulated in the next lemma. For a proof we refer to [**46**, p. 72].

LEMMA 1.16. *Let X be the solution to*(1.6). *Then,*

1) $\Re e \lambda_i(A) \leq 0$ *if* $X \succ 0$ *and* $Q \succeq 0$,
2) A *is stable if* $X \succ 0$ *and* $Q \succ 0$,
3) A *is stable if* $X \succeq 0$, $Q \succeq 0$, *and* (Q, A) *is detectable.*

1.5. Balanced realizations. In this section we concentrate on a very useful class of realizations for a given transfer matrix that is often used in control engineering and signal processing.

LEMMA 1.17. *Suppose that*

$$\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$$

is a state-space realization of a (not necessarily stable) transfer matrix $G(s)$. Let there exist a symmetric matrix

$$P = P^T = \begin{pmatrix} P_1 & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{m_x \times m_x},$$

where $P_1 \in \mathbb{R}^{k \times k}$, $k \in \{1, \ldots, m_x\}$, is non-singular so that

$$AP + PA^T + BB^T = 0.$$

We write

$$\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right) = \left(\begin{array}{cc|c} A_{11} & A_{12} & B_1 \\ A_{21} & A_{22} & B_2 \\ \hline C_1 & C_2 & D \end{array}\right)$$

with $A_{11} \in \mathbb{R}^{k \times k}$, $A_{12} \in \mathbb{R}^{k \times (m_x - k)}$, $A_{21} \in \mathbb{R}^{m_x - k) \times (m_x - k)}$, $B_1 \in \mathbb{R}^{k \times m_u}$, $B_2 \in \mathbb{R}^{(m_x - k) \times m_u}$, $C_1 \in \mathbb{R}^{m_y \times k}$, and $C_2 \in \mathbb{R}^{m_y \times (m_x - k)}$. Then,

$$\left(\begin{array}{c|c} A_{11} & B_1 \\ \hline C_1 & D \end{array}\right)$$

is also a realization of G. Moreover, (A_{11}, B_1) is controllable if A_{11}, is stable.

For a proof of Lemma 1.17 we refer to [**46**, p. 73]. An analogous result can be formulated regarding the observability of the system; see also [**46**, p. 73].

LEMMA 1.18. *Let*

$$\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$$

be a state-space realization of a (not necessarily stable) transfer matrix G. Supose that there is a symmetric matrix

$$Q = Q^T = \begin{pmatrix} Q_1 & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{m_x \times m_x},$$

where $Q_1 \in \mathbb{R}^{k\times k}$, $k \in \{1, \ldots, m_x\}$, is non-singular so that

$$QA + A^T Q + C^T C = 0.$$

Setting

$$\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right) = \left(\begin{array}{cc|c} A_{11} & A_{12} & B_1 \\ A_{21} & A_{22} & B_2 \\ \hline C_1 & C_2 & D \end{array}\right)$$

as in Lemma 1.17 *the matrix*

$$\left(\begin{array}{c|c} A_{11} & B_1 \\ \hline C_1 & D \end{array}\right)$$

is also a realization of G. Moreover, (C_1, A_{11}) is observable if A_{11} is stable.

Due to Lemmas 1.17 and 1.18 a minimal realization can be derived from a non-minimal one by elimination all states corresponding to the zero block diagonal term of the controllability Gramian P and the observability Gramian Q. In the case, where P is not block diagonal, we can proceed as described in [**46**, p. 74].

It turns out that controllability (or observability) Gramian alone does not describe very well the dominance of the system states in the input/output behavior. This motivates the introduction of a balanced realization giving balanced Gramians both for controllability and for observability. Let

$$G(s) = \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$$

be stable, i.e., A is stable. By P and Q we denote the controllability Gramian and observability Gramian, respectively. Then, by Lemma 1.15, the matrices P and Q satisfy the Lyapunov equations

$$AP + PA^T + BB^T = 0, \tag{1.9a}$$

$$A^T Q + QA + C^T C = 0 \tag{1.9b}$$

and P, $Q \succeq 0$ hold. The pair (A, B) is controllable if and only if $P \succ 0$. Moreover, (C, A) is observable if and only if $Q \succ 0$. Assume that we transform the state by utilizing a non-singular matrix $T \in \mathbb{R}^{m_x \times m_x}$, i.e., $\hat{x} = Tx$. Then, we derive the realization

$$\hat{G}(s) = \left(\begin{array}{c|c} \hat{A} & \hat{B} \\ \hline \hat{C} & \hat{D} \end{array}\right) = \left(\begin{array}{c|c} TAT^{-1} & TB \\ \hline CT^{-1} & D \end{array}\right).$$

It follows from (1.9a) and $\hat{P} = TPT^{-1}$ that

$$\begin{aligned} \hat{A}\hat{P} + \hat{P}\hat{A}^T + \hat{B}\hat{B}^T &= (TAT^{-1})(TPT^T) + (TPT^T)(T^{-1}AT)^T + (TB)(TB)^T \\ &= T\big(AP + PA^T + BB^T\big)T^T = 0 \end{aligned}$$

is satisfied. Analogously, we derive from (1.9a) and $Q = T^{-T}QT^{-1}$

$$\begin{aligned}\hat{A}^T\hat{Q} + \hat{Q}\hat{A} + \hat{C}^T\hat{C} &= (TAT^{-1})^T\,(T^{-T}QT^{-1}) + (T^{-T}QT^{-1})\,(TAT^{-1}) \\ &\quad + (CT^{-1})^T(CT^{-1}) \\ &= T^{-T}\big(A^TQ + QA + C^TC\big)T^{-1} = 0.\end{aligned}$$

Thus, the Gramians are transformed to

$$\hat{P} = TPT^T \quad \text{and} \quad \hat{Q} = T^TQT^{-1}.$$

Furthermore, $\hat{P}\hat{Q} = TPQT^{-1}$. This implies that $\hat{P}\hat{Q}$ is similar to PQ and therefore the eigenvalues of $\hat{P}\hat{Q}$ are the same as for PQ.

Next we consider a specific transformation T, which gives the eigenvalue decomposituion of the symmetric matrix PQ, i.e., we have

$$PQ = T^{-1}DT \quad \text{with} \quad D = \operatorname{diag}(\lambda_1, \ldots, \lambda_{m_x}) \in \mathbb{R}^{m_x \times m_x}.$$

Then, the columns of T^{-1} are eigenvectors of PQ corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_{m_x}$. Since P and Q are positive semi-definite, $PQ \succeq 0$ follows from Theorem 1.19 below. Therefore, $\lambda_i \geq 0$ holds for $1 \leq i \leq m_x$.

THEOREM 1.19. *Suppose that P and Q are two symmetric and positive semi-definite matrices. Then, there exists a regular matrix T such that*

$$TPT^T = \begin{pmatrix} \Sigma_1 & & & \\ & \Sigma_2 & & \\ & & 0 & \\ & & & 0 \end{pmatrix} \quad \textit{and} \quad T^{-T}QT^{-1} = \begin{pmatrix} \Sigma_1 & & & \\ & 0 & & \\ & & \Sigma_3 & \\ & & & 0 \end{pmatrix} \tag{1.10}$$

respectively, where Σ_1, Σ_2, Σ_3 are diagonal and positive definite matrices.

Theorem 1.19 is proved in [**46**, pp. 76-77]. A consequence of this theorem is the fact that the product of two positive semi-definite matrices is similar to a positive semi-definite matrix. Moreover, for any stable system

$$G(s) = \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$$

there exists a non-singular T such that

$$\hat{G}(s) = \left(\begin{array}{c|c} TAT^{-1} & TB \\ \hline CT^{-1} & D \end{array}\right)$$

has a controllability Gramian P and observability matrix Q satisfying (1.10). In case of a minimal realization the eigenvectors in the columns of T^{-1} can always be chosen such that

$$\hat{P}TPT^T = \Sigma \quad \text{and} \quad \hat{Q} = (T^{-1})^TQT^{-1} = \Sigma,$$

where $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_{m_x})$ and $\Sigma^2 = D = \operatorname{diag}(\lambda_1, \ldots, \lambda_{m_x})$. This new realization with controllability and observability Gramians $\hat{P} = \hat{Q} = \Sigma$ is called a *balanced realization*. The values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{m_x}$ are called the *Hankel singular values* of the system. In [**46**, p. 78] an algorithm is presented how we can compute a balanced realization from a minimal realization.

1.6. Model reduction by balanced truncation. The goal of this section is to reduce the order of a multivariable dynamical system, where we focus on the *balanced truncation method*.

First we introduce the following spaces of complex-valued matrix functions:

1) $\mathcal{L}_\infty(\jmath\mathbb{R})$ or simply $\mathcal{L}_\infty$ is a Banach space of matrix- (or scalar-) valued functions that are essentially bounded on $\jmath\mathbb{R}$ with the norm

$$\|G\|_\infty = \operatorname*{ess\,sup}_{\omega \in \mathbb{R}} \overline{\sigma}\big(G(\jmath\omega)\big),$$

where $\overline{\sigma}(G(\jmath\mathbb{R}))$ denotes the largest singular value of the matrix $G(\jmath\omega)$. The rational subspace of $\mathcal{L}_\infty$ — denoted by $\mathcal{RL}_\infty(\jmath\mathbb{R})$ or simply $\mathcal{RL}_\infty$ — consists of all proper and real rational transfer matrices with no poles on the imaginary axis.

2) By $\mathcal{H}_\infty$ we denote the (closed) subspace of $\mathcal{L}_\infty$ containing all functions of $\mathcal{L}_\infty$ that areanalytic and bounded in the open right half plane. On $\mathcal{H}_\infty$ we utilize the norm

$$\|G\|_\infty = \sup_{\Re es>0} \overline{\sigma}\big(G(s)\big) = \sup_{\omega\in\mathbb{R}} \overline{\sigma}\big(G(\jmath\omega)\big).$$

The real rational subspace of $\mathcal{H}_\infty$ is denoted by $\mathcal{RH}_\infty$ consisting of all proper and real rational stable transfer functions.

Let

$$G(s) = \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right) \in \mathcal{RH}_\infty$$

be a balanced realization, i.e., its controllability and observability Gramians are equal and diagonal. We denote by Σ the balanced Gramians. Then, we have

$$A\Sigma + \Sigma A^T + BB^T = 0, \tag{1.11a}$$

$$A^T\Sigma + \Sigma A + C^TC = 0. \tag{1.11b}$$

We suppose that

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \quad \text{with } \Sigma_1 \in \mathbb{R}^{k\times k},\ \Sigma_2 \in \mathbb{R}^{(m_x-k)\times(m_x-k)},$$

$k \in \{1,\dots,m_x\}$, and let

$$G(s) = \left(\begin{array}{cc|c} A_{11} & A_{12} & B_1 \\ A_{21} & A_{22} & B_2 \\ \hline C_1 & C_2 & D \end{array}\right)$$

as in Lemma 1.17. Then, if Σ_1 and Σ_2 have no diagonal entries in common, it follows that both subsystems (A_{ii}, B_i, C_i), $i = 1, 2$, are asymptotically stable [**46**, pp. 157-158]. If

$$\Sigma_1 = \operatorname{diag}\left(\sigma_1 I_{s_1},\dots,\sigma_r I_{s_r}\right) \in \mathbb{R}^{k\times k}, \qquad k = \sum_{i=1}^{r} s_i \le m_x,$$

$$\Sigma_2 = \operatorname{diag}\left(\sigma_{r+1} I_{s_{r+1}},\dots,\sigma_N I_{s_N}\right) \in \mathbb{R}^{(m_x-k)\times(m_x-k)}, \qquad \sum_{i=r+1}^{N} s_i = m_x - k,$$

where σ_i has the multiplicity s_i, $i = 1,\dots,N$, with $\sigma_1 > \sigma_2 > \ldots > \sigma_N$ then the truncated system

$$G_r(s) = \left(\begin{array}{c|c} A_{11} & B_1 \\ \hline C_1 & D \end{array}\right)$$

is balanced and asymptotically stable. Furthermore,

$$\|G - G_r\|_\infty \le 2\sum_{i=r+1}^{N} \sigma_i$$

and, in particular,

$$\|G - G_{N-1}\|_\infty = 2\sigma_N,$$

see [**46**, pp. 159-160].

2. The POD Method

In this section we introduce the POD method in the Euclidean space $\mathbb{R}^m$ and study the close connection to the SVD of rectangular matrices; see [**20**]. We also refer to the monograph [**12**].

2.1. POD and singular value decomposition. Let $Y = [y_1, \dots, y_n]$ be a real-valued $m \times n$ matrix of rank $d \leq \min\{m, n\}$ with columns $y_j \in \mathbb{R}^m$, $1 \leq j \leq n$. Consequently,

$$\bar{y} = \frac{1}{n} \sum_{j=1}^{n} y_j \tag{1.12}$$

can be viewed as the column-averaged mean of the matrix Y.

SVD [**33**] guarantees the existence of real numbers $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d > 0$ and orthogonal matrices $U \in \mathbb{R}^{m\times m}$ with columns $\{u_i\}_{i=1}^m$ and $V \in \mathbb{R}^{n\times n}$ with columns $\{v_i\}_{i=1}^n$ such that

$$U^T Y V = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} =: \Sigma \in \mathbb{R}^{m\times n}, \tag{1.13}$$

where $D = \operatorname{diag}(\sigma_1, \dots, \sigma_d) \in \mathbb{R}^{d\times d}$ and the zeros in (1.13) denote matrices of appropriate dimensions. Moreover the vectors $\{u_i\}_{i=1}^d$ and $\{v_i\}_{i=1}^d$ satisfy

$$Y v_i = \sigma_i u_i \quad \text{and} \quad Y^T u_i = \sigma_i v_i \quad \text{for } i = 1, \dots, d. \tag{1.14}$$

They are eigenvectors of YY^T and Y^TY, respectively, with eigenvalues $\lambda_i = \sigma_i^2 > 0$, $i = 1, \dots, d$. The vectors $\{u_i\}_{i=d+1}^m$ and $\{v_i\}_{i=d+1}^n$ (if $d < m$ respectively $d < n$) are eigenvectors of YY^T and Y^TY with eigenvalue 0.

From (1.13) we deduce that

$$Y = U \Sigma V^T.$$

It follows that Y can also be expressed as

$$Y = U^d D (V^d)^T, \tag{1.15}$$

where $U^d \in \mathbb{R}^{m\times d}$ and $V^d \in \mathbb{R}^{n\times d}$ are given by

$$\begin{aligned} U^d_{ij} &= U_{ij} \quad \text{for } 1 \leq i \leq m,\ 1 \leq j \leq d, \\ V^d_{ij} &= V_{ij} \quad \text{for } 1 \leq i \leq n,\ 1 \leq j \leq d. \end{aligned}$$

Setting $B^d = D(V^d)^T \in \mathbb{R}^{d\times n}$ we can write (1.15) in the form

$$Y = U^d B^d \quad \text{with } B^d = D(V^d)^T \in \mathbb{R}^{d\times n}.$$

Thus, the column space of Y can be represented in terms of the d linearly independent columns of U^d. The coefficients in the expansion for the columns y_j, $j = 1, \dots, n$, in the basis $\{u_i\}_{i=1}^d$ are given by the jth-column of B^d. Since U is orthogonal, we find that

$$\begin{aligned} y_j = \sum_{i=1}^{d} B^d_{ij} U^d_{\cdot,i} = \sum_{i=1}^{d} \big(D(V^d)^T\big)_{ij} u_i = \sum_{i=1}^{d} \big(\underbrace{(U^d)^T U^d}_{=I^d \in \mathbb{R}^{d\times d}} D(V^d)^T \big)_{ij} u_i \\ \overset{(1.15)}{=} \sum_{i=1}^{d} \big((U^d)^T Y\big)_{ij} u_i = \sum_{i=1}^{d} \Big(\underbrace{\sum_{k=1}^{m} U^d_{ki} Y_{kj}}_{=u_i^T y_j} \Big) u_i = \sum_{i=1}^{d} \langle u_i, y_j \rangle_{\mathbb{R}^m} u_i, \end{aligned}$$

where $\langle \cdot, \cdot \rangle_{\mathbb{R}^m}$ denotes the canonical inner product in $\mathbb{R}^m$. Thus,

$$y_j = \sum_{i=1}^{d} \langle y_j, u_i \rangle_{\mathbb{R}^m} u_i \quad \text{for } j = 1, \dots, n \tag{1.16}$$

Let us now interprete SVD in terms of POD. One of the central issues of POD is the reduction of data expressing their *essential information* by means of a few basis vectors. The problem of approximating all

spatial coordinate vectors y_j of Y simultaneously by a single, normalized vector as well as possible can be expressed as

$$(\mathbf{P}^1) \qquad \max_{u\in\mathbb{R}^m} \sum_{j=1}^{n} \big|\langle y_j, u\rangle_{\mathbb{R}^m}\big|^2 \quad \text{subject to (s.t.)} \quad \|u\|_{\mathbb{R}^m}^2 = 1,$$

where $\|u\|_{\mathbb{R}^m} = \sqrt{\langle u, u\rangle_{\mathbb{R}^m}}$ for $u \in \mathbb{R}^m$.

Note that $(\mathbf{P}^1)$ is a constrained optimization problem that can be solved by considering first-order necessary optimality conditions. For that purpose let $\mathcal{L} : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ be the Lagrange functional associated with $(\mathbf{P}^1)$, i.e.,

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

Suppose that $u \in \mathbb{R}^m$ is a solution to $(\mathbf{P}^1)$. Then, a first-order necessary optimality condition is given by

$$\nabla\mathcal{L}(u,\lambda) \overset{!}{=} 0 \quad \text{in } \mathbb{R}^m \times \mathbb{R}.$$

We compute the gradient of $\mathcal{L}$ with respect to u:

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial u_i}(u,\lambda) &= \frac{\partial}{\partial u_i}\left(\sum_{j=1}^{n}\left|\sum_{k=1}^{m} Y_{kj}u_k\right|^2 + \lambda\left(1 - \sum_{k=1}^{m} u_k^2\right)\right) \\
&= 2\sum_{j=1}^{n}\left(\sum_{k=1}^{m} Y_{kj}u_k\right)Y_{ij} - 2\lambda u_i \\
&= 2\sum_{k=1}^{m}\Bigg(\underbrace{\sum_{j=1}^{n} Y_{ij}Y_{jk}^T}_{=(YY^T)_{ik}}\, u_k\Bigg) - 2\lambda u_i.
\end{aligned}$$

Thus,

$$\nabla_u\mathcal{L}(u,\lambda) = 2\big(YY^Tu - \lambda u\big) \overset{!}{=} 0 \quad \text{in } \mathbb{R}^m. \tag{1.17}$$

Equation (1.17) yields the eigenvalue problem

$$YY^Tu = \lambda u \quad \text{in } \mathbb{R}^m. \tag{1.18a}$$

Notice that $YY^T \in \mathbb{R}^{m\times m}$ is a symmetric matrix satisfying

$$u^T(YY^T)u = (Y^Tu)^TY^Tu = \|Y^Tu\|_{\mathbb{R}^n}^2 \geq 0 \quad \text{for all } u \in \mathbb{R}^m.$$

Thus, YY^T is positive semi-definite. It follows that YY^T possesses m non-negative eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0$ and the corresponding eigenvectors can be chosen such that they are pairwise orthonormal.

From $\frac{\partial\mathcal{L}}{\partial\lambda}(u,\lambda) \overset{!}{=} 0$ in $\mathbb{R}$ we infer the constraint

$$\|u\|_{\mathbb{R}^m} = 1. \tag{1.18b}$$

Due to SVD the vector u_1 solves (1.18) and

$$\begin{aligned}\sum_{j=1}^{n}\left|\langle y_j,u_1\rangle_{\mathbb{R}^m}\right|^2&=\sum_{j=1}^{n}\langle y_j,u_1\rangle_{\mathbb{R}^m}\langle y_j,u_1\rangle_{\mathbb{R}^m}=\sum_{j=1}^{n}\left\langle\langle y_j,u_1\rangle_{\mathbb{R}^m}y_j,u_1\right\rangle_{\mathbb{R}^m}\\&=\left\langle\sum_{j=1}^{n}\langle y_j,u_1\rangle_{\mathbb{R}^m}y_j,u_1\right\rangle_{\mathbb{R}^m}=\left\langle\sum_{j=1}^{n}\left(\sum_{k=1}^{m}Y_{kj}(u_1)_k\right)y_j,u_1\right\rangle_{\mathbb{R}^m}\\&=\left\langle\sum_{k=1}^{m}\left(\sum_{j=1}^{n}Y_{\cdot,j}Y_{jk}^T(u_1)_k\right),u_1\right\rangle_{\mathbb{R}^m}=\left\langle YY^Tu_1,u_1\right\rangle_{\mathbb{R}^m}\\&=\lambda_1\left\langle u_1,u_1\right\rangle_{\mathbb{R}^m}=\lambda_1\left\|u_1\right\|_{\mathbb{R}^m}^2=\lambda_1.\end{aligned}$$

We next prove that u_1 solves $(\mathbf{P}^1)$. Suppose that $\tilde{u}\in\mathbb{R}^m$ is an arbitrary vector with $\|\tilde{u}\|_{\mathbb{R}^m}=1$. Since $\{u_i\}_{i=1}^m$ is an orthonormal basis in $\mathbb{R}^m$, we have

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

Thus,

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

Consequently, u_1 solves $(\mathbf{P}^1)$ and $\operatorname{argmax}(\mathbf{P}^1)=\sigma_1^2=\lambda_1$.

If we look for a second vector, orthogonal to u_1 that again describes the data set $\{y_i\}_{i=1}^n$ as well as possible then we need to solve

$$(\mathbf{P}^2)\qquad\max_{u\in\mathbb{R}^m}\sum_{j=1}^{n}\left|\langle y_j,u\rangle_{\mathbb{R}^m}\right|^2\quad\text{s.t.}\quad\|u\|_{\mathbb{R}^m}=1\text{ and }\langle u,u_1\rangle_{\mathbb{R}^m}=0.$$

SVD implies that u_2 is a solution to $(\mathbf{P}^2)$ and $\operatorname{argmax}(\mathbf{P}^2)=\sigma_2^2=\lambda_2$. In fact, u_2 solves the first-order necessary optimality conditions (1.18) and for

$$\tilde{u}=\sum_{i=1}^{m}\langle\tilde{u},u_i\rangle_{\mathbb{R}^m}u_i\in\operatorname{span}\{u_1\}^{\perp}$$

we have

$$\sum_{j=1}^{n} \left|\langle y_j, \tilde{u}\rangle_{\mathbb{R}^m}\right|^2 \leq \lambda_2 = \sum_{j=1}^{n} \left|\langle y_j, u_2\rangle_{\mathbb{R}^m}\right|^2.$$

Clearly this procedure can be continued by finite induction. We summarize our results in the following theorem.

THEOREM 2.1. *Let $Y = [y_1, \ldots, y_n] \in \mathbb{R}^{m\times n}$ be a given matrix with rank $d \leq \min\{m, n\}$. Further, let $Y = U\Sigma V^T$ be the singular value decomposition of Y, where $U = [u_1, \ldots, u_m] \in \mathbb{R}^{m\times m}$, $V = [v_1, \ldots, v_n] \in \mathbb{R}^{n\times n}$ are orthogonal matrices and the matrix $\Sigma \in \mathbb{R}^{m\times n}$ has the form as* (1.13)*. Then, for any $\ell \in \{1, \ldots, d\}$ the solution to*

$$\max_{\tilde{u}_1,\ldots,\tilde{u}_\ell \in \mathbb{R}^m} \sum_{i=1}^{\ell}\sum_{j=1}^{n} \left|\langle y_j, \tilde{u}_i\rangle_{\mathbb{R}^m}\right|^2 \quad \text{s.t.} \quad \langle \tilde{u}_i, \tilde{u}_j\rangle_{\mathbb{R}^m} = \delta_{ij} \text{ for } 1 \leq i, j \leq \ell \tag{1.19}$$

is given by the singular vectors $\{u_i\}_{i=1}^{\ell}$, i.e., by the first ℓ columns of U. Moreover,

$$\operatorname{argmax}(1.19) = \sum_{i=1}^{\ell} \sigma_i^2 = \sum_{i=1}^{\ell} \lambda_i.$$

PROOF. Since (1.19) is an equality constrained optimization problem, we introduce the Lagrangian

$$\mathcal{L} : \underbrace{\mathbb{R}^m \times \ldots \times \mathbb{R}^m}_{\ell\text{-times}} \times \mathbb{R}^{\ell\times\ell}$$

by

$$\mathcal{L}(\psi_1, \ldots, \psi_\ell, \Lambda) = \sum_{i=1}^{\ell}\sum_{j=1}^{n} \left|\langle y_j, \psi_i\rangle_{\mathbb{R}^m}\right|^2 + \sum_{i,j=1}^{\ell} \lambda_{ij}\left(\delta_{ij} - \langle \psi_i, \psi_j\rangle_{\mathbb{R}^m}\right)$$

for $\psi_1, \ldots, \psi_\ell \in \mathbb{R}^m$ and $\Lambda = ((\lambda_{ij})) \in \mathbb{R}^{\ell\times\ell}$. First-order necessary optimality conditions for (1.19) are given by

$$\frac{\partial \mathcal{L}}{\partial \psi_k}(\psi_1, \ldots, \psi_\ell, \Lambda)\delta\psi_k = 0 \quad \text{for all } \delta\psi_k \in \mathbb{R}^m \text{ and } k \in \{1, \ldots, \ell\}. \tag{1.20}$$

From

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \psi_k}(\psi_1, \ldots, \psi_\ell, \Lambda)\delta\psi_k &= 2\sum_{i=1}^{\ell}\sum_{j=1}^{n} \langle y_j, \psi_i\rangle_{\mathbb{R}^m}\langle y_j, \delta\psi_k\rangle_{\mathbb{R}^m}\delta_{ik} \\
&\quad - \sum_{i,j=1}^{\ell} \lambda_{ij}\langle \psi_i, \delta\psi_k\rangle_{\mathbb{R}^m}\delta_{jk} - \sum_{i,j=1}^{\ell} \lambda_{ij}\langle \delta\psi_k, \psi_j\rangle_{\mathbb{R}^m}\delta_{ki} \\
&= 2\sum_{j=1}^{n} \langle y_j, \psi_k\rangle_{\mathbb{R}^m}\langle y_j, \delta\psi_k\rangle_{\mathbb{R}^m} - \sum_{i=1}^{\ell} (\lambda_{ik} + \lambda_{ki})\,\langle \psi_i, \delta\psi_k\rangle_{\mathbb{R}^m} \\
&= \left\langle 2\sum_{j=1}^{n} \langle y_j, \psi_k\rangle_{\mathbb{R}^m}\, y_j - \sum_{i=1}^{\ell} (\lambda_{ik} + \lambda_{ki})\,\psi_i, \delta\psi_k \right\rangle_{\mathbb{R}^m}
\end{aligned}$$

and (1.20) we infer that

$$\sum_{j=1}^{n} \langle y_j, \psi_k\rangle_{\mathbb{R}^m}\, y_j = \frac{1}{2}\sum_{i=1}^{\ell} (\lambda_{ik} + \lambda_{ki})\,\psi_i \quad \text{in } \mathbb{R}^m \text{ and for all } k \in \{1, \ldots, \ell\}. \tag{1.21}$$

Note that

$$YY^T\psi = \sum_{j=1}^{n} \langle y_j, \psi\rangle_{\mathbb{R}^m}\, y_j \quad \text{for } \psi \in \mathbb{R}^m.$$

Thus, condition (1.21) can be expressed as

$$YY^T\psi_k = \frac{1}{2}\sum_{i=1}^{\ell}(\lambda_{ik}+\lambda_{ki})\,\psi_i \quad \text{in } \mathbb{R}^m \text{ and for all } k\in\{1,\dots,\ell\}. \tag{1.22}$$

Now we proceed by induction. For $\ell = 1$ we have $k = 1$. It follows from (1.22) that

$$YY^T\psi_1 = \lambda_1\psi_1 \quad \text{in } \mathbb{R}^m \tag{1.23}$$

with $\lambda_1 = \lambda_{11}$. Next we suppose that for $\ell \geq 1$ the first-order optimality conditions are given by

$$YY^T\psi_k = \lambda_k\psi_k \quad \text{in } \mathbb{R}^m \text{ and for all } k\in\{1,\dots,\ell\}. \tag{1.24}$$

We want to show that the first-order necessary optimality conditions for a POD basis $\{\psi_i\}_{i=1}^{\ell+1}$ of rank $\ell+1$ are given by

$$YY^T\psi_k = \lambda_k\psi_k \quad \text{in } \mathbb{R}^m \text{ and for all } k\in\{1,\dots,\ell+1\}. \tag{1.25}$$

By assumption we have (1.24). Thus, we only have to prove that

$$YY^T\psi_{\ell+1} = \lambda_{\ell+1}\psi_{\ell+1} \quad \text{in } \mathbb{R}^m. \tag{1.26}$$

Due to (1.22) we have

$$YY^T\psi_{\ell+1} = \frac{1}{2}\sum_{i=1}^{\ell+1}(\lambda_{i,\ell+1}+\lambda_{\ell+1,i})\,\psi_i \quad \text{in } \mathbb{R}^m. \tag{1.27}$$

Since $\{\psi_i\}_{i=1}^{\ell+1}$ is a POD basis we have $\langle\psi_{\ell+1},\psi_j\rangle_{\mathbb{R}^m} = 0$ for $1\leq j\leq\ell$. Using (1.24) and the symmetry of YY^T we have for any $j\in\{1,\dots,\ell\}$

$$\begin{aligned}0 &= \lambda_j\langle\psi_{\ell+1},\psi_j\rangle_{\mathbb{R}^m} = \langle\psi_{\ell+1},YY^T\psi_j\rangle_{\mathbb{R}^m} = \langle YY^T\psi_{\ell+1},\psi_j\rangle_{\mathbb{R}^m}\\ &= \frac{1}{2}\sum_{i=1}^{\ell+1}(\lambda_{i,\ell+1}+\lambda_{\ell+1,i})\langle\psi_i,\psi_j\rangle_{\mathbb{R}^m} = (\lambda_{j,\ell+1}+\lambda_{\ell+1,j}).\end{aligned}$$

This gives

$$\lambda_{\ell+1,i} = -\lambda_{i,\ell+1} \quad \text{for any } i\in\{1,\dots,\ell\}. \tag{1.28}$$

Inserting (1.28) into (1.27) we obtain

$$\begin{aligned}YY^T\psi_{\ell+1} &= \frac{1}{2}\sum_{i=1}^{\ell}(\lambda_{i,\ell+1}+\lambda_{\ell+1,i})\,\psi_i + \lambda_{\ell+1,\ell+1}\,\psi_{\ell+1}\\ &= \frac{1}{2}\sum_{i=1}^{\ell}(\lambda_{i,\ell+1}-\lambda_{i,\ell+1})\,\psi_i + \lambda_{\ell+1,\ell+1}\,\psi_{\ell+1} = \lambda_{\ell+1,\ell+1}\,\psi_{\ell+1}.\end{aligned}$$

Setting $\lambda_{\ell+1} = \lambda_{\ell+1,\ell+1}$ we obtain (1.26).
Summarizing, the necessary optimaity conditions for (1.19) are given by the symmetric $m\times m$ eigenvalue problem

$$YY^T u_i = \lambda_i u_i \quad \text{for } i = 1,\dots,\ell. \tag{1.29}$$

It follows from SVD that $\{u_i\}_{i=1}^{\ell}$ solves (1.29). The proof that $\{u_i\}_{i=1}^{\ell}$ is a solution to (1.19) and that $\operatorname{argmax}(1.19) = \sum_{i=1}^{\ell}\sigma_i^2$ holds is analogous to the proof for $(\mathbf{P}^1)$; see Exercise 1.1). □

Motivated by the previous theorem we give the next definition.

Definition 2.2. *For $\ell\in\{1,\dots,d\}$ the vectors $\{u_i\}_{i=1}^{\ell}$ are called* POD basis of rank ℓ.

The following result states that for every $\ell\leq d$ the approximation of the columns of Y by the first ℓ singular vectors $\{u_i\}_{i=1}^{\ell}$ is optimal in the mean among all rank ℓ approximations to the columns of Y.

COROLLARY 2.3 (Optimality of the POD basis). *Let all hypotheses of Theorem 2.1 be satisfied. Suppose that $\hat{U}^d \in \mathbb{R}^{m\times d}$ denotes a matrix with pairwise orthonormal vectors $\hat{u}_i$ and that the expansion of the columns of Y in the basis $\{\hat{u}_i\}_{i=1}^d$ be given by*

$$Y = \hat{U}^d C^d, \quad \text{where } C^d_{ij} = \langle \hat{u}_i, y_j\rangle_{\mathbb{R}^m} \text{ for } 1 \le i \le d, 1 \le j \le n.$$

Then for every $\ell \in \{1,\dots,d\}$ we have

$$\|Y - U^\ell B^\ell\|_F \le \|Y - \hat{U}^\ell C^\ell\|_F. \tag{1.30}$$

In (1.30), $\|\cdot\|_F$ denotes the Frobenius norm given by

$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |A_{ij}|^2} = \sqrt{\operatorname{trace}\left(A^T A\right)} \quad \text{for } A \in \mathbb{R}^{m\times n},$$

the matrix U^ℓ denotes the first ℓ columns of U, B^ℓ the first ℓ rows of B and similarly for $\hat{U}^\ell$ and C^ℓ.

REMARK 2.4. Notice that

$$\begin{aligned}\|Y - \hat{U}^\ell C^\ell\|_F^2 &= \sum_{i=1}^m \sum_{j=1}^n \left| Y_{ij} - \sum_{k=1}^\ell \hat{U}^\ell_{ik} C_{kj}\right|^2 = \sum_{j=1}^n \sum_{i=1}^m \left| Y_{ij} - \sum_{k=1}^\ell \langle \hat{u}_k, y_j\rangle_{\mathbb{R}^m} \hat{U}^\ell_{ik}\right|^2 \\ &= \sum_{j=1}^n \left\| y_j - \sum_{k=1}^\ell \langle y_j, \hat{u}_k\rangle_{\mathbb{R}^m} \hat{u}_k\right\|^2_{\mathbb{R}^m}.\end{aligned}$$

Analogously,

$$\|Y - U^\ell B^\ell\|_F^2 = \sum_{j=1}^n \left\| y_j - \sum_{k=1}^\ell \langle y_j, u_k\rangle_{\mathbb{R}^m} u_k\right\|^2_{\mathbb{R}^m}.$$

Thus, (1.30) implies that

$$\sum_{j=1}^n \left\| y_j - \sum_{k=1}^\ell \langle y_j, u_k\rangle_{\mathbb{R}^m} u_k\right\|^2_{\mathbb{R}^m} \le \sum_{j=1}^n \left\| y_j - \sum_{k=1}^\ell \langle y_j, \hat{u}_k\rangle_{\mathbb{R}^m} \hat{u}_k\right\|^2_{\mathbb{R}^m}$$

for any other set $\{\hat{u}_i\}_{i=1}^\ell$ of ℓ pairwise orthonormal vectors. Hence, the POD basis of rank ℓ can also be determined by solving

$$\min_{\tilde{u}_1,\dots,\tilde{u}_\ell \in \mathbb{R}^m} \sum_{j=1}^n \left\| y_j - \sum_{i=1}^\ell \langle y_j, \tilde{u}_i\rangle_{\mathbb{R}^m} \tilde{u}_i\right\|^2_{\mathbb{R}^m} \text{ s.t. } \langle \tilde{u}_i, \tilde{u}_j\rangle_{\mathbb{R}^m} = \delta_{ij},\ 1 \le i,j \le \ell. \tag{1.31}$$

◊

PROOF OF COROLLARY 2.3. Note that (see Exercise 1.2))

$$\|Y - \hat{U}^\ell C^\ell\|_F^2 = \|\hat{U}^d (C^d - C_0^\ell)\|_F^2 = \|C^d - C_0^\ell\|_F^2 = \sum_{i=\ell+1}^d \sum_{j=1}^n |C^d_{ij}|^2,$$

where $C_0^\ell \in \mathbb{R}^{d\times n}$ results from $C \in \mathbb{R}^{d\times n}$ by replacing the last $d-\ell$ rows by 0. Similarly,

$$\begin{aligned}\|Y - U^\ell B^\ell\|_F^2 &= \|U^k(B^d - B_0^\ell)\|_F^2 = \|B^d - B_0^\ell\|_F^2 = \sum_{i=\ell+1}^{d}\sum_{j=1}^{n} \left|B_{ij}^d\right|^2 \\ &= \sum_{i=\ell+1}^{d}\sum_{j=1}^{n} \left|\langle y_j, u_i\rangle_{\mathbb{R}^m}\right|^2 \\ &= \sum_{i=\ell+1}^{d}\sum_{j=1}^{n} \left\langle \langle y_j, u_i\rangle_{\mathbb{R}^m} y_j, u_i\right\rangle_{\mathbb{R}^m} = \sum_{i=\ell+1}^{d} \langle YY^T u_i, u_i\rangle_{\mathbb{R}^m} \\ &= \sum_{i=\ell+1}^{d} \sigma_i^2,\end{aligned} \tag{1.32}$$

By Theorem 2.1 the vectors $u_1, \ldots, u_\ell$ solve (1.19). From (1.32),

$$\|Y\|_F^2 = \|\hat{U}^d C^d\|_F^2 = \|C^d\|_F^2 = \sum_{i=1}^{d}\sum_{j=1}^{n} \left|C_{ij}^d\right|^2$$

and

$$\|Y\|_F^2 = \|U^d B^d\|_F^2 = \|B^d\|_F^2 = \sum_{i=1}^{d}\sum_{j=1}^{n} \left|B_{ij}^d\right|^2 = \sum_{i=1}^{d} \sigma_i^2$$

we infer that

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

which gives (1.30). □

REMARK 2.5. It follows from Corollary 2.3 that the POD basis of rank ℓ is optimal in the sense of representing in the mean the columns $\{y_j\}_{j=1}^n$ of Y as a linear combination by an orthonormal basis of rank ℓ:

$$\sum_{i=1}^{\ell}\sum_{j=1}^{n} \left|\langle y_j, u_i\rangle_{\mathbb{R}^m}\right|^2 = \sum_{i=1}^{\ell} \sigma_i^2 = \sum_{i=1}^{\ell} \lambda_i \ge \sum_{i=1}^{\ell}\sum_{j=1}^{n} \left|\langle y_j, \hat{u}_i\rangle_{\mathbb{R}^m}\right|^2$$

for any other set of orthonormal vectors $\{\hat{u}_i\}_{i=1}^\ell$. ◊

The next corollary states that the POD coefficients are uncorrelated.

COROLLARY 2.6 (Uncorrelated POD coefficients). *Let all hypotheses of Theorem* 2.1 *hold. Then.*

$$\sum_{j=1}^{n} \langle y_j, u_i\rangle_{\mathbb{R}^m} \langle y_j, u_k\rangle_{\mathbb{R}^m} = \sum_{j=1}^{n} B_{ij}^\ell B_{kj}^\ell = \sigma_i^2 \delta_{ik} \quad \textit{for } 1 \le i, k \le \ell.$$

PROOF. The claim follows from (1.29) and $\langle u_i, u_k\rangle_{\mathbb{R}^m} = \delta_{ik}$ for $1 \le i, k \le \ell$:

$$\sum_{j=1}^{n} \langle y_j, u_i\rangle_{\mathbb{R}^m} \langle y_j, u_k\rangle_{\mathbb{R}^m} = \Big\langle \underbrace{\sum_{j=1}^{n} \langle y_j, u_i\rangle_{\mathbb{R}^m} y_j}_{=YY^T u_i}, u_k \Big\rangle_{\mathbb{R}^m} = \langle \sigma_i^2 u_i, u_k\rangle_{\mathbb{R}^m} = \sigma_i^2 \delta_{ik}.$$

□

Next we turn to the practical computation of a POD-basis of rank ℓ. If $n < m$ then one can determine the POD basis of rank ℓ as follows: Compute the eigenvectors $v_1, \dots, v_\ell \in \mathbb{R}^n$ by solving the symmetric $n \times n$ eigenvalue problem

$$Y^T Y v_i = \lambda_i v_i \quad \text{for } i = 1, \dots, \ell \tag{1.33}$$

and set, by (1.14),

$$u_i = \frac{1}{\sqrt{\lambda_i}} Y v_i \quad \text{for } i = 1, \dots, \ell.$$

For historical reasons [**38**] this method of determing the POD-basis is sometimes called the *method of snapshots*. On the other hand, if $m < n$ holds, we can obtain the POD basis by solving the $m \times m$ eigenvalue problem (1.29).

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 system Y, which is expressed by

$$\mathcal{E}(\ell) = \frac{\sum_{i=1}^{\ell} \lambda_i}{\sum_{i=1}^{d} \lambda_i}.$$

Let us mention that POD is also called *Principal Component Analysis* (PCA) and *Karhunen-Loève Decomposition*.

Let us endow the Euclidean space $\mathbb{R}^m$ with the weighted inner product

$$\langle u, \tilde{u} \rangle_W = u^T W \tilde{u} = \langle u, W\tilde{u} \rangle_{\mathbb{R}^m} = \langle Wu, \tilde{u} \rangle_{\mathbb{R}^m} \quad \text{for } u,\, \tilde{u} \in \mathbb{R}^m, \tag{1.34}$$

where $W \in \mathbb{R}^{m \times m}$ is a symmetric, positive-definite matrix. Furthermore, let $\|u\|_W = \sqrt{\langle u, u \rangle_W}$ for $u \in \mathbb{R}^m$ be the associated induced norm. For the choice $W = I$, the inner product (1.34) coincides the Euclidean inner product.

Example 2.7. Let us motivate the weighted inner product by an example. Suppose that $\Omega = (0, 1) \subset \mathbb{R}$ holds. We consider the space $L^2(\Omega)$ of square integrable functions on Ω:

$$L^2(\Omega) = \left\{ \varphi : \Omega \to \mathbb{R} \,\middle|\, \int_\Omega |\varphi|^2 \, \mathrm{d}x < \infty \right\}.$$

Recall that $L^2(\Omega)$ is a Hilbert space endowed with the inner product

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

and the induced norm $\|\varphi\|_{L^2(\Omega)} = \sqrt{\langle \varphi, \varphi \rangle_{L^2(\Omega)}}$ for $\varphi \in L^2(\Omega)$. For the step size $h = 1/(m-1)$ let us introduce a spatial grid in Ω by

$$x_i = (i-1)h \quad \text{for } i = 1, \dots, m.$$

For any $\varphi,\, \tilde{\varphi} \in L^2(\Omega)$ we introduce a discrete inner product by trapezoidal approximation:

$$\langle \varphi, \tilde{\varphi} \rangle_{L^2_h(\Omega)} = h\left(\frac{\varphi_1^h \tilde{\varphi}_1^h}{2} + \sum_{i=2}^{m-1} \left(\varphi_i^h \tilde{\varphi}_i^h \right) + \frac{\varphi_m^h \tilde{\varphi}_m^h}{2} \right), \tag{1.35}$$

where

$$\varphi_i^h = \begin{cases} \dfrac{2}{h} \displaystyle\int_0^{h/2} \varphi(x) \, \mathrm{d}x & \text{for } i = 1, \\ \dfrac{1}{h} \displaystyle\int_{x_i - h/2}^{x_i + h/2} \varphi(x) \, \mathrm{d}x & \text{for } i = 2, \dots, m-1, \\ \dfrac{2}{h} \displaystyle\int_{1-h/2}^{1} \varphi(x) \, \mathrm{d}x & \text{for } i = m \end{cases}$$

and the $\tilde{\varphi}_i^h$'s are defined analogously. Setting $W = \operatorname{diag}(h/2, h, \dots, h, h/2) \in \mathbb{R}^{m\times m}$, $\varphi^h = (\varphi_1^h, \dots, \varphi_m^h)^T \in \mathbb{R}^m$ and $\tilde{\varphi}^h = (\tilde{\varphi}_1^h, \dots, \tilde{\varphi}_m^h)^T \in \mathbb{R}^m$ we find

$$\langle \varphi, \tilde{\varphi} \rangle_{L_h^2(\Omega)} = \langle \varphi^h, \tilde{\varphi}^h \rangle_W = (\varphi^h)^T W \tilde{\varphi}^h.$$

Thus, the discrete L^2-inner product can be written as a weighted inner product of the form (1.34). ◊

Now we replace $(\mathbf{P}^1)$ by

$$(\mathbf{P}_W^1) \qquad \max_{u\in\mathbb{R}^m} \sum_{j=1}^n \big|\langle y_j, u\rangle_W\big|^2 \quad \text{s.t.} \quad \|u\|_W = 1.$$

Analogously to Section 1.1 we treat $(\mathbf{P}_W^1)$ as an equality constrained optimization problem. The Lagrangian $\mathcal{L} : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ for $(\mathbf{P}_W^1)$ is given by

$$\mathcal{L}(u,\lambda) = \sum_{j=1}^n \big|\langle y_j, u\rangle_W\big|^2 + \lambda\big(1 - \|u\|_W^2\big) \quad \text{for } (u,\lambda) \in \mathbb{R}^m \times \mathbb{R}.$$

Suppose that $u \in \mathbb{R}^m$ is a solution to $(\mathbf{P}_W^1)$. Then, a first-order necessary optimality condition is given by

$$\nabla \mathcal{L}(u,\lambda) \overset{!}{=} 0 \quad \text{in } \mathbb{R}^m \times \mathbb{R}.$$

We compute the gradient of $\mathcal{L}$ with respect to u: Since W is symmetric, we derive

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial u_i}(u,\lambda) &= \frac{\partial}{\partial u_i}\left(\sum_{j=1}^n \left|\sum_{k=1}^m\sum_{\nu=1}^m Y_{j\nu}^T W_{\nu k} u_k\right|^2 + \lambda\left(1 - \sum_{k=1}^m\sum_{\nu=1}^m u_\nu W_{\nu k} u_k\right)\right) \\
&= 2\sum_{j=1}^n \left(\sum_{k=1}^m\sum_{\nu=1}^m Y_{j\nu}^T W_{\nu k} u_k\right)\left(\sum_{\mu=1}^m Y_{j\mu}^T W_{\mu i}\right) \\
&\quad - \lambda\left(\sum_{\nu=1}^m u_\nu W_{\nu i} + \sum_{k=1}^m W_{ik} u_k\right) \\
&= 2\sum_{k=1}^m\sum_{\nu=1}^m\sum_{\mu=1}^m W_{i\mu}\sum_{j=1}^n Y_{\mu j} Y_{j\nu}^T W_{\nu k} u_k - 2\lambda\left(\sum_{k=1}^m W_{ik} u_k\right) \\
&= 2\Big(WYY^TWu - \lambda W u\Big)_i.
\end{aligned}$$

Thus,

$$\nabla_u \mathcal{L}(u,\lambda) = 2\big(WYY^TWu - \lambda W u\big) \overset{!}{=} 0 \quad \text{in } \mathbb{R}^m. \tag{1.36}$$

Equation (1.36) yields the generalized eigenvalue problem

$$(WY)(WY)^T u = \lambda W u. \tag{1.37}$$

Since W is symmetric and positive definite, W possesses an eigenvalue decomposition of the form $W = QDQ^T$, where $D = \operatorname{diag}(\eta_1, \dots, \eta_m)$ contains the eigenvalues $\eta_1 \geq \dots \geq \eta_m > 0$ of W and $Q \in \mathbb{R}^{m\times m}$ is an orthogonal matrix. We define

$$W^\alpha = Q \operatorname{diag}(\eta_1^\alpha, \dots, \eta_m^\alpha) Q^T \quad \text{for } \alpha \in \mathbb{R}.$$

Note that $(W^\alpha)^{-1} = W^{-\alpha}$ and $W^{\alpha+\beta} = W^\alpha W^\beta$ for α, $\beta \in \mathbb{R}$; see Exercise 1.3). Moreover, we have

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

and $\|u\|_W = \|W^{1/2}u\|_{\mathbb{R}^m}$ for $u \in \mathbb{R}^m$.

Setting $\bar{u} = W^{1/2}u \in \mathbb{R}^m$ and $\bar{Y} = W^{1/2}Y \in \mathbb{R}^{m\times n}$ and multiplying (1.37) by $W^{-1/2}$ from the left we deduce the symmetric, $m \times m$ eigenvalue problem

$$\bar{Y}\bar{Y}^T\bar{u} = \lambda\bar{u} \quad \text{in } \mathbb{R}^m. \tag{1.38a}$$

From $\frac{\partial \mathcal{L}}{\partial \lambda}(u,\lambda) \stackrel{!}{=} 0$ in $\mathbb{R}$ we infer the constraint $\|u\|_W = 1$ that can be expressed as

$$\|\bar{u}\|_{\mathbb{R}^m} = 1. \tag{1.38b}$$

Thus, the first-order optimality conditions (1.38) for $(\mathbf{P}_W^1)$ are — as for $(\mathbf{P}^1)$ (compare (1.18)) — an $m \times m$ eigenvalue problem, but the matrix Y as well as the vector u have to be weighted by the matrix $W^{1/2}$.

It can be shown (see Exersice 1.4.1)) that

$$u_1 = W^{-1/2}\bar{u}_1$$

solves $(\mathbf{P}_W^1)$, where $\bar{u}_1$ is an eigenvector of $\bar{Y}\bar{Y}^T$ corresponding to the largest eigenvalue λ_1 with $\|\bar{u}_1\|_{\mathbb{R}^m} = 1$. Due to SVD the vector u_1 can be also determined by solving the symmetric $n \times n$ eigenvalue problem

$$\bar{Y}^T\bar{Y}\bar{v}_1 = \lambda_1\bar{v}_1$$

where $\bar{Y}^T\bar{Y} = Y^TWY$, and setting

$$u_1 = W^{-1/2}\bar{u}_1 = \frac{1}{\sqrt{\lambda_1}}\, W^{-1/2}\bar{Y}\bar{v}_1 = \frac{1}{\sqrt{\lambda_1}} Y\bar{v}_1. \tag{1.39}$$

As in Section 1.1 we can continue by looking at a second vector $u \in \mathbb{R}^m$ with $\langle u, u_1\rangle_W = 0$ that maximizes $\sum_{j=1}^n |\langle y_j, u\rangle_W|^2$. Let us generalize Theorem 2.1 as follows.

THEOREM 2.8. *Let $Y \in \mathbb{R}^{m\times n}$ be a given matrix with rank $d \le \min\{m,n\}$, W a symmetric, positive definite matrix, $\bar{Y} = W^{1/2}Y$ and $\ell \in \{1,\dots,d\}$. Further, let $\bar{Y} = \bar{U}\Sigma\bar{V}^T$ be the singular value decomposition of $\bar{Y}$, where $\bar{U} = [\bar{u}_1,\dots,\bar{u}_m] \in \mathbb{R}^{m\times m}$, $\bar{V} = [\bar{v}_1,\dots,\bar{v}_n] \in \mathbb{R}^{n\times n}$ are orthogonal matrices and the matrix Σ has the form*

$$\bar{U}^T\bar{Y}\bar{V} = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} = \Sigma \in \mathbb{R}^{m\times n}.$$

Then the solution to

$$\max_{\tilde{u}_1,\dots,\tilde{u}_\ell \in \mathbb{R}^m} \sum_{i=1}^{\ell}\sum_{j=1}^{n} \left|\langle y_j, \tilde{u}_i\rangle_W\right|^2 \quad s.t. \quad \langle \tilde{u}_i, \tilde{u}_j\rangle_W = \delta_{ij} \text{ for } 1 \le i,j \le \ell \tag{$\mathbf{P}_W^\ell$}$$

is given by the vectors $u_i = W^{-1/2}\bar{u}_i$, $i = 1,\dots,\ell$. Moreover,

$$\operatorname{argmax}(\mathbf{P}_W^\ell) = \sum_{i=1}^{\ell}\sigma_i^2 = \sum_{i=1}^{\ell}\lambda_i. \tag{1.40}$$

PROOF. Using similar arguments as in the proof of Theorem 2.1 one can prove that $\{u_i\}_{i=1}^\ell$ solves $(\mathbf{P}_W^\ell)$; see Exersice 1.4). □

REMARK 2.9. Due to SVD and $\bar{Y}^T\bar{Y} = Y^TWY$ the POD basis $\{u_i\}_{i=1}^\ell$ of rank ℓ can be determined by the method of snapshots as follows: Solve the symmetric $n \times n$ eigenvalue problem

$$Y^TWY\bar{v}_i = \lambda_i\bar{v}_i \quad \text{for } i = 1,\dots,\ell,$$

and set

$$u_i = W^{-1/2}\bar{u}_i = \frac{1}{\sqrt{\lambda_i}}\, W^{-1/2}\big(\bar{Y}\bar{v}_i\big) = \frac{1}{\sqrt{\lambda_i}}\, W^{-1/2}W^{1/2}Y\bar{v}_i = \frac{1}{\sqrt{\lambda_i}}\, Y\bar{v}_i$$

for $i = 1,\dots,\ell$. Notice that

$$\langle u_i, u_j\rangle_W = u_i^TWu_j = \frac{\delta_{ij}\lambda_j}{\sqrt{\lambda_i\lambda_j}} \quad \text{for } 1 \le i,j \le \ell.$$

For $m \gg n$ the method of snapshots turns out to be faster than computing the POD basis via (1.38). Notice that the matrix $W^{1/2}$ is also not required for the method of snapshots. ◊

2.2. POD for nonlinear dynamical systems. For $T > 0$ we consider the semi-linear initial value problem

$$\dot{y}(t) = Ay(t) + f(t, y(t)) \quad \text{for } t \in (0, T], \tag{1.41a}$$

$$y(0) = y_0, \tag{1.41b}$$

where $y_0 \in \mathbb{R}^m$ is a chosen initial condition, $A \in \mathbb{R}^{m\times m}$ is a given matrix, $f : [0, T] \times \mathbb{R}^m \to \mathbb{R}^m$ is continuous in both arguments and locally Lipschitz-continuous with respect to the second argument. It is well known that (1.41) has a unique (classical) solution $y \in C^1(0, T; \mathbb{R}^m) \cap C([0, T]; \mathbb{R}^m)$ given by the implicit integral representation

$$y(t) = e^{tA}y_0 + \int_0^t e^{(t-s)A} f(s, y(s))\,\mathrm{d}s$$

with $e^{tA} = \sum_{i=0}^{\infty} t^n A^n/(n!)$. Let $0 \le t_1 < t_2 < \ldots < t_n \le T$ be a given time grid in the interval $[0, T]$. For simplicity of the presentation, the time grid is assumed to be equidistant with step-size $\Delta t = T/(n-1)$, i.e., $t_j = (j-1)\Delta t$. We suppose that we know the solution to (1.41) at the given time instances t_j, $j \in \{1, \ldots, n\}$. Our goal is to determine a POD basis of rank $\ell \le n$ that desribes the ensemble

$$y_j = y(t_j) = e^{t_j A}y_0 + \int_0^{t_j} e^{(t_j-s)A} f(s, y(s))\,\mathrm{d}s, \quad j = 1, \ldots, n,$$

as well as possible with respect to the weighted inner product:

$$(\hat{\mathbf{P}}_W^{n,\ell}) \qquad \min_{\tilde{u}_1,\ldots,\tilde{u}_\ell \in \mathbb{R}^m} \sum_{j=1}^{n} \alpha_j \Big\| y_j - \sum_{i=1}^{\ell} \langle y_j, \tilde{u}_i\rangle_W \tilde{u}_i \Big\|_W^2 \quad \text{s.t.} \quad \langle \tilde{u}_i, \tilde{u}_j\rangle_W \text{ for } 1 \le i, j \le \ell,$$

where the α_j's denote non-negative weights which will be specified later on. Note that for $\alpha_j = 1$ for $j = 1, \ldots, n$ and $W = I$ problem $(\hat{\mathbf{P}}_W^{n,\ell})$ coincides with (1.31).

Example 2.10. Let us consider the following one-dimensional heat equation:

$$\theta_t(t, x) = \theta_{xx}(t, x) \qquad \text{for all } (t, x) \in Q = (0, T) \times \Omega, \tag{1.42a}$$

$$\theta_x(t, 0) = \theta_x(t, 1) = 0 \qquad \text{for all } t \in (0, T), \tag{1.42b}$$

$$\theta(0, x) = \theta_0(x) \qquad \text{for all } x \in \Omega = (0, 1) \subseteq \mathbb{R}, \tag{1.42c}$$

where $\theta_0 \in C(\overline{\Omega})$ is a given initial condition. To solve (1.42) numerically we apply a classical finite difference approximation for the spatial variable x. In Example 2.7 we have introduced the spatial grid $\{x_i\}_{i=1}^m$ in the interval $[0, 1]$. Let us denote by $y_i : [0, T] \to \mathbb{R}$ the numerical approximation for $\theta(\cdot\,, x_i)$ for $i = 1, \ldots, m$. The second partial derivative θ_{xx} in (1.42a) and the boundary conditions (1.42b) are discretized by centered difference quotients of second-order so that we obtain the following ordinary differential equations for the time-dependent functions y_i:

$$\begin{cases} \dot{y}_1(t) &= \dfrac{-2y_1(t) + 2y_2(t)}{h^2}, \\ \dot{y}_i(t) &= \dfrac{y_{i-1}(t) - 2y_i(t) + y_{i+1}(t)}{h^2}, \quad i = 2, \ldots, m-1, \\ \dot{y}_m(t) &= \dfrac{-2y_m(t) + 2y_{m-1}(t)}{h^2} \end{cases} \tag{1.43a}$$

for $t \in (0, T]$. From (1.42c) we infer the initial condion

$$y_i(0) = \theta_0(x_i), \quad i = 1, \ldots, m. \tag{1.43b}$$

Introducing the matrix

$$A = \frac{1}{h^2}\begin{pmatrix} -2 & 2 & & & 0 \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ 0 & & & 2 & -2 \end{pmatrix} \in \mathbb{R}^{m\times m}$$

and the vectors

$$y(t) = \begin{pmatrix} y_1(t) \\ \vdots \\ y_m(t) \end{pmatrix} \text{ for } t \in [0,T], \quad y_0 = \begin{pmatrix} \theta_0(x_1) \\ \vdots \\ \theta_0(x_m) \end{pmatrix} \in \mathbb{R}^m$$

we can express (1.43) in the form

$$\begin{aligned} \dot{y}(t) &= Ay(t) \quad \text{for } t \in (0,T], \\ y(0) &= y_0 \end{aligned} \tag{1.44}$$

Setting $f \equiv 0$ the linear initial-value problem coincides with (1.41). Note that now the vector $y(t)$, $t \in [0,T]$, represents a function in Ω evaluated at m grid points. Therefore, we should supply $\mathbb{R}^m$ by a weighted inner product representing a discretized inner product in an appropriate function space. Here we choose the inner product introduced in (1.35); see Example 2.7. Next we choose a time grid $\{t_j\}_{j=1}^n$ in the interval $[0,T]$ and define $y_j = y(t_j)$ for $j = 1,\dots,n$. If we are interested in finding a POD basis of rank $\ell \le n$ that desribes the ensemble $\{y_j\}_{j=1}^n$ as well as possible, we end up with $(\hat{\mathbf{P}}_W^{n,\ell})$. ◊

To solve $(\hat{\mathbf{P}}_W^{n,\ell})$ we apply the techniques used in Sections 1.1 and 1.2, i.e., we use the Lagrangian framework. Thus, we introduce the Lagrange functional

$$\mathcal{L} : \underbrace{\mathbb{R}^m \times \ldots \times \mathbb{R}^m}_{\ell-\text{times}} \times \mathbb{R}^{\ell\times\ell} \to \mathbb{R}$$

by

$$\mathcal{L}(u_1,\dots,u_\ell,\Lambda) = \sum_{j=1}^n \alpha_j \Big\| y_j - \sum_{i=1}^\ell \langle y_j, u_i\rangle_W u_i \Big\|_W^2 + \sum_{i=1}^\ell \sum_{j=1}^\ell \Lambda_{ij}\big(1 - \langle u_i, u_j\rangle_W\big)$$

for $u_1,\dots,u_\ell \in \mathbb{R}^m$ and $\Lambda \in \mathbb{R}^{\ell\times\ell}$ with elements Λ_{ij}, $1 \le i,j \le \ell$. It turns out that the solution to $(\hat{\mathbf{P}}_W^{n,\ell})$ is given by the first-order necessary optimality condions

$$\nabla_{u_i}\mathcal{L}(u_1,\dots,u_\ell,\Lambda) \stackrel{!}{=} 0 \quad \text{in } \mathbb{R}^m,\ 1 \le i \le \ell, \tag{1.45a}$$

and

$$\langle u_i, u_j\rangle_W \stackrel{!}{=} \delta_{ij}, \quad 1 \le i,j \le \ell. \tag{1.45b}$$

From (1.45a) we derive

$$YDY^TWu_i = \lambda_i u_i \quad \text{for } i = 1,\dots,\ell, \tag{1.46}$$

where $D = \operatorname{diag}(\alpha_1,\dots,\alpha_n) \in \mathbb{R}^{n\times n}$. Inserting $u_i = W^{-1/2}\bar{u}_i$ in (1.46) and multiplying (1.46) by $W^{1/2}$ from the left yield

$$W^{1/2}YDY^TW^{1/2}\bar{u}_i = \lambda_i\bar{u}_i. \tag{1.47a}$$

From (1.45b) we find

$$\langle \bar{u}_i, \bar{u}_j\rangle_{\mathbb{R}^m} = \bar{u}_i^T\bar{u}_j = u_i^TWu_j = \langle u_i, u_j\rangle_W = \delta_{ij}, \quad 1 \le i,j \le \ell. \tag{1.47b}$$

Setting $\bar{Y} = W^{1/2}YD^{1/2} \in \mathbb{R}^{m\times n}$ and using $W^T = W$ as well as $D^T = D$ we infer from (1.47) that the solution $\{u_i\}_{i=1}^\ell$ to $(\hat{\mathbf{P}}_W^{n,\ell})$ is given by the symmetric $m \times m$ eigenvalue problem

$$\bar{Y}\bar{Y}^T\bar{u}_i = \lambda_i\bar{u}_i,\ 1 \le i \le \ell \quad \text{and} \quad \langle \bar{u}_i, \bar{u}_j\rangle_{\mathbb{R}^m} = \delta_{ij},\ 1 \le i,j \le \ell.$$

Note that

$$\bar{Y}^T\bar{Y} = D^{1/2}Y^TWYD^{1/2} \in \mathbb{R}^{n\times n}.$$

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

$$\bar{Y}^T\bar{Y}\bar{v}_i = \lambda_i\bar{v}_i,\ 1 \le i \le \ell \quad \text{and} \quad \langle \bar{v}_i, \bar{v}_j\rangle_{\mathbb{R}^n} = \delta_{ij},\ 1 \le i,j \le \ell.$$

Then we set (by SVD)

$$u_i = W^{-1/2}\bar{u}_i = \frac{1}{\sqrt{\lambda_i}}\,W^{-1/2}\bar{Y}\bar{v}_i = \frac{1}{\sqrt{\lambda_i}}\,YD^{1/2}\bar{v}_i, \quad 1 \le i \le \ell;$$

compare (1.39).

Note that

$$\langle u_i, u_j\rangle_W = u_i^TWu_j = \frac{1}{\sqrt{\lambda_i\lambda_j}}\,\bar{v}_i^T\underbrace{D^{1/2}Y^TWYD^{1/2}}_{=\bar{Y}^T\bar{Y}}\bar{v}_j = \frac{\lambda_i}{\sqrt{\lambda_i\lambda_j}}\,\bar{v}_i^T\bar{v}_j = \frac{\lambda_i\delta_{ij}}{\sqrt{\lambda_i\lambda_j}}$$

for $1 \le i,j \le \ell$, i.e., the POD basis vectors $u_1, \dots, u_\ell$ are orthonormal in $\mathbb{R}^m$ with respect to the inner product $\langle\cdot\,,\cdot\rangle_W$.

2.3. Continuous POD for nonlinear dynamical systems. Of course, the snapshot ensemble $\{y_j\}_{j=1}^n$ for $(\hat{\mathbf{P}}_W^{n,\ell})$ and therefore the snapshot set span $\{y_1, \dots, y_n\}$ depend on the chosen time instances $\{t_j\}_{j=1}^n$. Consequently, the POD basis vectors $\{u_i\}_{i=1}^\ell$ and the corresponding eigenvalues $\{\lambda_i\}_{i=1}^\ell$ depend also on the time instances, i.e.,

$$u_i = u_i^n \quad \text{and} \quad \lambda_i = \lambda_i^n, \quad 1 \le i \le \ell.$$

Moreover, we have not discussed so far what is the motivation to introduce the non-negative weights $\{\alpha_j\}_{j=1}^n$ in $(\hat{\mathbf{P}}_W^{n,\ell})$. For this reason we proceed by investigating the following two questions:

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

To address these two questions we will introduce a *continuous version* of POD. Let $y : [0,T] \to \mathbb{R}^m$ be the unique solution to (1.41). If we are interested to find a POD basis of rank ℓ that describes the whole trajectory $\{y(t)\,|\,t \in [0,T]\} \subset \mathbb{R}^m$ as good as possible we have to consider the following minimization problem

$$(\hat{\mathbf{P}}_W^\ell) \qquad \begin{aligned} &\min_{\tilde{u}_1,\dots,\tilde{u}_\ell \in \mathbb{R}^m} \int_0^T \Big\| y(t) - \sum_{i=1}^\ell \langle y(t), \tilde{u}_i\rangle_W\,\tilde{u}_i \Big\|_W^2\,\mathrm{d}t \\ &\text{s.t. } \langle \tilde{u}_i, \tilde{u}_j\rangle_W = \delta_{ij},\ 1 \le i,j \le \ell, \end{aligned}$$

To solve $(\hat{\mathbf{P}}_W^\ell)$ we use similar arguments as in Sections 1.1 and 1.2. For $\ell = 1$ we obtain instead of $(\hat{\mathbf{P}}_W^\ell)$ the minimization problem

$$\min_{\tilde{u}\in\mathbb{R}^m} \int_0^T \Big\| y(t) - \langle y(t), \tilde{u}\rangle_W\,\tilde{u} \Big\|_W^2\,\mathrm{d}t \quad \text{s.t.} \quad \|\tilde{u}\|_W^2 = 1, \tag{1.48}$$

Suppose that $\{\tilde{u}_i\}_{i=2}^m$ are chosen in such a way that $\{\tilde{u}, \tilde{u}_2, \dots, \tilde{u}_m\}$ is an orthonormal basis in $\mathbb{R}^m$ with respect to the inner product $\langle\cdot\,,\cdot\rangle_W$. Then we have

$$y(t) = \langle y(t), \tilde{u}\rangle_W\,\tilde{u} + \sum_{i=2}^m \langle y(t), \tilde{u}_i\rangle_W\,\tilde{u}_i \quad \text{for all } t \in [0,T].$$

Thus,

$$\int_0^T \Big\| y(t) - \langle y(t), \tilde{u} \rangle_W \tilde{u} \Big\|_W^2 \,\mathrm{d}t = \int_0^T \Big\| \sum_{i=2}^m \langle y(t), \tilde{u} \rangle_W \tilde{u}_i \Big\|_W^2 \,\mathrm{d}t$$
$$= \sum_{i=2}^m \int_0^T \big| \langle y(t), \tilde{u}_i \rangle_W \big|^2 \,\mathrm{d}t$$

we conclude that (1.48) is equivalent with the following maximization problem

$$\max_{\tilde{u} \in \mathbb{R}^m} \int_0^T \big| \langle y(t), \tilde{u} \rangle_W \big|^2 \,\mathrm{d}t \quad \text{s.t.} \quad \|\tilde{u}\|_W^2 = 1. \tag{1.49}$$

The Lagrange functional $\mathcal{L} : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ associated with (1.49) is given by

$$\mathcal{L}(u, \lambda) = \int_0^T \big| \langle y(t), u \rangle_W \big|^2 \,\mathrm{d}t + \lambda\big(1 - \|u\|_W^2\big) \quad \text{for } (u, \lambda) \in \mathbb{R}^m \times \mathbb{R}.$$

First-order necessary optimality conditions are given by

$$\nabla \mathcal{L}(u, \lambda) \overset{!}{=} 0 \quad \text{in } \mathbb{R}^m \times \mathbb{R}.$$

Therefore, we compute the partial derivative of $\mathcal{L}$ with respect to the ith component u_i of the vector u:

$$\frac{\partial \mathcal{L}}{\partial u_i}(u, \lambda) = \frac{\partial}{\partial u_i} \left(\int_0^T \Big| \sum_{k=1}^m \sum_{\nu=1}^m y_k(t) W_{k\nu} u_\nu \Big|^2 \,\mathrm{d}t + \lambda\Big(1 - \sum_{k=1}^m \sum_{\nu=1}^m u_k W_{k\nu} u_\nu\Big)\right)$$
$$= 2 \int_0^T \Big(\sum_{k=1}^m \sum_{\nu=1}^m y_k(t) W_{k\nu} u_\nu \Big) \sum_{\mu=1}^m y_\mu(t) W_{\mu i} \,\mathrm{d}t - 2\lambda \sum_{k=1}^m W_{ik} u_k$$
$$= 2\bigg(\int_0^T \langle y(t), u \rangle_W W y(t) \,\mathrm{d}t - \lambda W u \bigg)_i$$

for $i \in \{1, \ldots, m\}$. Thus,

$$\nabla_u \mathcal{L}(u, \lambda) = 2\bigg(\int_0^T \langle y(t), u \rangle_W W y(t) \,\mathrm{d}t - \lambda W u \bigg) \overset{!}{=} 0 \quad \text{in } \mathbb{R}^m,$$

which gives

$$\int_0^T \langle y(t), u \rangle_W W y(t) \,\mathrm{d}t = \lambda W u \quad \text{in } \mathbb{R}^m. \tag{1.50}$$

Multiplying (1.50) by W^{-1} from the left yields

$$\int_0^T \langle y(t), u \rangle_W y(t) \,\mathrm{d}t = \lambda u \quad \text{in } \mathbb{R}^m. \tag{1.51}$$

We define the operator $\mathcal{R} : \mathbb{R}^m \to \mathbb{R}^m$ as

$$\mathcal{R} u = \int_0^T \langle y(t), u \rangle_W \, y(t) \,\mathrm{d}t \quad \text{for } u \in \mathbb{R}^m. \tag{1.52}$$

LEMMA 2.11. *The operator $\mathcal{R}$ is linear and bounded (i.e., continuous). Moreover,*

1) *$\mathcal{R}$ is non-negative:*

$$\langle \mathcal{R} u, u \rangle_W \geq 0 \quad \textit{for all } u \in \mathbb{R}^m.$$

2) *$\mathcal{R}$ is self-adjoint (or symmetric):*

$$\langle \mathcal{R} u, \tilde{u} \rangle_W = \langle u, \mathcal{R} \tilde{u} \rangle_W \quad \textit{for all } u, \tilde{u} \in \mathbb{R}^m.$$

PROOF. For arbitrary u, $\tilde{u} \in \mathbb{R}^m$ and α, $\tilde{\alpha} \in \mathbb{R}$ we have

$$\begin{aligned}\mathcal{R}\big(\alpha u + \tilde{\alpha}\tilde{u}\big) &= \int_0^T \langle y(t), \alpha u + \tilde{\alpha}\tilde{u}\rangle_W \, y(t)\,\mathrm{d}t \\ &= \int_0^T \left(\alpha \langle y(t), u\rangle_W + \tilde{\alpha}\langle y(t), \tilde{u}\rangle_W\right) y(t)\,\mathrm{d}t \\ &= \alpha \int_0^T \langle y(t), u\rangle_W \, y(t)\,\mathrm{d}t + \tilde{\alpha}\int_0^T \langle y(t), \tilde{u}\rangle_W \, y(t)\,\mathrm{d}t = \alpha\mathcal{R}u + \tilde{\alpha}\mathcal{R}\tilde{u},\end{aligned}$$

so that $\mathcal{R}$ is linear. From the Cauchy-Schwarz inequality we derive

$$\begin{aligned}\|\mathcal{R}u\|_W &\le \int_0^T \big\|\langle y(t), u\rangle_W \, y(t)\big\|_W \,\mathrm{d}t = \int_0^T \big|\langle y(t), u\rangle_W\big| \, \|y(t)\|_W \,\mathrm{d}t \\ &\le \int_0^T \|y(t)\|_W^2 \|u\|_W \,\mathrm{d}t = \left(\int_0^T \|y(t)\|_W^2 \,\mathrm{d}t\right)\|u\|_W = \|y\|^2_{L^2(0,T;\mathbb{R}^m)}\|u\|_W\end{aligned}$$

for an arbitrary $u \in \mathbb{R}^m$. Since $y \in C([0,T];\mathbb{R}^m) \subset L^2(0,T;\mathbb{R}^m)$ holds, the norm $\|y\|_{L^2(0,T;\mathbb{R}^m)}$ is bounded. Therefore, $\mathcal{R}$ is bounded. Since

$$\begin{aligned}\langle \mathcal{R}u, u\rangle_W &= \left(\int_0^T \langle y(t), u\rangle_W \, y(t)\,\mathrm{d}t\right)^T Wu = \int_0^T \langle y(t), u\rangle_W \, y(t)^T Wu\,\mathrm{d}t \\ &= \int_0^T \big|\langle y(t), u\rangle_W\big|^2 \,\mathrm{d}t \ge 0\end{aligned}$$

for all $u \in \mathbb{R}^m$ holds, $\mathcal{R}$ is non-negative. Finally, we infer from

$$\begin{aligned}\langle \mathcal{R}u, \tilde{u}\rangle_W &= \int_0^T \langle y(t), u\rangle_W \langle y(t), \tilde{u}\rangle_W \,\mathrm{d}t = \left\langle \int_0^T \langle y(t), \tilde{u}\rangle_W y(t)\,\mathrm{d}t, u\right\rangle_W \\ &= \langle \mathcal{R}\tilde{u}, u\rangle_W = \langle u, \mathcal{R}\tilde{u}\rangle_W\end{aligned}$$

for all u, $\tilde{u} \in \mathbb{R}^m$ that $\mathcal{R}$ is self-adjoint. □

Utilizing the operator $\mathcal{R}$ we can write (1.51) as the eigenvalue problem

$$\mathcal{R}u = \lambda u \quad \text{in } \mathbb{R}^m.$$

It follows from Lemma 2.11 that $\mathcal{R}$ possesses eigenvectors $\{u_i\}_{i=1}^m$ and associated real eigenvalues $\{\lambda_i\}_{i=1}^m$ such that

$$\mathcal{R}u_i = \lambda_i u_i \text{ for } 1 \le i \le m \quad \text{and} \quad \lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_m \ge 0. \tag{1.53}$$

Note that

$$\int_0^T \big|\langle y(t), u_i\rangle_W\big|^2 \,\mathrm{d}t = \int_0^T \big\langle \langle y(t), u_i\rangle_W y(t), u_i\big\rangle_W \,\mathrm{d}t = \langle \mathcal{R}u_i, u_i\rangle_W = \lambda_i \, \|u_i\|_W^2 = \lambda_i$$

for $i \in \{1, \ldots, m\}$ so that u_1 solves (1.48).

Proceeding as in Sections 1.1 and 1.2 we obtain the following result.

THEOREM 2.12. *Let $y \in C([0,T];\mathbb{R}^m)$ be the unique solution to (1.41). Then the POD basis of rank ℓ solving the minimization problem $(\hat{\mathbf{P}}_W^\ell)$ is given by the eigenvectors $\{u_i\}_{i=1}^\ell$ of $\mathcal{R}$ corresponding to the ℓ largest eigenvalues $\lambda_1 \ge \ldots \ge \lambda_\ell$.*

REMARK 2.13 (Methods of snapshots). Let us introduce the linear and bounded operator $\mathcal{Y} : L^2(0,T) \to \mathbb{R}^m$ by

$$\mathcal{Y}v = \int_0^T v(t)y(t)\,\mathrm{d}t \quad \text{for } v \in L^2(0,T).$$

The adjoint $\mathcal{Y}^\star : \mathbb{R}^m \to L^2(0,T)$ satisfying

$$\langle \mathcal{Y}^\star u, v\rangle_{L^2(0,T)} = \langle u, \mathcal{Y}v\rangle_W \quad \text{for all } (u,v) \in \mathbb{R}^m \times L^2(0,T)$$

is given as

$$(\mathcal{Y}^\star u)(t) = \langle u, y(t)\rangle_W \quad \text{for } u \in \mathbb{R}^m \text{ and almost all } t \in [0,T].$$

Then we have

$$\mathcal{Y}\mathcal{Y}^\star u = \int_0^T \langle u, y(t)\rangle_W \, y(t)\,\mathrm{d}t = \int_0^T \langle y(t), u\rangle_W \, y(t)\,\mathrm{d}t = \mathcal{R}u$$

for all $u \in \mathbb{R}^m$, i.e., $\mathcal{R} = \mathcal{Y}\mathcal{Y}^*$ holds. Furthermore,

$$(\mathcal{Y}^\star\mathcal{Y}v)(t) = \left\langle \int_0^T v(s)y(s)\,\mathrm{d}t, y(t)\right\rangle_W = \int_0^T \langle y(s), y(t)\rangle_W \, v(s)\,\mathrm{d}s =: (\mathcal{K}v)(t)$$

for all $v \in L^2(0,T)$ and almost all $t \in [0,T]$. Thus, $\mathcal{K} = \mathcal{Y}^\star\mathcal{Y}$. It can be shown that the operator $\mathcal{K}$ is linear, bounded, non-negartive and self-adjoint. Moreover, $\mathcal{K}$ is compact. Therefore, the POD basis can also be computed as follows: Solve

$$\mathcal{K}v_i = \lambda_i v_i \text{ for } 1 \le i \le \ell, \quad \lambda_1 \ge \ldots \ge \lambda_\ell > 0, \quad \int_0^T v_i(t)v_j(t)\,\mathrm{d}t = \delta_{ij} \tag{1.54}$$

and set

$$u_i = \frac{1}{\sqrt{\lambda_i}}\,\mathcal{Y}v_i = \frac{1}{\sqrt{\lambda_i}}\int_0^T v_i(t)y(t)\,\mathrm{d}t \quad \text{for } i = 1,\ldots,\ell.$$

Note that (1.54) is a symmetric eigenvalue problem in the infinite-dimensional function space $L^2(0,T)$. For the functional analytic theory we refer, e.g., to [**36**]. ◊

Let us turn back to the optimality conditions (1.46). For any $u \in \mathbb{R}^m$ and $i \in \{1,\ldots,m\}$ we derive

$$\begin{aligned}\left(YDY^TWu\right)_i &= \sum_{\nu=1}^m\sum_{j=1}^m\sum_{k=1}^m \alpha_j Y_{ij}Y_{kj}W_{k\nu}u_\nu = \sum_{j=1}^n \alpha_j Y_{ij}\,\langle y_j, u\rangle_W \\ &= \sum_{j=1}^n \alpha_j\,\langle y_j, u\rangle_W (y_j)_i,\end{aligned}$$

where $(y_j)_i$ stands for the ith component of the vector $y_j \in \mathbb{R}^m$. Thus,

$$YDY^TWu = \sum_{j=1}^n \alpha_j\,\langle y_j, u\rangle_W y_j =: \mathcal{R}^n u.$$

Note that the operator $\mathcal{R}^n : \mathbb{R}^m \to \mathbb{R}^m$ is linear and bounded. Moreover,

$$\langle \mathcal{R}^n u, u\rangle_W = \left\langle \sum_{j=1}^n \alpha_j\,\langle y_j, u\rangle_W y_j, u\right\rangle_W = \sum_{j=1}^n \alpha_j\left|\langle y_j, u\rangle_W\right|^2 \ge 0$$

holds for all $u \in \mathbb{R}^m$ so that $\mathcal{R}^n$ is non-negative. Further,

$$\begin{aligned}\langle \mathcal{R}^n u, \tilde{u}\rangle_W &= \left\langle \sum_{j=1}^n \alpha_j\,\langle y_j, u\rangle_W y_j, \tilde{u}\right\rangle_W = \sum_{j=1}^n \alpha_j\,\langle y_j, u\rangle_W\langle y_j, \tilde{u}\rangle_W \\ &= \left\langle \sum_{j=1}^n \alpha_j\,\langle y_j, \tilde{u}\rangle_W y_j, u\right\rangle_W = \langle \mathcal{R}^n\tilde{u}, u\rangle_W = \langle u, \mathcal{R}^n\tilde{u}\rangle_W\end{aligned}$$

for all $u, \tilde{u} \in \mathbb{R}^m$, i.e., $\mathcal{R}^n$ is self-adjoint. Therefore, $\mathcal{R}^n$ has the same properties as the operator $\mathcal{R}$. Summarizing, we have

$$\mathcal{R}^n u_i^n = \lambda_i^n u_i^n, \qquad \lambda_1^n \ge \ldots \lambda_\ell^n \ge \ldots \lambda_{d(n)}^n > \lambda_{d(n)+1}^n = \ldots = \lambda_m^n = 0, \tag{1.55a}$$

$$\mathcal{R}u_i = \lambda_i u_i, \qquad \lambda_1 \ge \ldots \lambda_\ell \ge \ldots \lambda_d > \lambda_{d+1} = \ldots = \lambda_m = 0. \tag{1.55b}$$

Let us note that

$$\int_0^T \|y(t)\|_W^2 \,\mathrm{d}t = \sum_{i=1}^{d} \lambda_i = \sum_{i=1}^{m} \lambda_i. \tag{1.56}$$

In fact,

$$\mathcal{R}u_i = \int_0^T \langle y(t), u_i\rangle_W \, y(t)\,\mathrm{d}t \quad \text{for every } i \in \{1,\dots,m\}.$$

Taking the inner product with u_i, using (1.55b) and summing over i we arrive at

$$\sum_{i=1}^{d} \int_0^T \left|\langle y(t), u_i\rangle_W\right|^2 \mathrm{d}t = \sum_{i=1}^{d} \langle \mathcal{R}u_i, u_i\rangle_W = \sum_{i=1}^{d} \lambda_i = \sum_{i=1}^{m} \lambda_i.$$

Expanding $y(t) \in \mathbb{R}^m$ in terms of $\{u_i\}_{i=1}^m$ we have

$$y(t) = \sum_{i=1}^{m} \langle y(t), u_i\rangle_W \, u_i$$

and hence

$$\int_0^T \|y(t)\|_W^2 \,\mathrm{d}t = \sum_{i=1}^{m} \int_0^T \left|\langle y(t), u_i\rangle_W\right|^2 \mathrm{d}t = \sum_{i=1}^{m} \lambda_i,$$

which is (1.56). Analogously, we obtain

$$\sum_{j=1}^{n} \alpha_j \, \|y(t_j)\|_W^2 = \sum_{i=1}^{d(n)} \lambda_i^n = \sum_{i=1}^{m} \lambda_i^n \quad \text{for every } n \in \mathbb{N}. \tag{1.57}$$

For convenience we do not indicate the dependence of α_j on n. Let $y \in C([0,T];\mathbb{R}^m)$ hold. To ensure

$$\sum_{j=1}^{n} \alpha_j \, \|y(t_j)\|_W^2 \to \int_0^T \|y(t)\|_W^2 \,\mathrm{d}t \quad \text{as } \Delta t \to 0 \tag{1.58}$$

we have to choose the α_j's appropriately. Here we take the trapezoidal weights

$$\alpha_1 = \frac{\Delta t}{2},\ \alpha_j = \Delta t \text{ for } 2 \le j \le n-1,\ \alpha_n = \frac{\Delta t}{2}. \tag{1.59}$$

Suppose that we have

$$\lim_{n\to\infty} \|\mathcal{R}^n - \mathcal{R}\|_{L(\mathbb{R}^m)} = \lim_{n\to\infty} \sup_{\|u\|_W = 1} \|\mathcal{R}^n u - \mathcal{R}u\|_W = 0 \tag{1.60}$$

provided $y \in C^1([0,T];\mathbb{R}^m)$ is satisfied. In (1.60) $L(\mathbb{R}^m)$ denotes the Banach space of all linear and bounded operators mapping from $\mathbb{R}^m$ into itself. Combining (1.58) with (1.56) and (1.57) we find

$$\sum_{i=1}^{m} \lambda_i^n \to \sum_{i=1}^{m} \lambda_i \quad \text{as } n \to \infty. \tag{1.61}$$

Now choose and fix

$$\ell \quad \text{such that} \quad \lambda_\ell \neq \lambda_{\ell+1}. \tag{1.62}$$

Then by spectral analysis of compact operators ([**19**, pp. 212–214]) and (1.60) it follows that

$$\lambda_i^n \to \lambda_i \quad \text{for } 1 \le i \le \ell \text{ as } n \to \infty. \tag{1.63}$$

Combining (1.61) and (1.63) there exists $\bar{n} \in \mathbb{N}$ such that

$$\sum_{i=\ell+1}^{m} \lambda_i^n \le 2 \sum_{i=\ell+1}^{m} \lambda_i \quad \text{for all } n \ge \bar{n},$$

if $\sum_{i=\ell+1}^{m} \lambda_i \neq 0$. Moreover, for ℓ as above, $\bar{n}$ can also be chosen such that

$$\sum_{i=\ell+1}^{d(n)} \left|\langle y_0, u_i^n\rangle_W\right|^2 \leq 2 \sum_{i=\ell+1}^{m} \left|\langle y_0, u_i\rangle_W\right|^2 \quad \text{for all } n \geq \bar{n}, \tag{1.64}$$

provided that $\sum_{i=\ell+1}^{m} |\langle y_0, u_i\rangle_W|^2 \neq 0$ (1.60) hold. Recall that the vector $y_0 \in \mathbb{R}^m$ stands for the initial condition in (1.41b). Then we have

$$\|y_0\|_W^2 = \sum_{i=1}^{m} \left|\langle y_0, u_i\rangle_W\right|^2. \tag{1.65}$$

If $t_1 = 0$ holds, we have $y_0 \in \operatorname{span}\{y_j\}_{j=1}^n$ for every n and

$$\|y_0\|_W^2 = \sum_{i=1}^{d(n)} \left|\langle y_0, u_i^n\rangle_W\right|^2. \tag{1.66}$$

Therefore, for $\ell < d(n)$ by (1.65) and (1.66)

$$\begin{aligned}\sum_{i=\ell+1}^{d(n)} \left|\langle y_0, u_i^n\rangle_W\right|^2 &= \sum_{i=1}^{d(n)} \left|\langle y_0, u_i^n\rangle_W\right|^2 - \sum_{i=1}^{\ell} \left|\langle y_0, u_i^n\rangle_W\right|^2 + \sum_{i=1}^{\ell} \left|\langle y_0, u_i\rangle_W\right|^2 \\ &\quad + \sum_{i=\ell+1}^{m} \left|\langle y_0, u_i\rangle_W\right|^2 - \sum_{i=1}^{m} \left|\langle y_0, u_i\rangle_W\right|^2 \\ &= \sum_{i=1}^{\ell} \left(\left|\langle y_0, u_i\rangle_W\right|^2 - \left|\langle y_0, u_i^n\rangle_W\right|^2\right) + \sum_{i=\ell+1}^{m} \left|\langle y_0, u_i\rangle_W\right|^2.\end{aligned}$$

As a consequence of (1.60) and (1.62) we have $\lim_{n\to\infty} \|u_i^n - u_i\|_W = 0$ for $i = 1, \ldots, \ell$ and hence (1.64) follows.

Summarizing we have the following theorem.

THEOREM 2.14. *Assume that $y \in C^1([0,T];\mathbb{R}^m)$ is the unique solution to (1.41). Let $\{(u_i^n, \lambda_i^n)\}_{i=1}^m$ and $\{(u_i, \lambda_i)\}_{i=1}^m$ be the eigenvector-eigenvalue pairs given by (1.55). Suppose that $\ell \in \{1, \ldots, m\}$ is fixed such that (1.62) and*

$$\sum_{i=\ell+1}^{m} \lambda_i \neq 0, \quad \sum_{i=\ell+1}^{m} \left|\langle y_0, u_i\rangle_W\right|^2 \neq 0$$

hold. Then we have

$$\lim_{n\to\infty} \|\mathcal{R}^n - \mathcal{R}\|_{L(\mathbb{R}^m)} = 0. \tag{1.67}$$

This implies

$$\begin{aligned}&\lim_{n\to\infty} \left|\lambda_i^n - \lambda_i\right| = \lim_{n\to\infty} \|u_i^n - u_i\|_W = 0 \quad \textit{for } 1 \leq i \leq \ell, \\ &\lim_{n\to\infty} \sum_{i=\ell+1}^{m} \left(\lambda_i^n - \lambda_i\right) = 0 \quad \textit{and} \quad \lim_{n\to\infty} \sum_{i=\ell+1}^{m} \left|\langle y_0, u_i^n\rangle_W\right|^2 = \sum_{i=\ell+1}^{m} \left|\langle y_0, u_i\rangle_W\right|^2.\end{aligned}$$

PROOF. We only have to verify (1.67). For that purpose we choose an arbitrary $u \in \mathbb{R}^m$ with $\|u\|_W = 1$ and introduce $f_u : [0,T] \to \mathbb{R}^m$ by

$$f_u(t) = \langle y(t), u\rangle_W \, y(t) \quad \text{for } t \in [0,T].$$

Then, we have $f_u \in C^1([0,T];\mathbb{R}^m)$ with

$$\dot{f}_u(t) = \langle \dot{y}(t), u\rangle_W \, y(t) + \langle y(t), u\rangle_W \, \dot{y}(t) \quad \text{for } t \in [0,T]$$

By Taylor expansion there exist $\tau_{j1}(t), \tau_{j2}(t) \in [t_j, t_{j+1}]$ depending on t

$$\begin{aligned}\int_{t_j}^{t_{j+1}} f_u(t)\,\mathrm{d}t &= \frac{1}{2}\int_{t_j}^{t_{j+1}} f_u(t_j) + \dot{f}_u(\tau_{j1}(t))(t-t_j)\,\mathrm{d}t \\ &\quad + \frac{1}{2}\int_{t_j}^{t_{j+1}} f_u(t_{j+1}) + \dot{f}_u(\tau_{j2}(t))(t-t_{j+1})\,\mathrm{d}t \\ &= \frac{\Delta t}{2}\left(f_u(t_j)+f_u(t_{j+1})\right) + \frac{1}{2}\int_{t_j}^{t_{j+1}} \dot{f}_u(\tau_{j1}(t))(t-t_j)\,\mathrm{d}t \\ &\quad + \frac{1}{2}\int_{t_j}^{t_{j+1}} \dot{f}_u(\tau_{j2}(t))(t-t_{j+1})\,\mathrm{d}t.\end{aligned}$$

Hence,

$$\begin{aligned}\left\|\mathcal{R}^n u - \mathcal{R}u\right\|_W &= \left\|\sum_{j=1}^{n}\alpha_j f_u(t_j) - \int_0^T f_u(t)\,\mathrm{d}t\right\|_W \\ &= \left\|\sum_{j=1}^{n-1}\left(\frac{\Delta t}{2}\left(f_u(t_j)+f_u(t_{j+1})\right) - \int_{t_j}^{t_{j+1}} f_u(t)\,\mathrm{d}t\right)\right\|_W \\ &\le \frac{1}{2}\sum_{j=1}^{n-1}\int_{t_j}^{t_{j+1}} \left\|\dot{f}_u(\tau_{j1}(t))\right\|_W |t-t_j| + \left\|\dot{f}_u(\tau_{j2}(t))\right\|_W |t-t_{j+1}|\,\mathrm{d}t \\ &\le \frac{1}{2}\max_{t\in[0,T]}\left\|\dot{f}_u(t)\right\|_W \sum_{j=1}^{n-1}\left(\frac{(t-t_j)^2}{2} - \frac{(t_{j+1}-t)^2}{2}\Big|_{t=t_j}^{t=t_{j+1}}\right) \\ &= \frac{\Delta t}{2}\max_{t\in[0,T]}\left\|\dot{f}_u(t)\right\|_W \sum_{j=1}^{n-1}\Delta t = \frac{\Delta t\, T}{2}\max_{t\in[0,T]}\left\|\dot{f}_u(t)\right\|_W \\ &\le \frac{\Delta t\, T}{2}\max_{t\in[0,T]}\left\|\dot{f}_u(t)\right\|_W \\ &= \frac{\Delta t\, T}{2}\max_{t\in[0,T]}\left\|\langle \dot{y}(t), u\rangle_W\, y(t) + \langle y(t), u\rangle_W\, \dot{y}(t)\right\|_W \\ &= \Delta t\, T \max_{t\in[0,T]}\|\dot{y}(t)\|_W \|y(t)\|_W \le \Delta t\, T\, \|y\|^2_{C^1([0,T];\mathbb{R}^m)}.\end{aligned}$$

Consequently,

$$\left\|\mathcal{R}^n - \mathcal{R}\right\|_{L(\mathbb{R}^m)} = \sup_{\|u\|_W=1}\|\mathcal{R}^n u - \mathcal{R}u\|_W \le 2\Delta t\,\|y\|^2_{C^1([0,T];\mathbb{R}^m)} \overset{\Delta t\to 0}{\longrightarrow} 0$$

which is (1.67). □

2.4. POD for parabolic problems. Let V and H be real separable Hilbert spaces and suppose that V is dense in H with compact embedding. By $\langle\cdot,\cdot\rangle_H$ we denote the inner product in H. The inner product in V is given by a symmetric bounded, coercive, bilinear form $a : V \times V \to \mathbb{R}$:

$$\langle \varphi, \psi\rangle_V = a(\varphi,\psi) \quad \text{for all } \varphi,\psi \in V \tag{1.68}$$

with associated norm given by $\|\cdot\|_V = \sqrt{a(\cdot,\cdot)}$. Since V is continuously injected into H, there exists a constant $c_V > 0$ such that

$$\|\varphi\|_H \le c_V\, \|\varphi\|_V \quad \text{for all } \varphi \in V. \tag{1.69}$$

We associate with a the linear operator A:

$$\langle A\varphi, \psi\rangle_{V',V} = a(\varphi,\psi) \quad \text{for all } \varphi,\psi \in V,$$

where $\langle \cdot\,, \cdot \rangle_{V',V}$ denotes the duality pairing between V and its dual. Then A is an isomorphism from V onto V'. Alternatively, A can be considered as a linear unbounded self-adjoint operator in H with domain

$$D(A) = \{\varphi \in V : A\varphi \in H\}.$$

By identifying H and its dual H' it follows that

$$D(A) \hookrightarrow V \hookrightarrow H = H' \hookrightarrow V',$$

each embedding being continuous and dense, when $D(A)$ is endowed with the graph norm of A.

We introduce the continuous operator $R : V \to V'$, which maps $D(A)$ into H and satisfies

$$\begin{aligned} \|R\varphi\|_H &\leq c_R \, \|\varphi\|_V^{1-\delta_1} \|A\varphi\|_H^{\delta_1} \quad \text{for all } \varphi \in D(A), \\ |\langle R\varphi, \varphi\rangle_{V',V}| &\leq c_R \, \|\varphi\|_V^{1+\delta_2} \|\varphi\|_H^{1-\delta_2} \quad \text{for all } \varphi \in V \end{aligned}$$

for a constant $c_R > 0$ and for $\delta_1, \delta_2 \in [0, 1)$. We also assume that $A + R$ is coercive on V, i.e., there exists a constant $\eta > 0$ such that

$$a(\varphi, \varphi) + \langle R\varphi, \varphi\rangle_{V',V} \geq \eta \, \|\varphi\|_V^2 \quad \text{for all } \varphi \in V.$$

Moreover, let $B : V \times V \to V'$ be a bilinear continuous operator mapping $D(A) \times D(A)$ into H such that there exist constants $c_B > 0$ and $\delta_3, \delta_4, \delta_5 \in [0, 1)$ satisfying

$$\begin{aligned} \langle B(\varphi, \psi), \psi\rangle_{V',V} &= 0, \\ \big|\langle B(\varphi, \psi), \phi\rangle_{V',V}\big| &\leq c_B \, \|\varphi\|_H^{\delta_3} \|\varphi\|_V^{1-\delta_3} \|\psi\|_V \|\phi\|_V^{\delta_3} \|\phi\|_H^{1-\delta_3}, \\ \|B(\varphi, \chi)\|_H + \|B(\chi, \varphi)\|_H &\leq c_B \, \|\varphi\|_V \|\chi\|_V^{1-\delta_4} \|A\chi\|_H^{\delta_4}, \\ \|B(\varphi, \chi)\|_H &\leq c_B \, \|\varphi\|_H^{\delta_5} \|\varphi\|_V^{1-\delta_5} \|\chi\|_V^{1-\delta_5} \|A\chi\|_H^{\delta_5}, \end{aligned}$$

for all $\varphi, \psi, \phi \in V$, for all $\chi \in D(A)$. To simplify the notation we set $B(\varphi) = B(\varphi, \varphi)$ for $\varphi \in V$.

For given $f \in L^2(0, T; H)$ and $y_0 \in V$ we consider the nonlinear evolution problem

$$\frac{d}{dt} \, \langle y(t), \varphi\rangle_H + a(y(t), \varphi) + \big\langle B(y(t)) + Ry(t), \varphi\big\rangle_{V',V} = \langle f(t), \varphi\rangle_H \tag{1.70a}$$

for all $\varphi \in V$ and $t \in (0, T]$ a.e. and

$$y(0) = y_0 \quad \text{in } H. \tag{1.70b}$$

It follows from Theorem 2.1 in [**39**, p. 111] that (1.70) has a unique solution satisfying

$$y \in C([0, T]; V) \cap L^2(0, T; D(A)) \cap H^1(0, T; H). \tag{1.71}$$

Next we discuss the POD method for (1.70). We denote by y the unique solution to (1.70) satisfying (1.71). Moreover, we suppose that $f \in C([0, T]; H)$. For given $n \in \mathbb{N}$ let

$$0 = t_0 < t_2 < \ldots < t_n \leq T$$

denote a grid in the interval $[0, T]$ and set $\delta t_j = t_j - t_{j-1}$, $j = 1, \ldots, n$. Define

$$\Delta t = \max\,(\delta t_1, \ldots, \delta t_n) \quad \text{and} \quad \delta t = \min\,(\delta t_1, \ldots, \delta t_n).$$

Suppose that the snapshots $y(t_j)$ of (1.70) at the given time instances t_j, $j = 0, \ldots, n$, are known. We set

$$\mathcal{V} = \text{span}\,\{y(t_0), \ldots, y(t_n)\},$$

and refer to $\mathcal{V}$ as the ensemble consisting of the snapshots $\{y(t_j)\}_{j=0}^n$, at least one of which is assumed to be nonzero. Notice that $\mathcal{V} \subset V$ by construction. Throughout the remainder of this section we let X denote either the space V or H.

Let $\{\psi_i\}_{i=1}^d$ denote an orthonormal basis for $\mathcal{V}$ with $d = \dim \mathcal{V}$. Then each member of the ensemble can be expressed as

$$y(t_j) = \sum_{i=1}^{d} \langle y(t_j), \psi_i\rangle_X \psi_i \quad \text{for } j = 0, \ldots, n. \tag{1.72}$$

The method of POD consists in choosing an orthonormal basis such that for every $\ell \in \{1, \dots, d\}$ the mean square error between the elements $y(t_j)$, $0 \le j \le n$, and the corresponding ℓ-th partial sum of (1.72) is minimized on average:

$$\min_{\{\psi_i\}_{i=1}^{\ell}} \sum_{j=0}^{n} \alpha_j \Big\| y(t_j) - \sum_{i=1}^{\ell} \langle y(t_j), \psi_i \rangle_X \psi_i \Big\|_X^2 \tag{1.73}$$
$$\text{subject to } \langle \psi_i, \psi_j \rangle_X = \delta_{ij} \quad \text{for } 1 \le i \le \ell, 1 \le j \le i.$$

Here $\{\alpha_j\}_{j=0}^n$ are positive weights, which for our purposes are chosen to be

$$\alpha_0 = \frac{\delta t_1}{2}, \quad \alpha_j = \frac{\delta t_j + \delta t_{j+1}}{2} \text{ for } j = 1, \dots, n-1, \quad \text{and} \quad \alpha_n = \frac{\delta t_n}{2}.$$

A solution $\{\psi_i\}_{i=1}^{\ell}$ to (1.73) is called POD basis of rank ℓ. The subspace spanned by the first ℓ POD basis functions is denoted by V^ℓ.

The solution of (1.73) is characterized by the necessary optimality condition. For that purpose we endow $\mathbb{R}^{n+1}$ with the weighted inner product

$$\langle v, w \rangle_{\mathbb{R}^{n+1}} = \sum_{j=0}^{n} \alpha_j v_j w_j \quad \text{for } v = (v_0, \dots, v_n)^{\mathsf{T}},\ w = (w_0, \dots, w_n)^{\mathsf{T}} \in \mathbb{R}^{n+1}.$$

Let us introduce the bounded linear operator $\mathcal{Y}_n : \mathbb{R}^{n+1} \to X$ by

$$\mathcal{Y}_n v = \sum_{j=0}^{n} \alpha_j v_j y(t_j) \quad \text{for } v \in \mathbb{R}^{n+1}.$$

Then the adjoint $\mathcal{Y}_n^* : X \to \mathbb{R}^{n+1}$ is given by

$$\mathcal{Y}_n^* z = (\langle z, y(t_0) \rangle_X, \dots, \langle z, y(t_n) \rangle_X)^{\mathsf{T}} \quad \text{for } z \in X.$$

It follows that $\mathcal{R}_n = \mathcal{Y}_n \mathcal{Y}_n^* \in \mathcal{L}(X)$ and $\mathcal{K}_n = \mathcal{Y}_n^* \mathcal{Y}_n \in \mathbb{R}^{(n+1)\times(n+1)}$ are given by

$$\mathcal{R}_n z = \sum_{j=0}^{n} \alpha_j \langle z, y(t_j) \rangle_X y(t_j) \text{ for } z \in X \quad \text{and} \quad \big(\mathcal{K}_n\big)_{ij} = \langle y(t_j), y(t_i) \rangle_X,$$

respectively. Here $\mathcal{L}(X)$ denotes the Banach space of all bounded linear operators on X.

Using a Lagrangian framework we derive the following optimality conditions for the optimization problem (1.73):

$$\mathcal{R}_n \psi = \lambda \psi,$$

compare e.g. [**12, 43**]. Note that $\mathcal{R}_n$ is a bounded, self-adjoint and nonnegative operator. Moreover, since the image of $\mathcal{R}_n$ has finite dimension, $\mathcal{R}_n$ is also compact. By Hilbert–Schmidt theory (see e.g. [**36**, p. 203]) there exist an orthonormal basis $\{\psi_i\}_{i\in\mathbb{N}}$ for X and a sequence $\{\lambda_i\}_{i\in\mathbb{N}}$ of nonnegative real numbers so that

$$\mathcal{R}_n \psi_i = \lambda_i \psi_i, \quad \lambda_1 \ge \ldots \ge \lambda_d > 0 \quad \text{and } \lambda_i = 0 \text{ for } i > d.$$

Moreover, $\mathcal{V} = \operatorname{span}\{\psi_i\}_{i=1}^d$.

Note that $\{\lambda_i\}_{i\in\mathbb{N}}$ as well as $\{\psi_i\}_{i\in\mathbb{N}}$ depend on n. Contents permitting the notation of this dependence is dropped.

Remark 2.15. Setting

$$v_i = \frac{1}{\sqrt{\lambda_i}}\, \mathcal{Y}_n^* \psi_i \quad \text{for } i = 1, \dots, d$$

we find $\mathcal{K}_n v_i = \lambda_i v_i$ and $\langle v_i, v_j \rangle_{\mathbb{R}^{n+1}} = \delta_{ij}$ for $1 \le i, j \le d$. Thus, $\{v_i\}_{i=1}^d$ is an orthonormal basis of eigenvectors of $\mathcal{K}_n$ for the image of $\mathcal{K}_n$. Conversely, if $\{v_i\}_{i=1}^d$ is a given orthonormal basis for the image of $\mathcal{K}_n$, then it follows that the first d eigenfunctions of $\mathcal{R}_n$ can be determined by

$$\psi_i = \frac{1}{\sqrt{\lambda_i}}\, \mathcal{Y}_n v_i \quad \text{for } i = 1, \dots, d. \ \diamondsuit$$

The sequence $\{\psi_i\}_{i=1}^{\ell}$ solves the optimization problem (1.73). This fact as well as the error formula below were proved in [**12, 43**], for example.

PROPOSITION 2.16. *Let $\lambda_1 \geq \ldots \geq \lambda_d > 0$ denote the positive eigenvalues of $\mathcal{R}^n$ with the associated eigenvectors $\psi_1, \ldots, \psi_d \in X$. Then, $\{\psi_i^n\}_{i=1}^{\ell}$ is a POD basis of rank $\ell \leq d$, and we have the error formula*

$$\sum_{j=0}^{n} \alpha_j \Big\| y(t_j) - \sum_{i=1}^{\ell} \langle y(t_j), \psi_i \rangle_X \psi_i \Big\|_X^2 = \sum_{i=\ell+1}^{d} \lambda_i. \tag{1.74}$$

The eigenvalues $\{\lambda_i\}_{i\in\mathbb{N}}$ depend on the time instances $\{t_j\}_{j=0}^n$. Next we investigate $\sum_{i=\ell+1}^d \lambda_i$ as Δt tends to zero, i.e., $n \to \infty$. Let us define the bounded linear operator $\mathcal{Y} : L^2(0,T;\mathbb{R}) \to X$ by

$$\mathcal{Y}\varphi = \int_0^T \varphi(t) y(t)\, dt \quad \text{for } \varphi \in L^2(0,T;\mathbb{R}).$$

The adjoint $\mathcal{Y}^* : X \to L^2(0,T;\mathbb{R})$ is given by

$$\big(\mathcal{Y}^* z\big)(t) = \langle z, y(t) \rangle_X \quad \text{for } z \in X.$$

For $\mathcal{R} = \mathcal{Y}\mathcal{Y}^* \in \mathcal{L}(X)$ we find

$$\mathcal{R}z = \int_0^T \langle z\,, y(t) \rangle_X y(t)\, dt \quad \text{for } z \in X.$$

Notice that $\mathcal{R}_n\varphi$ is the trapezoidal approximation for the integral $\mathcal{R}\varphi$. If $y_t \in L^2(0,T;X)$ then we obtain

$$\lim_{\Delta t \to \infty} \|\mathcal{R}_n - \mathcal{R}\|_{\mathcal{L}(X)} = 0. \tag{1.75}$$

Let us mention that as far as the following analysis is concerned any other choice of positive weights α_j is possible provided that (1.75) hold.

We proceed to investigate the relationship between $\mathcal{R}_n$ and $\mathcal{R}$. Notice that $\mathcal{R}$ is self-adjoint and nonnegative. Since $y \in C([0,T];V)$, the Kolmogorov compactness criterion in $L^2(0,T;\mathbb{R})$ implies that $\mathcal{Y}^* : X \to L^2(0,T;X)$ is compact. Boundedness of $\mathcal{Y}$ implies that $\mathcal{R}$ is a compact operator as well. From the Hilbert–Schmidt theorem it follows that there exists a complete orthonormal basis $\{\psi_i^\infty\}_{i\in\mathbb{N}}$ for X and a sequence $\{\lambda_i^\infty\}_{i\in\mathbb{N}}$ of nonnegative real numbers so that

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

REMARK 2.17. Analogous to Remark 2.15 we set

$$v_i^\infty = \frac{1}{\sqrt{\lambda_i^\infty}}\, \mathcal{Y}^* \psi_i^\infty = \frac{1}{\sqrt{\lambda_i^\infty}}\, \langle \psi_i^\infty, y(t) \rangle_X\, dt \quad \text{for } i \in \{j \in \mathbb{N} : \lambda_j^\infty > 0\}.$$

Let $\mathcal{K} = \mathcal{Y}^*\mathcal{Y} \in \mathcal{L}(L^2(0,T;\mathbb{R}))$ be given by

$$\mathcal{K}\varphi = \int_0^T \langle y(s), y(t) \rangle_X \varphi(s)\, ds \quad \text{for } \varphi \in L^2(0,T;\mathbb{R}).$$

It follows that

$$\big(\mathcal{K} v_i^\infty\big)(t) = \lambda_i^\infty v_i^\infty(t)$$

and consequently, the v_i^∞'s are the eigenfunctions of $\mathcal{K}$ for $i \in \mathbb{N}$ with $\lambda_i^\infty > 0$. $\Diamond$

Henceforth we denote by $\{\lambda_i^n\}_{i=1}^{d(n)}$ the positive eigenvalues of $\mathcal{R}_n$ with associated eigenfunctions $\{\psi_i^n\}_{i=1}^{d(n)}$. Similarly $\{\lambda_i^\infty\}_{i\in\mathbb{N}}$ denote the positive eigenvalues of $\mathcal{R}$ with associated eigenfunctions $\{\psi_i^\infty\}_{i\in\mathbb{N}}$. In each case the eigenvalues are considered according to their multiplicity. Now choose and fix

$$\ell \quad \text{such that} \quad \lambda_\ell^\infty \neq \lambda_{\ell+1}^\infty. \tag{1.76}$$

It follows from spectral analysis of compact operators ([**19**, pp. 212–214]) and from [**22**] that there exists $\overline{\Delta t} > 0$ such that

$$\sum_{i=\ell+1}^{\infty} \lambda_i^n \leq 2 \sum_{i=\ell+1}^{\infty} \lambda_i^\infty \quad \text{for all } \Delta t \leq \overline{\Delta t}, \tag{1.77}$$

if $\sum_{i=\ell+1}^{\infty} \lambda_i^\infty \neq 0$. Moreover, for ℓ as above, $\overline{\Delta t}$ can also be chosen such that

$$\sum_{i=\ell+1}^{d(n)} \left|\langle \psi_i^n, y_0 \rangle_X\right|^2 \leq 2 \sum_{i=\ell+1}^{\infty} \left|\langle \psi_i^\infty, y_0 \rangle_X\right|^2 \quad \text{for all } \Delta \leq \overline{\Delta t}, \tag{1.78}$$

provided that $\sum_{i=\ell+1}^{\infty} |\langle y_0, \psi_i^\infty \rangle_X|^2 \neq 0$. As a consequence of (1.75) and (1.76) we have $\lim_{\Delta t \to 0} \psi_i^n = \psi_i^\infty$ for $i = 1, \ldots, \ell$ and hence (1.78) follows.

2.5. POD for parameter-dependent elliptic systems. As in the previous 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. Since V is continuously injected into H, there exists a constant $c_V > 0$ satisfying (1.69).

For $\mu_a, \mu_b \in \mathbb{R}$ with $\mu_a < \mu_b$ we introduce the interval $\mathcal{I} = [\mu_a, \mu_b]$ containing the admissible values for the parameters. Then we define the parametrized bilinear form $a : V \times V \times \mathcal{I} \to \mathbb{R}$ as

$$a(\varphi, \phi; \mu) = \langle \varphi, \phi \rangle_V + \mu \langle \varphi, \phi \rangle_H \quad \text{for } \varphi, \phi \in V \text{ and } \mu \in \mathcal{I}.$$

For any $\mu \in \mathcal{I}$ we obtain

$$|a(\varphi, \phi; \mu)| \leq \left(1 + c_V^2 \max\{|\mu_a|, |\mu_b|\}\right) \|\varphi\|_V \|\phi\|_V \quad \text{for all } \varphi, \phi \in V,$$

i.e., the bilinear form $a(\cdot, \cdot; \mu)$ is continuous on $V \times V$ for any $\mu \in \mathcal{I}$. Since

$$a(\varphi, \varphi; \mu) = \|\varphi\|_V^2 + \mu \|\varphi\|_H^2 \quad \text{for all } \varphi \in V \text{ and } \mu \in \mathcal{I},$$

it follows that $a(\cdot, \cdot; \mu)$ is coercive on $V \times V$ for every $\mu \in \mathcal{I}$ provided

$$\eta_a = 1 + 2c_V^2 \min\{0, \mu_a\} > 0. \tag{1.79}$$

Let $f \in V'$ be given. For given parameter $\mu \in \mathcal{I}$ we consider the following variational problem: Find $u = u(\mu) \in V$ such that

$$a(u, \varphi; \mu) = \langle f, \varphi \rangle_{V',V} \quad \text{for all } \varphi \in V, \tag{1.80}$$

where $\langle \cdot, \cdot \rangle_{V',V}$ stands for the duality pairing of V and its dual space V'.

If (1.79) holds, it follows from the Lax-Milgram lemma [**5**] that for every $\mu \in \mathcal{I}$ there exists a unique solution $u = u(\mu) \in V$ to (1.80).

Together with (1.80) we will consider a discretized variational problem, where we apply POD for the discretization of V. We follow the arguments for time-dependent systems. Henceforth, we denote by $u = u(\mu) \in V$ the associated solution to (1.80) for chosen parameter $\mu \in \mathcal{I}$. We define the bounded linear operator $\mathcal{Y} : L^2(\mathcal{I}) \to V$ by

$$\mathcal{Y}\varphi = \int_{\mathcal{I}} \varphi(\mu) u(\mu) \, \mathrm{d}\mu \quad \text{for } \varphi \in L^2(\mathcal{I}).$$

Its Hilbert space adjoint $\mathcal{Y}^* : V \to L^2(\mathcal{I})$ is given by

$$\left(\mathcal{Y}^* z\right)(\mu) = \langle z, u(\mu) \rangle_V \quad \text{for } z \in V \text{ and } \mu \in \mathcal{I}.$$

Furthermore, we find that the bounded, linear, symmetric and non-negative operator $\mathcal{R} = \mathcal{Y}\mathcal{Y}^* : V \to V$ has the form

$$\mathcal{R}z = \int_{\mathcal{I}} \langle z, u(\mu) \rangle_V u(\mu) \, \mathrm{d}\mu \quad \text{for } z \in V. \tag{1.81}$$

The operator $\mathcal{K} = \mathcal{Y}^*\mathcal{Y} : L^2(\mathcal{I}) \to L^2(\mathcal{I})$ is given by

$$\big(\mathcal{K}\varphi\big)(\bar{\mu}) = \int_{\mathcal{I}} \langle u(\mu), u(\bar{\mu})\rangle_V \varphi(\mu)\,\mathrm{d}\mu \quad \text{for } \varphi \in L^2(\mathcal{I}). \tag{1.82}$$

It follows that $\mathcal{K} = \mathcal{Y}^*\mathcal{Y}$ is compact and, therefore, $\mathcal{R} = \mathcal{Y}\mathcal{Y}^*$ is compact as well. From the Hilbert-Schmidt theorem it follows that there exists a complete orthonormal basis $\{\psi_i\}_{i\in\mathbb{N}}$ for V and a sequence $\{\lambda_i\}_{i\in\mathbb{N}}$ of non-negative real numbers so that

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

Remark 2.18. Analogous to the theory of singular value decomposition for matrices, we find that the bounded, linear, symmetric and non-negative operator $\mathcal{K}$ (see (1.82) has the same eigenvalues $\{\lambda_i\}_{i\in\mathbb{N}}$ as the operator $\mathcal{R}$ and the eigenfunctions

$$v_i(t) = \frac{1}{\sqrt{\lambda_i}}\,\big(\mathcal{C}^*\psi_i\big)(\mu) = \frac{1}{\sqrt{\lambda_i}}\,\langle \psi_i, u(\mu)\rangle_V$$

for $i \in \{j \in \mathbb{N} : \lambda_j > 0\}$ and almost all $\mu \in D$. ◊

For given $\ell \in \mathbb{N}$ we introduce the mapping

$$\mathfrak{J} : \underbrace{V \times \ldots \times V}_{\ell-\text{times}} \to \mathbb{R}$$

by

$$\mathfrak{J}(\psi_1, \ldots, \psi_\ell) = \int_{\mathcal{I}} \Big\| u(\mu) - \sum_{i=1}^{\ell} \langle u(\mu), \psi_i\rangle_V \psi_i \Big\|_V^2 \,\mathrm{d}\mu. \tag{1.83}$$

In the following proposition [**12**, Section 3.3] we formulate properties of the eigenvalues and eigenfunctions of $\mathcal{R}$.

Proposition 2.19. *Let $\{\lambda_i\}_{i\in\mathbb{N}}$ and $\{\psi_i\}_{i\in\mathbb{N}}$ denote the eigenvalues and eigenfunctions, respectively, of $\mathcal{R}$ introduced in* (1.81)*. Then, for every $\ell \in \mathbb{N}$ the first ℓ eigenfunctions $\psi_1, \ldots, \psi_\ell \in V$ solve the minimization problem*

$$\min \mathfrak{J}(\tilde{\psi}_1, \ldots, \tilde{\psi}_\ell) \quad \textit{s.t.} \quad \langle \tilde{\psi}_j, \tilde{\psi}_i\rangle_V = \delta_{ij} \quad \textit{for } 1 \leq i, j \leq \ell, \tag{1.84}$$

where $\mathfrak{J}$ is defined in (1.83)*. Moreover,*

$$\mathfrak{J}(\psi_1, \ldots, \psi_\ell) = \sum_{i=\ell+1}^{\infty} \lambda_i \quad \textit{for any } \ell \in \mathbb{N}.$$

In applications the weak solution to (1.80) is not known for all parameters $\mu \in \mathcal{I}$, but only for a given grid in $\mathcal{I}$. For that purpose let

$$\mu_a = \mu_1 < \mu_2 < \ldots < \mu_n = \mu_b \tag{1.85}$$

be a grid in $\mathcal{I}$ and let $u_i = u(\mu_i)$, $1 \leq i \leq n$, denote the corresponding solutions to (1.79) for the grid points μ_i. We define the snapshot set $\mathcal{V}^n = \operatorname{span}\{u_1, \ldots, u_n\} \subset V$ and determine a POD basis of rank $\ell \leq n$ for $\mathcal{V}^n$ by solving

$$\min \sum_{j=1}^{n} \alpha_j \Big\| u_j - \sum_{i=1}^{\ell} \langle u_j, \psi_i\rangle_V \psi_i \Big\|_V^2 \quad \text{s.t.} \quad \langle \psi_i, \psi_j\rangle_V = \delta_{ij},\ 1 \leq i, j \leq \ell \tag{1.86}$$

where the α_j's are non-negative weights. The solution to (1.86) is given by the solution to the eigenvalue problem

$$\mathcal{R}^n\psi_i^n = \lambda_i^n\psi_i^n, \quad i = 1, \ldots, \ell,$$

with

$$\mathcal{R}^n\psi = \sum_{j=1}^{n} \alpha_j \langle u_j, \psi\rangle_V u_j \quad \text{for } \psi \in V.$$

In contrast to $\mathcal{R}$ introduced in (1.81) the operator $\mathcal{R}^n$ and therefore its eigenvalues and eigenfunctions depend on the grid $\{\mu_j\}_{j=1}^n$. Furthermore, the image space of $\mathcal{R}^n$ has finite dimension $d^n \leq n$, whereas, in general, the image space of the operator $\mathcal{R}$ is infinite-dimensional. Since $\mathcal{R}^n$ is a linear, bounded, compact, non-negative, self-adjoint operator, there exist eigenvalues $\{\lambda_i^n\}_{i=1}^{d^n}$ and orthonormal eigenfunctions $\{\psi_i^n\}_{i=1}^{\ell}$ with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{d^n} > 0$ and

$$\sum_{j=1}^{n} \alpha_j \left\| u_j - \sum_{i=1}^{\ell} \langle u_j, \psi_i\rangle_V \psi_i \right\|_V^2 = \sum_{i=\ell+1}^{d^n} \lambda_i^n.$$

REMARK 2.20 (Snapshot POD [**38**]). Let us supply $\mathbb{R}^n$ with the weighted inner product

$$\langle v, w\rangle_{\mathbb{R}^n} = \sum_{i=1}^{n} \alpha_i v_i w_i \quad \text{for } v = (v_1, \ldots, v_n)^T, w = (w_1, \ldots, w_n)^T \in \mathbb{R}^n.$$

If the α_i's are trapezoidal weights corresponding to the parameter grid $\{\mu_i\}_{i=1}^n$ then the inner product $\langle \cdot, \cdot\rangle_{\mathbb{R}^n}$ is a discrete version of the inner product in $L^2(\mathcal{I})$. We define the symmetric non-negative matrix $\mathcal{K}^n \in \mathbb{R}^{n\times n}$ with the elements $\langle u_i, u_j\rangle_V$, $1 \leq i, j \leq n$, and consider the eigenvalue problem

$$\mathcal{K}^n v_i^n = \lambda_i^n v_i^n,\ 1 \leq i \leq \ell \quad \text{and} \quad \langle v_i^n, v_j^n\rangle_{\mathbb{R}^n} = \delta_{ij},\ 1 \leq i, j \leq \ell \leq d^n \tag{1.87}$$

From singular value decomposition it follows that $\mathcal{K}^n$ has the same eigenvalues $\{\lambda_i^n\}_{i=1}^{d^n}$ as the operator $\mathcal{R}^n$; compare Remark 2.18 and [**22**]. Furthermore, the POD basis functions are given by the formula

$$\psi_i = \frac{1}{\sqrt{\lambda_i^n}} \sum_{j=1}^{n} \alpha_j (v_i^n)_j u_j \quad \text{for } i = 1, \ldots, \ell,$$

where $(v_i^n)_j$ denotes the jth-component of the eigenvector $v_i^n \in \mathbb{R}^n$. ◊

3. Balanced truncation and POD method

In Section 1.4 the controllability Gramian L_c and the observability Gramian L_o have been introduced. Both Gramians can be computed by solving the Lyapunov equations (1.8) and (1.7), respectively. In another approach, see [**26, 32**], L_c and L_o can be derived from numerical simulations. This gives the possibility to extend the balanced truncation method to nonlinear systems. In this context, L_c and L_o are called *empirical Gramians*.

Let us consider the linear, time-invariant system (1.1) with m_u inputs. We write $B = (b_1|\ldots|b_{m_u}) \in \mathbb{R}^{m_x\times m_u}$ with $b_i \in \mathbb{R}^{m_x}$ for $1 \leq i \leq m_x$. Then, its solution/response $x^i(t) \in \mathbb{R}^{m_x}$ to unit impulses $u^i(t) = \delta(t)e_i$, where $e_i \in \mathbb{R}^{m_x}$ denotes the i-th canonical unit vector and

$$\int_0^T \delta(s) f(s)\,\mathrm{d}s = f(0) \quad \text{for any continuous } f : [0, T] \to \mathbb{R},$$

satisfies

$$\begin{aligned} x^i(t) &= e^{tA}x_0 + \int_0^T e^{(t-s)A} B u^i(t)\,\mathrm{d}s = e^{tA}x_0 + \int_0^T \delta(s) e^{(t-s)A} B e_i\,\mathrm{d}s \\ &= e^{tA}x_0 + \int_0^T \delta(s) e^{(t-s)A} b_i\,\mathrm{d}s = e^{tA}x_0 + e^{tA}b_i. \end{aligned}$$

In particular, for $x_0 = 0$ we find $x^i(t) = e^{tA}b_i$ for $1 \leq i \leq m_u$. From Lemma 1.15 we have

$$\begin{aligned} L_c &= \int_0^\infty e^{tA}BB^T e^{tA^T}\,\mathrm{d}t = \int_0^\infty e^{tA}\big(b_1|\ldots|b_{m_u}\big)\big(b_1|\ldots|b_{m_u}\big)^T e^{tA^T}\,\mathrm{d}t \\ &= \int_0^\infty \big(e^{tA}b_1|\ldots|e^{tA}b_{m_u}\big)\big(e^{tA}b_1|\ldots|e^{tA}b_{m_u}\big)^T\,\mathrm{d}t \\ &= \int_0^\infty \big(x^1(t)|\ldots|x^{m_u}(t)\big)\big(x^1(t)|\ldots|x^{m_u}(t)\big)^T\,\mathrm{d}t \\ &= \int_0^\infty x^1(t)\big(x^1(t)\big)^T + \ldots + x^{m_u}(t)\big(x^{m_u}(t)\big)^T\,\mathrm{d}t. \end{aligned}$$

Hence, the controllability matrix can also be computed from m_u the simulations $x^i(t)$ with the inputs $u^i(t) = \delta(t)e_i$, $1 \leq i \leq m_u$.

Let us introduce the snapshot set

$$\mathcal{V} = \bigcup_{i=1}^{m_u} \operatorname{span}\left\{x^i(t)\,|\,t \in [0,\infty)\right\} \subset \mathbb{R}^{m_x}.$$

We consider the following problem

$$\begin{cases} \min\limits_{\psi_1,\ldots,\psi_\ell} \displaystyle\int_0^\infty \sum_{j=1}^{m_u} \Big\| x^j(t) - \sum_{i=1}^{\ell} \langle x^j(t), \psi_i\rangle_{\mathbb{R}^{m_x}}\,\psi_i \Big\|^2_{\mathbb{R}^{m_x}}\,\mathrm{d}t \\ \text{s.t. } \langle \psi_i, \psi_j\rangle_{\mathbb{R}^{m_x}} = \delta_{ij} \text{ for } 1 \leq i,j \leq \ell \end{cases} \tag{1.1}$$

The first-order necessary optimality conditions for (1.1) are given by the symmetric eigenvalue problem

$$\tilde{\mathcal{R}}\psi_i = \lambda_i\psi_i \quad \text{for } 1 \leq i \leq \ell$$

with $\tilde{\mathcal{R}} = L_c$. Hence, the POD eigenvectors for the snapshot set $\mathcal{V}$ of the impulse responses are the eigenvectors of the controllability Gramian corresponding to the largest eigenvalues. Therefore, $\psi_1, \ldots, \psi_\ell$ are the most controllable modes of the realization.

Utilizing the dual equations an analogous approach can be used for the computation of the observability Gramian [**37**].

CHAPTER 2

Reduced-Order Modelling with POD

If the POD basis is computed, it can be used to derive a so-called *low-dimensional approximation* or a *reduced-order model* (ROM). This is the focus of this chapter, which is organized as follows: In Section 1 we consider a ROM for dynamical systems in $\mathbb{R}^m$ and derive error estimates in Section 2. A ROM for parabolic problems (see previous chapter in Section 2.4) are studied in Section 3. Error estimates are presented from the works [**21, 22, 23**]. A ROM for elliptic problems is investigated in Section 4. For the error analysis estimates are presented from [**15**].

1. ROM for dynamical systems

Suppose that we have determined a POD basis $\{u_j\}_{j=1}^{\ell}$ of rank $\ell \in \{1,\dots,m\}$ in $\mathbb{R}^m$. Then we make the ansatz

$$y^\ell(t) = \sum_{j=1}^{\ell} \underbrace{\langle y^\ell(t), u_j\rangle_W}_{=:\mathrm{y}_j^\ell(t)}\, u_j \quad \text{for all } t\in[0,T], \tag{2.1}$$

where the Fourier coefficients y_j^ℓ, $1\le j\le \ell$, are functions mapping $[0,T]$ into $\mathbb{R}$. Since

$$y(t) = \sum_{j=1}^{m} \langle y(t), u_j\rangle_W\, u_j \quad \text{for all } t\in[0,T]$$

holds, $y^\ell(t)$ is an approximation for $y(t)$ provided $\ell < m$. Inserting (2.1) into (1.41) yields

$$\sum_{j=1}^{\ell} \dot{\mathrm{y}}_j^\ell(t) u_j = \sum_{j=1}^{\ell} \mathrm{y}_j^\ell(t)\, Au_j + f(t, y^\ell(t)), \quad t\in(0,T], \tag{2.2a}$$

$$\sum_{j=1}^{\ell} \mathrm{y}_j^\ell(0) u_j = y_0 \tag{2.2b}$$

Note that (2.2) is an initial-value problem in $\mathbb{R}^m$ for $\ell \le m$ coefficient functions $\mathrm{y}_j^\ell(t)$, $1\le j\le\ell$ and $t\in[0,T]$, so that the coefficients are overdetermined. Therefore, we assume that (2.2) holds after projection on the ℓ dimensional subspace $V^\ell = \operatorname{span}\{u_j\}_{j=1}^{\ell}$. From (2.2a) and $\langle u_j, u_i\rangle_W = \delta_{ij}$ we infer that

$$\dot{\mathrm{y}}_i^\ell(t) = \sum_{j=1}^{\ell} \mathrm{y}_j^\ell(t)\, \langle Au_j, u_i\rangle_W + \langle f(t, y^\ell(t)), u_i\rangle_W \tag{2.3}$$

for $1\le i\le\ell$ and $t\in(0,T]$. Let us introduce the matrix

$$\mathrm{A} = \big((\mathrm{a}_{ij})\big) \in \mathbb{R}^{\ell\times\ell} \quad \text{with} \quad \mathrm{a}_{ij} = \langle Au_j, u_i\rangle_W,$$

the vector-valued mapping

$$\mathrm{y}^\ell = \begin{pmatrix} \mathrm{y}_1^\ell \\ \vdots \\ \mathrm{y}_\ell^\ell \end{pmatrix} : [0,T] \to \mathbb{R}^\ell$$

and the non-linearity $\mathrm{F} = (\mathrm{F}_1, \dots, \mathrm{F}_\ell)^T : [0,T] \times \mathbb{R}^\ell \to \mathbb{R}^\ell$ by

$$F_i(t, \mathrm{y}) = \left\langle f\left(t, \sum_{j=1}^{\ell} \mathrm{y}_j u_j\right), u_i \right\rangle_W \quad \text{for } t \in [0,T] \text{ and } \mathrm{y} = (\mathrm{y}_1, \dots, \mathrm{y}_\ell) \in \mathbb{R}^\ell.$$

Then, (2.3) can be expressed as

$$\dot{\mathrm{y}}^\ell(t) = \mathrm{A}\mathrm{y}^\ell(t) + \mathrm{F}(t, \mathrm{y}^\ell(t)) \quad \text{for } t \in (0,T] \tag{2.4a}$$

From (2.2b) we derive

$$\mathrm{y}^\ell(0) = \mathrm{y}_0, \tag{2.4b}$$

where

$$\mathrm{y}_0 = \begin{pmatrix} \langle y_0, u_1 \rangle_W \\ \vdots \\ \langle y_0, u_\ell \rangle_W \end{pmatrix} \in \mathbb{R}^\ell$$

holds. System (2.4) is called the *POD-Galerkin projection* for (1.41). In case of $\ell \ll m$ the ℓ-dimensional system (2.4) is a low-dimensional approximation for (1.41). Therefore, (2.4) is a reduced-order model for (1.41).

2. Error estimation

In this section we focus on error analysis for POD Galerkin approximations. For a more detailed presentation we refer the reader to [**21, 22, 23**] and [**15**].

Let us suppose that $y \in C([0,T]; \mathbb{R}^m) \cap C^1(0,T; \mathbb{R}^m)$ is the unique solution to (1.41) and $\{u_i\}_{i=1}^\ell$ the POD basis of rank ℓ solving

$$\min \int_0^T \left\| y(t) - \sum_{i=1}^{\ell} \langle y(t), u_i \rangle_W \, u_i \right\|_W^2 \mathrm{d}t \quad \text{s.t.} \quad \langle u_j, u_i \rangle_W = \delta_{ij}, \ 1 \le i,j \le \ell. \tag{2.5}$$

The reduced-order model for (1.41) is given by (2.4). We are interested in estimating the error

$$\int_0^T \| y(t) - y^\ell(t) \|_W^2 \, \mathrm{d}t.$$

Let us introduce the finite-dimensional space

$$V^\ell = \operatorname{span}\{u_1, \dots, u_\ell\} \subset \mathbb{R}^m$$

and the projection $\mathcal{P}^\ell : \mathbb{R}^m \to V^\ell$ by

$$\mathcal{P}^\ell u = \sum_{i=1}^{\ell} \langle u, u_i \rangle_W \, u_i \quad \text{for } u \in \mathbb{R}^m.$$

Then,

$$\begin{aligned} \mathcal{P}^\ell(\alpha u + \tilde{\alpha}\tilde{u}) &= \sum_{i=1}^{\ell} \langle \alpha u + \tilde{\alpha}\tilde{u}, u_i \rangle_W \, u_i = \sum_{i=1}^{\ell} \left(\alpha \langle u, u_i \rangle_W + \tilde{\alpha} \langle \tilde{u}, u_i \rangle_W \right) u_i \\ &= \alpha \mathcal{P}^\ell u + \tilde{\alpha} \mathcal{P}^\ell \tilde{u} \end{aligned}$$

for all α, $\tilde{\alpha} \in \mathbb{R}$ and u, $\tilde{u} \in \mathbb{R}^m$ so that $\mathcal{P}^\ell$ is linear. Further,

$$\begin{aligned} \| \mathcal{P}^\ell \|_{L(\mathbb{R}^m)}^2 &= \sup_{\|u\|_W = 1} \| \mathcal{P}^\ell u \|_W^2 = \sup_{\|u\|_W = 1} \sum_{i=1}^{\ell} \left| \langle u, u_i \rangle_W \right|^2 \\ &\le \sup_{\|u\|_W = 1} \sum_{i=1}^{m} \left| \langle u, u_i \rangle_W \right|^2 = \sup_{\|u\|_W = 1} \|u\|_W^2 = 1, \end{aligned} \tag{2.6}$$

i.e., $\mathcal{P}^\ell$ is bounded and therefore continuous. In particular, (2.6) and $\|\mathcal{P}^\ell u\|_W = \|u\|_W$ for any $u \in V^\ell$ imply $\|\mathcal{P}^\ell\|_{L(\mathbb{R}^m)} = 1$.

Throughout we shall use the decomposition

$$y(t) - y^\ell(t) = y(t) - \mathcal{P}^\ell y(t) + \mathcal{P}^\ell y(t) - y^\ell(t) = \varrho^\ell(t) + \vartheta^\ell(t), \tag{2.7}$$

where $\varrho^\ell(t) = y(t) - \mathcal{P}^\ell y(t)$ and $\vartheta^\ell(t) = \mathcal{P}^\ell y(t) - y^\ell(t)$. Note that

$$\int_0^T \Big\| y(t) - \sum_{i=1}^{\ell} \langle y(t), u_i \rangle_W \, u_i \Big\|_W^2 \, \mathrm{d}t = \int_0^T \|y(t) - \mathcal{P}^\ell y(t)\|_W^2 \, \mathrm{d}t = \int_0^T \|\varrho^\ell(t)\|_W^2 \, \mathrm{d}t.$$

Since $\{u_i\}_{i=1}^\ell$ is a POD basis of rank ℓ we have

$$\int_0^T \|\varrho^\ell(t)\|_W^2 \, \mathrm{d}t = \sum_{i=\ell+1}^{m} \lambda_i. \tag{2.8}$$

Next we estimate the term $\vartheta^\ell(t)$. Utilizing (1.41a) and (2.4) we obtain for every $u^\ell \in V^\ell$ and $t \in (0, T]$

$$\begin{aligned} \langle \dot\vartheta^\ell(t), u^\ell \rangle_W &= \langle \mathcal{P}^\ell \dot y(t) - \dot y(t), u^\ell \rangle_W + \langle \dot y(t) - \dot y^\ell(t), u^\ell \rangle_W \\ &= \langle \mathcal{P}^\ell \dot y(t) - \dot y(t), u^\ell \rangle_W \\ &\quad + \langle A(y(t) - y^\ell(t)) + f(t, y(t)) - f(t, y^\ell(t)), u^\ell \rangle_W \end{aligned} \tag{2.9}$$

We choose $u^\ell = \vartheta^\ell(t) \in V^\ell$. Let

$$\|A\| = \max_{\|u\|_W = 1} \|Au\|_W$$

the matrix norm induced by the vector norm $\|\cdot\|_W$. Further,

$$\frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \|\vartheta^\ell(t)\|_W^2 = \langle \dot\vartheta^\ell(t), \vartheta^\ell(t) \rangle_W \quad \text{for every } t \in (0, T].$$

holds. Then, we infer from (2.9)

$$\begin{aligned} \frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \|\vartheta^\ell(t)\|_W^2 &\le \|A\| \big(\|\varrho^\ell(t)\|_W + \|\vartheta^\ell(t)\|_W \big) \|\vartheta^\ell(t)\|_W \\ &\quad + \|f(t, y(t)) - f(t, y^\ell(t))\|_W \|\vartheta^\ell(t)\|_W \\ &\quad + \|\mathcal{P}^\ell \dot y(t) - \dot y(t)\|_W \|\vartheta^\ell(t)\|_W. \end{aligned} \tag{2.10}$$

Suppose that f is Lipschitz-continuous with respect to the second argument, i.e., there exists a constant $L_f \ge 0$ satisfying

$$\|f(t, u) - f(t, \tilde u)\|_W \le L_f \, \|u - \tilde u\|_W \quad \text{for all } u, \, \tilde u \in \mathbb{R}^m \text{ and } t \in [0, T].$$

Moreover, we have

$$\|\mathcal{P}^\ell \dot y(t) - \dot y(t)\|_W^2 = \Big\| \sum_{i=\ell+1}^{m} \langle \dot y(t), u_i \rangle_W \, u_i \Big\|_W^2 = \sum_{i=\ell+1}^{m} \big| \langle \dot y(t), u_i \rangle_W \big|^2$$

for all $t \in (0,T)$. Consequently, (2.10) and (2.7) imply

$$\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\vartheta^\ell(t)\right\|_W^2 &\leq \frac{\|A\|}{2}\left(\left\|\varrho^\ell(t)\right\|_W^2 + \left\|\vartheta^\ell(t)\right\|_W^2\right) + \|A\|\left\|\vartheta^\ell(t)\right\|_W^2 \\
&\quad + L_f\left\|\varrho^\ell(t) + \vartheta^\ell(t)\right\|_W\left\|\vartheta^\ell(t)\right\|_W \\
&\quad + \frac{1}{2}\left(\left\|\mathcal{P}^\ell\dot{y}(t) - \dot{y}(t)\right\|_W^2 + \left\|\vartheta^\ell(t)\right\|_W^2\right) \\
&\leq \frac{\|A\|}{2}\left\|\varrho^\ell(t)\right\|_W^2 + \left(\frac{1}{2}\|A\| + \frac{1}{2} + L_f\right)\left\|\vartheta^\ell(t)\right\|_W^2 \\
&\quad + L_f\left\|\varrho^\ell(t)\right\|_W\left\|\vartheta^\ell(t)\right\|_W + \sum_{i=\ell+1}^m \left|\left\langle\dot{y}(t), u_i\right\rangle_W\right|^2 \\
&\leq \frac{\|A\| + L_f}{2}\left\|\varrho^\ell(t)\right\|_W^2 + \left(\frac{3}{2}\left(\|A\| + L_f\right) + \frac{1}{2}\right)\left\|\vartheta^\ell(t)\right\|_W^2 \\
&\quad + \sum_{i=\ell+1}^m \left|\left\langle\dot{y}(t), u_i\right\rangle_W\right|^2.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left\|\vartheta^\ell(t)\right\|_W^2 &\leq \Big(3\big(\|A\| + L_f\big) + 1\Big)\left\|\vartheta^\ell(t)\right\|_W^2 + \big(\|A\| + L_f\big)\left\|\varrho^\ell(t)\right\|_W^2 \\
&\quad + \sum_{i=\ell+1}^m \left|\left\langle\dot{y}(t), u_i\right\rangle_W\right|^2.
\end{aligned}$$

Using Gronwall's lemma (see Exercise 2.1)) and (2.8) we arrive at

$$\begin{aligned}
\left\|\vartheta^\ell(t)\right\|_W^2 &\leq c_1\left(\left\|\vartheta^\ell(0)\right\|_W^2 + \big(\|A\| + L_f\big)\int_0^t \left\|\varrho^\ell(s)\right\|_W^2 \mathrm{d}s\right) \\
&\quad + c_1\sum_{i=\ell+1}^m \int_0^t \left|\left\langle\dot{y}(s), u_i\right\rangle_W\right|^2 \mathrm{d}s \\
&\leq c_2\left(\left\|\vartheta^\ell(0)\right\|_W^2 + \sum_{i=\ell+1}^m \left(\lambda_i + \int_0^T \left|\left\langle\dot{y}(t), u_i\right\rangle_W\right|^2 \mathrm{d}t\right)\right)
\end{aligned} \tag{2.11}$$

where $c_1 = \exp(3(\|A\| + L_f) + 1)T)$ and $c_2 = c_1\max\{\|A\| + L_f, 1\}$.

THEOREM 2.1. *Let $y \in C([0,T];\mathbb{R}^m) \cap C^1(0,T;\mathbb{R}^m)$ be the unique solution to (1.41), $\ell \in \{1,\dots,m\}$ be fixed and $\{u_i\}_{i=1}^\ell$ a POD basis of rank ℓ solving (2.5). Let y^ℓ be the unique solution to the reduced-order model (2.4). Then*

$$\int_0^T \left\|y(t) - y^\ell(t)\right\|_W^2 \mathrm{d}t \leq C\sum_{i=\ell+1}^m \left(\lambda_i + \int_0^T \left|\left\langle\dot{y}(t), u_i\right\rangle_W\right|^2 \mathrm{d}t\right)$$

for a constant $C > 0$.

PROOF. From (2.8), (2.11) and $\vartheta^\ell(0) = \mathcal{P}^\ell y_0 - y^\ell(0) = 0$ we find

$$\begin{aligned}
\int_0^T \left\|y(t) - y^\ell(t)\right\|_W^2 \mathrm{d}t &= \int_0^T \left\|\varrho^\ell(t) + \vartheta^\ell(t)\right\|_W^2 \mathrm{d}t \\
&\leq 2\int_0^T \left\|\varrho^\ell(t)\right\|_W^2 + \left\|\vartheta^\ell(t)\right\|_W^2 \mathrm{d}t \\
&\leq 2\sum_{i=\ell+1}^m \lambda_i + c_3\sum_{i=\ell+1}^m \left(\lambda_i + \int_0^T \left|\left\langle\dot{y}(t), u_i\right\rangle_W\right|^2 \mathrm{d}t\right)
\end{aligned}$$

with $c_3 = 2c_2$. Setting $C = 2 + c_3$ the claim follows directly. □

REMARK 2.2. The term

$$\sum_{i=\ell+1}^{m} \int_0^T \left| \langle \dot{y}(t), u_i \rangle_W \right|^2 \mathrm{d}t$$

can not be estimated by the sum over the eigenvalues $\lambda_{\ell+1}, \ldots, \lambda_m$. If we replace (2.5) by

$$\min \int_0^T \left\| y(t) - \sum_{i=1}^{\ell} \langle y(t), u_i \rangle_W \, u_i \right\|_W^2 + \left\| \dot{y}(t) - \sum_{i=1}^{\ell} \langle \dot{y}(t), u_i \rangle_W \, u_i \right\|_W^2 \mathrm{d}t \tag{2.12a}$$

subject to

$$\langle u_j, u_i \rangle_W = \delta_{ij} \quad \text{for } 1 \le i, j \le \ell, \tag{2.12b}$$

we end up with the estimate

$$\int_0^T \| y(t) - y^\ell(t) \|_W^2 \, \mathrm{d}t \le \tilde{C} \sum_{i=\ell+1}^{m} \tilde{\lambda}_i$$

for a constant $\tilde{C} > 0$. In this case the time derivatives are also included in the snapshot ensemble. Of course, the operator $\mathcal{R}$ defined in (1.52) has to be replaced. It turns out that the POD basis $\{u_i\}_{i=1}^{\ell}$ is given by the eigenvalue problem

$$\tilde{\mathcal{R}} \tilde{u}_i = \tilde{\lambda}_i \tilde{u}_i \text{ for } 1 \le i \le m \quad \text{and} \quad \tilde{\lambda}_1 \ge \tilde{\lambda}_2 \ge \ldots \ge \tilde{\lambda}_m \ge 0$$

where the operator $\tilde{\mathcal{R}} : \mathbb{R}^m \to \mathbb{R}^m$ is defined by

$$\tilde{\mathcal{R}} u = \int_0^T \langle y(t), u \rangle_W \, y(t) + \langle \dot{y}(t), u \rangle_W \, \dot{y}(t) \, \mathrm{d}t$$

for $u \in \mathbb{R}^m$. ◊

From a practical point of view we do not have the information on the whole trajectory in $[0, T]$. Therefore, let $\Delta t = T/(n-1)$ be a fixed time step size and $t_j = (j-1)\Delta t$ for $1 \le j \le n$ a given time grid in $[0, T]$. To simplify the presentation we choose an equidistant grid. Of course, non-equidistant meshes can be treated analogously [**22**]. We compute a POD basis $\{u_i^n\}_{i=1}^{\ell}$ of rank ℓ by solving the constrained minimization problem ($\hat{\mathbf{P}}_W^{n,\ell}$). After the POD basis has been determined, we derive the reduced-order model as described in Section 2.1. Thus,

$$y^\ell(t) = \sum_{i=1}^{\ell} \mathrm{y}_j^\ell(t) u_i^n, \quad t \in [0, T],$$

solves the POD Galerkin projection of (1.41)

$$\langle \dot{y}^\ell(t), u_i^n \rangle_W = \langle A y^\ell(t) + f(t, y^\ell(t)), u_i^n \rangle_W \qquad \text{for } i = 1 \ldots, \ell \text{ and } t \in (0, T], \tag{2.13a}$$

$$\langle y^\ell(0), u_i^n \rangle_W = \langle y_0, u_i^n \rangle_W \qquad \text{for } i = 1 \ldots, \ell. \tag{2.13b}$$

To solve (2.13) we apply the implicit Euler method. By Y_j we denote an approximation for y^ℓ at the time t_j, $1 \le j \le n$. Then, the discrete system for the sequence $\{Y_j\}_{j=1}^n$ in $V_n^\ell = \operatorname{span}\{u_1^n, \ldots, u_\ell^n\}$ looks like

$$\left\langle \frac{Y_j - Y_{j-1}}{\Delta t}, u_i^n \right\rangle_W = \langle A Y_j + f(t, Y_j), u_i^n \rangle_W \qquad \text{for } i = 1 \ldots, \ell, \ 2 \le j \le n, \tag{2.14a}$$

$$\langle Y_1, u_i^n \rangle_W = \langle y_0, u_i^n \rangle_W \qquad \text{for } i = 1 \ldots, \ell. \tag{2.14b}$$

We are interested in estimating

$$\sum_{j=1}^{n} \alpha_j \| y(t_j) - Y_j \|_W^2.$$

Let us introduce the projection $\mathcal{P}_n^\ell : \mathbb{R}^m \to V_n^\ell$ by

$$\mathcal{P}_n^\ell = \sum_{i=1}^{\ell} \langle u, u_i^n \rangle_W \, u_i^n \quad \text{for } u \in \mathbb{R}^m. \tag{2.15}$$

It follows that $\mathcal{P}_n^\ell$ is linear and bounded (and therefore continuous). In particular, $\|\mathcal{P}_n^\ell\|_{L(\mathbb{R}^m)} = 1$.

We shall make use of the decomposition

$$y(t_j) - Y_j = y(t_j) - \mathcal{P}_n^\ell y(t_j) + \mathcal{P}_n^\ell y(t_j) - Y_j = \varrho_j^\ell + \vartheta_j^\ell,$$

where $\varrho_j^\ell = y(t_j) - \mathcal{P}_n^\ell y(t_j)$ and $\vartheta_j^\ell = \mathcal{P}_n^\ell y(t_j) - Y_j$. Note that

$$\sum_{j=1}^{n} \alpha_j \left\| y(t_j) - \sum_{i=1}^{\ell} \langle y(t_j), u_i^n \rangle_W \, u_i^n \right\|_W^2 = \sum_{j=1}^{n} \alpha_j \, \|y(t_j) - \mathcal{P}_n^\ell y(t_j)\|_W^2 = \sum_{j=1}^{n} \alpha_j \, \|\varrho_j^\ell\|_W^2.$$

Since $\{u_i^n\}_{i=1}^\ell$ is the POD basis of rank ℓ, we have

$$\sum_{j=1}^{n} \alpha_j \, \|\varrho_j^\ell\|_W^2 = \sum_{i=\ell+1}^{m} \lambda_i^n. \tag{2.16}$$

Next we estimate the terms ϑ_j^ℓ. Using the notation $\overline{\partial}\vartheta_j^\ell = (\vartheta_j^\ell - \vartheta_{j-1}^\ell)/\Delta t$ for $2 \le j \le n$ we obtain by (1.41a) and (2.14a)

$$\begin{aligned}
\langle \overline{\partial}\vartheta_j^\ell, u_i \rangle &= \left\langle \mathcal{P}_n^\ell \left(\frac{y(t_j) - y(t_{j-1})}{\Delta t} \right) - \frac{Y_j - Y_{j-1}}{\Delta t}, u_i^n \right\rangle_W \\
&= \langle \dot{y}(t_j) - (AY_j + f(t_j, Y_j))), u_i^n \rangle_W \\
&\quad + \left\langle \mathcal{P}_n^\ell \left(\frac{y(t_j) - y(t_{j-1})}{\Delta t} \right) - \dot{y}(t_j), u_i^n \right\rangle_W \\
&= \langle A(y(t_j) - Y_j) + f(t_j, y(t_j)) - f(t_j, Y_j), u_i^n \rangle_W \\
&\quad + \left\langle \mathcal{P}_n^\ell \left(\frac{y(t_j) - y(t_{j-1})}{\Delta t} \right) - \frac{y(t_j) - y(t_{j-1})}{\Delta t}, u_i^n \right\rangle_W \\
&\quad + \left\langle \frac{y(t_j) - y(t_{j-1})}{\Delta t} - \dot{y}(t_j), u_i^n \right\rangle_W \\
&= \langle A(y(t_j) - Y_j) + f(t_j, y(t_j)) - f(t_j, Y_j) + z_j^\ell + w_j^\ell, u_i^n \rangle_W
\end{aligned} \tag{2.17}$$

for $1 \le i \le \ell$ and $2 \le j \le n$, where

$$z_j^\ell = \mathcal{P}_n^\ell \left(\frac{y(t_j) - y(t_{j-1})}{\Delta t} \right) - \frac{y(t_j) - y(t_{j-1})}{\Delta t}, \quad w_j^\ell = \frac{y(t_j) - y(t_{j-1})}{\Delta t} - \dot{y}(t_j).$$

Multiplying (2.17) by $\langle \vartheta_j^\ell, u_i^n \rangle_W$ and adding all ℓ equations we arrive at

$$\langle \overline{\partial}\vartheta_j^\ell, \vartheta_j^\ell \rangle = \langle A(y(t_j) - Y_j) + f(t_j, y(t_j)) - f(t_j, Y_j) + z_j^\ell + w_j^\ell, \vartheta_j^\ell \rangle_W \tag{2.18}$$

for $j = 2, \dots, n$. Note that

$$\begin{aligned}
2 \langle u - \tilde{u}, u \rangle_W &= 2 \|u\|_W^2 - 2 \langle \tilde{u}, u \rangle_W \\
&= \|u\|_W^2 + \|u\|_W^2 - 2 \langle \tilde{u}, u \rangle_W + \|\tilde{u}\|_W^2 - \|\tilde{u}\|_W^2 \\
&= \|u\|_W^2 - \|\tilde{u}\|_W^2 + \|u - \tilde{u}\|_W^2
\end{aligned}$$

for all $u, \tilde{u} \in \mathbb{R}^m$. Choosing $u = \vartheta_j^\ell$ and $\tilde{u} = \vartheta_{j-1}^\ell$ we infer from (2.18)

$$2 \langle \overline{\partial}\vartheta_j^\ell, \vartheta_j^\ell \rangle = \frac{1}{\Delta t} \left(\|\vartheta_j^\ell\|_W^2 - \|\vartheta_{j-1}^\ell\|_W^2 + \|\vartheta_j^\ell - \vartheta_{j-1}^\ell\|_W^2 \right). \tag{2.19}$$

Inserting (2.19) into (2.18) and using the Cauchy-Schwarz inequality we obtain

$$\begin{aligned}\|\vartheta_j^\ell\|_W^2 \leq \|\vartheta_{j-1}^\ell\|_W^2 + \Delta t\,\|A\|\big(\|\varrho_j^\ell\|_W + \|\vartheta_j^\ell\|_W\big)\,\|\vartheta_j^\ell\|_W \\ + \Delta t\,\Big(\|f(t_j,y(t_j)) - f(t_j,Y_j)\|_W + \|z_j^\ell\|_W + \|w_j^\ell\|_W\Big)\|\vartheta_j^\ell\|_W.\end{aligned}$$

Suppose that f is Lipschitz-continuous with respect to the second argument. Then there exists a constant $L_f \geq 0$ such that

$$\|f(t_j,y(t_j)) - f(t_j,Y_j)\|_W \leq L_f\,\|y(t_j) - Y_j\|_W \quad \text{for } j = 2,\dots,n.$$

Hence, by Young's inequality we find

$$\|\vartheta_j^\ell\|_W^2 \leq \|\vartheta_{j-1}^\ell\|_W^2 + \Delta t\Big(c_1\,\|\varrho_j^\ell\|_W^2 + c_2\,\|\vartheta_j^\ell\|_W^2 + \|z_j^\ell\|_W^2 + \|w_j^\ell\|_W^2\Big),$$

where $c_1 = \max\{\|A\|, L_f\}$ and $c_2 = \max\{3\,\|A\|, 3L_f, 2\}$. Suppose that

$$0 < \Delta t \leq \frac{1}{2c_2} \tag{2.20}$$

holds. With (2.20) holding we have

$$0 \leq 1 - 2c_2\Delta t < 1 - c_2\Delta t \quad \text{and} \quad 1 - c_2\Delta t \geq 1 - \frac{1}{2} = \frac{1}{2}.$$

Thus,

$$\frac{1}{1 - c_2\Delta t} = \frac{1 - c_2\Delta t + c_2\Delta t}{1 - c_2\Delta t} = 1 + \frac{c_2\Delta t}{1 - c_2\Delta t}. \leq 1 + 2c_2\Delta t \tag{2.21}$$

Using (2.21) we infer that

$$\|\vartheta_j^\ell\|_W^2 \leq (1 + 2c_2\Delta t)\Big(\|\vartheta_{j-1}^\ell\|_W^2 + \Delta t\big(\|z_j^\ell\|_W^2 + \|w_j^\ell\|_W^2 + c_1\,\|\varrho_j^\ell\|_W^2\big)\Big).$$

Summation on j yields

$$\|\vartheta_j^\ell\|_W^2 \leq (1 + 2c_2\Delta t)^j\left(\|\vartheta_0^\ell\|_W^2 + \Delta t\sum_{k=1}^{j}\left(\|z_k^\ell\|_W^2 + \|w_k^\ell\|_W^2 + c_1\,\|\varrho_k^\ell\|_W^2\right)\right).$$

Note that

$$(1 + 2c_2\Delta t)^j = \left(1 + \frac{2c_2 j\Delta t}{j}\right)^j \leq e^{2c_2 j\Delta t}.$$

Thus,

$$\|\vartheta_j^\ell\|_W^2 \leq e^{2c_2 j\Delta t}\left(\|\vartheta_0^\ell\|_W^2 + \Delta t\sum_{k=1}^{j}\left(\|z_k^\ell\|_W^2 + \|w_k^\ell\|_W^2 + c_1\,\|\varrho_k^\ell\|_W^2\right)\right).$$

We next estimate the term involving w_k^ℓ:

$$\begin{aligned}
\Delta t \sum_{k=1}^{j} \|w_k^\ell\|_W^2 &= \Delta t \sum_{k=1}^{j} \left\| \frac{y(t_k) - y(t_{k-1})}{\Delta t} - \dot{y}(t_k) \right\|_W^2 \\
&= \frac{1}{\Delta t} \sum_{k=1}^{j} \|y(t_k) - y(t_{k-1}) - \Delta t \dot{y}(t_k)\|_W^2 \\
&= \frac{1}{\Delta t} \sum_{k=1}^{j} \left\| \int_{t_{k-1}}^{t_k} (t_{k-1} - s) \ddot{y}(s) \, \mathrm{d}s \right\|_W^2 \\
&\leq \frac{1}{\Delta t} \sum_{k=1}^{j} \int_{t_{k-1}}^{t_k} |t_{k-1} - s|^2 \, \mathrm{d}s \int_{t_{k-1}}^{t_k} \|\ddot{y}(s)\|_W^2 \, \mathrm{d}s \\
&\leq \frac{(\Delta t)^2}{3} \sum_{k=1}^{j} \|\ddot{y}\|_{L^2(t_{k-1}, t_k; \mathbb{R}^m)}^2 = \frac{(\Delta t)^2}{3} \|\ddot{y}\|_{L^2(0, t_j; \mathbb{R}^m)}^2.
\end{aligned}$$

The term z_k^ℓ can be estimated as follows:

$$\begin{aligned}
\|z_k^\ell\|_W^2 &= \left\| \mathcal{P}_n^\ell \Big(\frac{y(t_k) - y(t_{k-1})}{\Delta t} \Big) - \frac{y(t_k) - y(t_{k-1})}{\Delta t} \right\|_W^2 \\
&= \left\| \mathcal{P}_n^\ell \Big(\frac{y(t_k) - y(t_{k-1})}{\Delta t} \Big) - \mathcal{P}_n^\ell \dot{y}(t_k) + \mathcal{P}_n^\ell \dot{y}(t_k) - \frac{y(t_k) - y(t_{k-1})}{\Delta t} \right\|_W^2 \\
&\leq 2 \|\mathcal{P}_n^\ell\|_{L(\mathbb{R}^m)}^2 \left\| \frac{y(t_k) - y(t_{k-1})}{\Delta t} - \dot{y}(t_k) \right\|_W^2 \\
&\quad + 2 \left\| \mathcal{P}_n^\ell \dot{y}(t_k) - \dot{y}(t_k) + \dot{y}(t_k) - \frac{y(t_k) - y(t_{k-1})}{\Delta t} \right\|_W^2 \\
&\leq 2 \|w_k^\ell\|_W^2 + 4 \|\mathcal{P}_n^\ell \dot{y}(t_k) - \dot{y}(t_k)\|_W^2 + 4 \left\| \dot{y}(t_k) - \frac{y(t_k) - y(t_{k-1})}{\Delta t} \right\|_W^2 \\
&= 4 \|\mathcal{P}_n^\ell \dot{y}(t_k) - \dot{y}(t_k)\|_W^2 + 6 \|w_k^\ell\|_W^2.
\end{aligned}$$

Recall that $\Delta t \leq 2\alpha_k$ for $1 \leq k \leq n$. Hence,

$$\Delta t \sum_{k=1}^{j} \|z_k^\ell\|_W^2 \leq 8 \sum_{k=1}^{n} \alpha_k \|\mathcal{P}_n^\ell \dot{y}(t_k) - \dot{y}(t_k)\|_W^2 + 2(\Delta t)^2 \|\ddot{y}\|_{L^2(0, t_j; \mathbb{R}^m)}^2.$$

Further, $\vartheta_0^\ell = \mathcal{P}_n^\ell y_0 - Y_1 = 0$ and $0 \leq j\Delta t \leq T$ for $j = 0, \ldots, n-1$. Summarizing

$$\|\vartheta_j^\ell\|_W^2 \leq c_3 \bigg(\sum_{k=1}^{n} 8\alpha_k \left(\|\mathcal{P}_n^\ell \dot{y}(t_k) - \dot{y}(t_k)\|_W^2 + 2c_1 \|\varrho_k^\ell\|_W^2 \right) + \frac{7}{3} (\Delta t)^2 \|\ddot{y}\|_{L^2(0, t_j; \mathbb{R}^m)}^2 \bigg),$$

where $c_3 = e^{2c_2 T} \max\{7/3, 2c_1, 8\}$ is independent of ℓ and $\{t_j\}_{j=1}^n$. From $\sum_{k=1}^n \alpha_k = T$ and (2.16) we infer

$$\begin{aligned}
\sum_{j=1}^{n} \alpha_j \|\vartheta_j^\ell\|_W^2 &\leq c_3 T \bigg(\sum_{j=1}^{n} \alpha_j \left(\|\mathcal{P}_n^\ell \dot{y}(t_j) - \dot{y}(t_j)\|_W^2 + \|\varrho_j^\ell\|_W^2 \right) \\
&\qquad + (\Delta t)^2 \|\ddot{y}\|_{L^2(0,T;\mathbb{R}^m)}^2 \bigg) \\
&\leq c_4 \bigg(\sum_{i=\ell+1}^{m} \left(\lambda_i^n + \sum_{j=1}^{n} \alpha_j \left| \langle \dot{y}(t_j), u_i^n \rangle_W \right|^2 \right) + (\Delta t)^2 \bigg)
\end{aligned} \tag{2.22}$$

with $c_4 = c_3 T \max\{1, \|\ddot{y}\|_{L^2(0,T;\mathbb{R}^m)}^2\}$.

THEOREM 2.3. *Let $y \in C([0,T];\mathbb{R}^m) \cap C^1(0,T;\mathbb{R}^m)$ be the unique solution to (1.41) satisfying $\ddot{y} \in L^2(0,T;\mathbb{R}^m)$ and $\ell \in \{1,\dots,m\}$ be fixed. Suppose that $\{u_i^n\}_{i=1}^{\ell}$ is a POD basis of rank ℓ solving $(\hat{\mathbf{P}}_W^{n,\ell})$. Assume that (2.14) possesses a unique solution $\{Y_j\}_{j=1}^n$. Then there exists a constant $C > 0$ such that*

$$\sum_{j=1}^{n} \alpha_j \left\| y(t_j) - Y_j \right\|_W^2 \leq C \left((\Delta t)^2 + \sum_{i=\ell+1}^{m} \left(\lambda_i^n + \sum_{j=1}^{n} \alpha_j \left| \langle \dot{y}(t_j), u_i^n \rangle_W \right|^2 \right) \right)$$

provided Δt is sufficiently small and f is Lipschitz-continuous with respect to the second argument.

PROOF. The claim follows directly from (2.16), (2.22), and

$$\begin{aligned} \sum_{j=1}^{n} \alpha_j \left\| y(t_j) - Y_j \right\|_W^2 &\leq 2 \sum_{j=1}^{n} \alpha_j \left(\left\| \vartheta_j^{\ell} \right\|_W^2 + \left\| \varrho_j^{\ell} \right\|_W^2 \right) \\ &\leq 2c_4 \left(\sum_{i=\ell+1}^{m} \left(\lambda_i^n + \sum_{j=1}^{n} \left| \langle \dot{y}(t_j), u_i^n \rangle_W \right|^2 \right) + (\Delta t)^2 \right) \\ &\quad + 2 \sum_{i=\ell+1}^{m} \lambda_i^n \end{aligned}$$

provided Δt is sufficiently small and f is Lipschitz-continuous with respect to the second argument. □

REMARK 2.4. Compared to the estimate in Theorem 2.1 we observe the term

$$\sum_{j=1}^{n} \alpha_j \left| \langle \dot{y}(t_j), u_i^n \rangle_W \right|^2 \tag{2.23}$$

instead of the term

$$\int_0^T \left| \langle \dot{y}(t), u_i \rangle_W \right|^2 \mathrm{d}t. \tag{2.24}$$

Note that (2.23) is the trapezoidal approximation of (2.24). Furthermore, the error $O((\Delta t)^2)$ appears in the estimate of Theorem 2.3 due to the Euler method. ◊

Next we address the fact that the eigenvalues $\{\lambda_i^n\}_{i=1}^m$ and the associated eigenvectors $\{u_i^n\}$ (i.e., the POD basis) depend on the chosen time grid $\{t_j\}_{j=1}^n$. We apply the asymptotic theory presented in Section 1.3. Then, it follows from Theorem 2.14 that there exists a number $\bar{n} \in \mathbb{N}$ satisfying

$$\begin{aligned} \sum_{i=\ell+1}^{m} \lambda_i^n &\leq 2 \sum_{i=\ell+1}^{m} \lambda_i, \\ \sum_{i=\ell+1}^{m} \sum_{j=1}^{n} \alpha_j \left| \langle \dot{y}(t_j), u_i^n \rangle_W \right|^2 &\leq 2 \sum_{i=\ell+1}^{m} \int_0^T \left| \langle \dot{y}(t), u_i \rangle_W \right|^2 \mathrm{d}t \end{aligned}$$

for $n \geq \bar{n}$ provided $\sum_{i=\ell+1}^m \lambda_i \neq 0$ and $\int_0^T \left| \langle \dot{y}(t), u_i \rangle_W \right|^2 \mathrm{d}t \neq 0$ hold. Thus, we infer from Theorems 2.1 and 2.3 the following result.

THEOREM 2.5. *Let all hypothesis of Theorems 2.14, 2.1 and 2.3 be satisfied. If $\int_0^T \left| \langle \dot{y}(t), u_i \rangle_W \right|^2 \mathrm{d}t \neq 0$, then there exists a constant $C > 0$ and a number $\bar{n} \in \mathbb{N}$ such that*

$$\sum_{j=1}^{n} \alpha_j \left\| y(t_j) - Y_j \right\|_W^2 \leq C \left((\Delta t)^2 + \sum_{i=\ell+1}^{m} \left(\lambda_i + \int_0^T \left| \langle \dot{y}(t), u_i \rangle \right|^2 \mathrm{d}t \right) \right)$$

for all $n \geq \bar{n}$.

3. ROM for evolution problems

This section is devoted to error estimates for the Galerkin POD method applied to (1.70) combined with the backward Euler method for the time integration. For more details and the proofs we refer the reader to [**21, 22, 23**].

3.1. Case $X = V$. Let us choose $X = V$ in the context of Chapter 2, Section 2.4. To study the backward Euler POD Galerkin method for (1.70), we introduce the Ritz projection $\mathcal{P}^\ell : V \to V^\ell$, $1 \le \ell \le d$, by

$$a(\mathcal{P}^\ell \varphi, \psi) = a(\varphi, \psi) \quad \text{for all } \psi \in V^\ell, \tag{2.25}$$

where $\varphi \in V$. Since the Hilbert space V is endowed with the inner product (1.68), $\mathcal{P}^\ell$ is the orthogonal projection of V on V^ℓ. In particular, this implies that $\mathcal{P}^\ell$ has norm one.

The POD Galerkin method for (1.70) is described next. For $m \in \mathbb{N}$ we introduce the time grid

$$0 = \tau_0 < \tau_1 < \ldots < \tau_m = T, \quad \delta\tau_j = \tau_j - \tau_{j-1} \text{ for } j = 1, \ldots, m,$$

and set

$$\delta\tau = \min\{\delta\tau_j : 1 \le j \le m\} \quad \text{and} \quad \Delta\tau = \max\{\delta\tau_j : 1 \le j \le m\}.$$

Throughout we assume that $\Delta\tau/\delta\tau$ is bounded uniformly with respect to m. To relate the two time discretizations $\{t_j\}_{j=0}^n$ and $\{\tau_j\}_{j=0}^m$ we set for every τ_k, $0 \le k \le m$, an associated index $\bar{k} = \operatorname{argmin}\{|\tau_k - t_j| : 0 \le j \le n\}$ and define $\sigma_n \in \{1, \ldots, n\}$ as the maximum of the occurrence of the same value $t_{\bar{k}}$ as k ranges over $0 \le k \le m$.

The problem consists in finding a sequence $\{Y_k\}_{k=0}^m$ in V^ℓ satisfying

$$\langle Y_0, \psi \rangle_H = \langle y_0, \psi \rangle_H \quad \text{for all } \psi \in V^\ell \tag{2.26a}$$

and

$$\langle \overline{\partial}_\tau Y_k, \psi \rangle_H + a(Y_k, \psi) + \langle B(Y_k) + RY_k, \psi \rangle_{V',V} = \langle f(\tau_k), \psi \rangle_H \tag{2.26b}$$

for all $\psi \in V^\ell$ and $k = 1, \ldots, m$, where we have set

$$\overline{\partial}_\tau Y_k = \frac{Y_k - Y_{k-1}}{\delta\tau_k}.$$

For every $k = 1, \ldots, m$ there exists at least one solution Y_k of (2.26b). If $\Delta\tau$ is sufficiently small, the sequence $\{Y_k\}_{k=1}^m$ is uniquely determined [**22**].

Our next goal is to present an error estimate for the expression

$$\sum_{k=0}^m \beta_k \, \|Y_k - y(\tau_k)\|_H^2,$$

where $y(\tau_k)$ is the solution of (1.70) at the time instances $t = \tau_k$, $k = 1, \ldots, m$, and the positive weights β_j are given by

$$\beta_0 = \frac{\delta\tau_1}{2}, \quad \beta_j = \frac{\delta\tau_j + \delta\tau_{j+1}}{2} \text{ for } j = 1, \ldots, m-1, \quad \text{and} \quad \beta_m = \frac{\delta\tau_m}{2}.$$

We make use of the following assumptions:

- (A1) $y_t \in L^2(0,T;V)$ *and* $y_{tt} \in L^2(0,T;H)$.
- (A2) *There exists a normed linear space* W *continuously embedded in* V *and a constant* $c_a > 0$ *such that* $y \in C([0,T];W)$ *and*

$$a(\varphi, \psi) \le c_a \, \|\varphi\|_H \|\psi\|_W \quad \textit{for all } \varphi \in V \textit{ and } \psi \in W. \tag{2.27}$$

- (A3) $y \in W^{2,2}(0,T;V)$.

EXAMPLE 3.1. For $V = H_0^1(\Omega)$, $H = L^2(\Omega)$, with Ω a bounded domain in $\mathbb{R}^l$ and

$$a(\varphi, \psi) = \int_\Omega \nabla\varphi \cdot \nabla\psi \, dx \text{ for all } \varphi, \psi \in H_0^1(\Omega),$$

choosing $W = H^2(\Omega) \cap H_0^1(\Omega)$ implies $a(\varphi, \psi) \le \|\varphi\|_W \|\psi\|_H$ for all $\varphi \in W$, $\psi \in V$, and (2.27) holds with $c_a = 1$. $\diamondsuit$

REMARK 3.2. Note that (A2) implies the existence of a constant $c_P > 0$ depending on ℓ and λ_ℓ such that

$$\|P^\ell\|_{\mathcal{L}(H)} \le c_P \quad \text{for all } 1 \le \ell \le d, \tag{2.28}$$

see [**22**]. $\diamond$

In [**22**] the following result is proved.

THEOREM 3.3. a) *Assume that* (A1), (A2) *hold and that* $\Delta\tau$ *is sufficiently small. Then there exists a constant* C *depending on* T*, but independent of the grids* $\{t_j\}_{j=0}^n$ *and* $\{\tau_j\}_{j=0}^m$*, such that*

$$\begin{aligned} &\sum_{k=0}^m \beta_k \, \|Y_k - y(\tau_k)\|_H^2 \\ &\le C \sum_{i=\ell+1}^d \Big(\big|\langle \psi_i, y_0 \rangle_V\big|^2 + \frac{\sigma_n}{\delta t}\Big(\frac{1}{\delta\tau} + \Delta\tau\Big)\lambda_i\Big) + C\sigma_n \Delta\tau \Delta t \, \|y_t\|^2_{L^2(0,T;V)} \\ &\quad + C\sigma_n(1 + c_P^2)\Delta\tau \left(\Delta t \|y_t\|^2_{L^2(0,T;H)} + (\Delta\tau + \Delta t)\|y_{tt}\|^2_{L^2(0,T;H)} \right). \end{aligned} \tag{2.29}$$

b) *If* (A3) *is satisfied and* $\Delta\tau$ *sufficiently small, then there exists a constant* C *depending on* T*, but independent of the grids* $\{t_j\}_{j=0}^n$ *and* $\{\tau_j\}_{j=0}^m$*, such that*

$$\begin{aligned} &\sum_{k=0}^m \beta_k \, \|Y_k - y(\tau_k)\|_H^2 \le C\sigma_n \Delta\tau(\Delta\tau + \Delta t)\|y_{tt}\|^2_{L^2(0,T;V)} \\ &+ C\Big(\sum_{i=\ell+1}^d \Big(\big|\langle \psi_i, y_0 \rangle_V\big|^2 + \frac{\sigma_n}{\delta t}\Big(\frac{1}{\delta\tau} + \Delta\tau\Big)\lambda_i\Big) + \sigma_n \Delta\tau \Delta t \|y_t\|^2_{L^2(0,T;V)}\Big). \end{aligned} \tag{2.30}$$

Compared to standard finite difference, finite element or spectral element approximation results the basic POD-Galerkin backward Euler convergence result of Theorem 3.3 has an unusual format. This is due, in part, to the fact that one can not rely on function space rate of convergence results, which are typically at the basis for approximation theory of partial differential equations. The terms in the second line of (2.29) depend (through ψ_i, λ_i, d) on the way in which the snapshots are taken, on the number ℓ of basis elements and on the relative location of the snapshots and the time discretization (through σ_n).

REMARK 3.4. In (2.29) and (2.30) the eigenvalues and eigenfunctions depend on n, i.e., $\lambda_i = \lambda_i^n$ and $\psi_i = \psi_i^n$. If ℓ satisfies (1.76) and $\sum_{i=\ell+1}^\infty \lambda_i^\infty \ne 0$ or $\sum_{i=\ell+1}^\infty |\langle \psi_i, y_0 \rangle_V|^2 \ne 0$, then by (1.77), (1.78) we have

$$\begin{aligned} &\sum_{i=\ell+1}^d \Big(\big|\langle \psi_i, y_0 \rangle_V\big|^2 + \frac{\sigma_n}{\delta t}\Big(\frac{1}{\delta\tau} + \Delta\tau\Big)\lambda_i\Big) \\ &\qquad \le 2 \sum_{i=\ell+1}^\infty \Big(\big|\langle \psi_i^\infty, y_0 \rangle_V\big|^2 + \frac{\sigma_n}{\delta t}\Big(\frac{1}{\delta\tau} + \Delta\tau\Big)\lambda_i^\infty\Big) \quad \text{for all } \Delta t \le \overline{\Delta t} \end{aligned}$$

and the dependence of the estimates of eigenvalues and eigenfunctions on n in (2.29) and (2.30) is thus eliminated. $\diamond$

Let us next derive some corollaries to the proof of Theorem 3.3. At first we consider the case, where the two grids coincide so that $n = m$ and $\tau_j = t_j$ for $j = 0, \dots, m$.

COROLLARY 3.5. *Suppose that the assumptions of Theorem* 3.3–a) *hold. If the two time discretizations coincide, then there exists a constant* $C > 0$ *depending on* T*, but independent of the grid* $\{\tau_j\}_{j=0}^m$*, such that*

$$\begin{aligned}\sum_{k=0}^{m} \beta_k \ \|Y_k - y(\tau_k)\|_H^2 &\le C(1+c_P^2)\Delta\tau^2 \ \|y_{tt}\|^2_{L^2(0,T;H)} \\ &+ C\Big(\sum_{i=\ell+1}^{d} \Big(\big|\langle \psi_i, y_0\rangle_V\big|^2 + \Big(\frac{1}{\delta\tau^2}+1\Big)\lambda_i\Big) + \Delta\tau^2 \ \|y_t\|_{L^2(0,T;V)}\Big).\end{aligned} \tag{2.31}$$

REMARK 3.6. Again, as in Theorem 3.3 b) compared to a), the factor $1 + c_P^2$ can be avoided in (2.31), if in place of (A1), (A2) we assume (A3) and replace the term $\|y_{tt}\|_{L^2(0,T;H)}$ by $\|y_{tt}\|_{L^2(0,T;V)}$. ◊

Let us briefly reflect on the behavior of the right-hand side of (2.29) and (2.30). First we note that if the number of POD elements for the Galerkin scheme coincides with the dimension of $\mathcal{V}$ then the first additive term on the right-hand side disappears. Secondly, if the number of snapshots is refined so that $\Delta t \to 0$ then the factor multiplying $\sum_{i=\ell+1}^{d} \lambda_i$ blows up. As noted above the term $\sum_{i=\ell+1}^{d} \lambda_i$ itself changes as the snapshots are refined. While computations for many concrete situations show that $\sum_{i=\ell+1}^{d} \lambda_i$ is small compared to $\Delta\tau$, the question nevertheless arises whether the term $1/(\delta\tau\delta t)$ can be avoided in the estimates. For this purpose we choose

$$\mathcal{V} = \text{span}\,\{y(t_0), \dots, y(t_n), \overline{\partial}_t y(t_1), \dots, \overline{\partial}_t y(t_n)\}, \tag{2.32}$$

where

$$\overline{\partial}_t y(t_j) = \frac{y(t_j) - y(t_{j-1})}{\delta t_j} \quad \text{for } j = 1, \dots, n,$$

(1.74) must be replaced by

$$\begin{aligned}\sum_{j=0}^{n} \alpha_j \Big\| y(t_j) - \sum_{i=1}^{\ell} \langle y(t_j), \hat\psi_i\rangle_V \hat\psi_i \Big\|_V^2 + \sum_{j=1}^{n} \alpha_j \Big\| \overline{\partial}_t y(t_j) - \sum_{i=1}^{\ell} \langle \overline{\partial}_t y(t_j), \hat\psi_i\rangle_V \hat\psi_i \Big\|_V^2 \\ = \sum_{i=\ell+1}^{d} \hat\lambda_i,\end{aligned}$$

where $\{\hat\lambda_i\}_{i\in\mathbb{N}}$, $\{\hat\psi_i\}_{i\in\mathbb{N}}$ are the eigenvalues and eigenfunctions of $\hat{\mathcal{R}}_n \in \mathcal{L}(V)$ given by

$$\hat{\mathcal{R}}_n z = \sum_{j=0}^{n} \alpha_j \left(\langle z, y(t_j)\rangle_V y(t_j) + \langle z, \overline{\partial}_t y(t_j)\rangle_V \overline{\partial}_t y(t_j)\right)$$

and satisfying

$$\hat{\mathcal{R}}_n \hat\psi_i = \hat\lambda_i \hat\psi_i, \quad \hat\lambda_1 \ge \ldots \ge \hat\lambda_{d(n)} > 0 \quad \text{and } \lambda_i = 0 \text{ for } i > d(n).$$

COROLLARY 3.7. *If in addition to the assumptions of Theorem* 3.3–a) *the snapshots set is taken as in* (2.32)*, then*

$$\begin{aligned}&\sum_{k=0}^{m} \beta_k \ \|Y_k - y(\tau_k)\|_H^2 \\ &\le C \sum_{i=\ell+1}^{d} \Big(\big|\langle \hat\psi_i, y_0\rangle_V\big|^2 + \frac{\sigma_n \Delta\tau}{\delta t}\hat\lambda_i\Big) + C\sigma_n\Delta\tau\Delta t \ \|y_t\|^2_{L^2(0,T;V)} \\ &\quad + C(1+c_P^2)\Delta\tau \left((\Delta\tau + \sigma_n\Delta t)\|y_{tt}\|^2_{L^2(0,T;H)} + \sigma_n\Delta t \|y_t\|^2_{L^2(0,T;H)}\right),\end{aligned}$$

where C *has the same properties as in Theorem* 3.3*.*

If we suppose that

$$\Delta t = O(\delta\tau) \quad \text{and} \quad \Delta\tau = O(\delta t), \tag{2.33}$$

then there exists a constant $c_1 > 0$ independent of $\{t_j\}_{j=0}^n$ and $\{\tau_j\}_{j=0}^m$ such that

$$\max\Big(\sigma_n, \frac{\sigma_n \Delta\tau}{\delta t}\Big) \le c_1. \tag{2.34}$$

For $y \in W^{2,2}(0,T;V)$, i.e., (A3) holds, we introduce the operator $\hat{\mathcal{R}} \in \mathcal{L}(V)$ corresponding to $\mathcal{R}$ by

$$\hat{\mathcal{R}} z = \int_0^T \langle z, y(t)\rangle_V y(t) + \langle z, y_t(t)\rangle_V y_t(t)\, dt \quad \text{for } z \in V.$$

Note that $\hat{\mathcal{R}} = \hat{\mathcal{Y}}\hat{\mathcal{Y}}^*$, where $\hat{\mathcal{Y}}^* : V \to W^{1,2}(0,T;\mathbb{R})$ is given by

$$(\hat{\mathcal{Y}}^* z)(t) = \langle z, y(t)\rangle_V.$$

Let us choose and fix ℓ such that

$$\hat\lambda_\ell^\infty \neq \hat\lambda_{\ell+1}^\infty. \tag{2.35}$$

Then, we have the following result, see [**22**].

COROLLARY 3.8. *Assume that $y \in W^{2,2}(0,T;V)$ and let the snapshots be chosen as in* (2.32)*. If* (2.33) *holds and ℓ satisfies* (2.35)*, then there exists a constant $C > 0$, independent of ℓ and the grids $\{t_j\}_{j=0}^n$ and $\{\tau_j\}_{j=0}^m$, and a $\overline{\Delta t} > 0$, depending on ℓ, such that*

$$\begin{aligned}\sum_{k=0}^m \beta_k \, \|Y_k - y(\tau_k)\|_H^2 \le C \sum_{i=\ell+1}^\infty \Big(\big|\langle y_0, \hat\psi_i^\infty\rangle_V\big|^2 + \hat\lambda_i^\infty\Big) \\ + C\Big(\Delta\tau\Delta t \, \|y_t\|^2_{L^2(0,T;V)} + \Delta\tau(\Delta\tau + \Delta t) \, \|y_{tt}\|^2_{L^2(0,T;V)}\Big)\end{aligned} \tag{2.36}$$

for all $\Delta t \le \overline{\Delta t}$.

REMARK 3.9. In (2.36) the first term on the right-hand side of the inequality reflects the spatial approximation error of the Galerkin-POD scheme and the second the approximation error due to the temporal backward Euler scheme. If the latter is replaced by the Crank-Nicolson method then, assuming $\Delta\tau = \Delta t$ and appropriate regularity on y, it can be shown with the techniques of this section that an estimate analogous to (2.36) holds with the first additive term on the right-hand side unchanged and the second one of fourth order in $\Delta\tau$. ◊

3.2. Case $X = H$. Suppose that the POD basis is constructed with respect to the H-norm. Differently from the situation, where the POD basis was constructed in V, the right-hand side of the estimate will involve the stiffness matrix

$$S = ((S_{ij})) \in \mathbb{R}^{d\times d} \quad \text{with} \quad S_{ij} = a(\psi_j, \psi_i).$$

THEOREM 3.10. *Suppose that* (A3) *holds and that $\Delta\tau$ is sufficiently small. Then there exists a constant $C > 0$ depending on T, but independent of the grids $\{t_j\}_{j=0}^n$ and $\{\tau_j\}_{j=0}^m$, such that*

$$\begin{aligned}\sum_{k=0}^m \beta_k \, \|Y_k - y(\tau_k)\|_H^2 \le C \sum_{i=\ell+1}^d \|S\|_2\Big(\big|\langle \psi_i, y_0\rangle_H\big|^2 + \frac{\sigma_n}{\delta t}\Big(\frac{1}{\delta\tau} + \Delta\tau\Big)\lambda_i\Big) \\ + C\sigma_n\Delta\tau\Big((\Delta\tau + \Delta t)\|y_{tt}\|^2_{L^2(0,T;V)} + \Delta t \, \|y_t\|^2_{L^2(0,T;V)}\Big).\end{aligned} \tag{2.37}$$

REMARK 3.11. Let us briefly discuss the asymptotic properties of the expression on the right-hand side of (2.37), which are restricted due to the appearance of $\delta t\delta\tau$ in the dominator, and the terms σ_n and $\|S\|_2$. As in section 3.1 the factor $1/\delta\tau$ can be eliminated by adding the set $\{\overline\partial y(t_j)\}_{j=1}^n$ to the set of snapshots. Assuming that $\Delta t = O(\delta\tau)$ and $\Delta\tau = O(\delta t)$ implies (2.34) and consequently, σ_n and $\sigma_n\Delta\tau/\delta t$ are uniformly bounded with respect to refinement of the t- and τ-grids. The factor $\|S\|_2$, which tends to infinity as $m \to \infty$ appears to be unavoidable in case the POD basis is computed in H. ◊

4. ROM for parameter-dependent PDEs

4.1. POD Galerkin scheme. In this section we discuss the ROM for (1.80). Let us fix $\ell \in \mathbb{N}$ and compute the first ℓ POD basis functions $\psi_1, \dots, \psi_\ell \in V$. Then we define the finite dimensional linear space

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

Endowed with the topology in V it follows that V^ℓ is a Hilbert space. Next we introduce the orthogonal projection $\mathcal{P}^\ell$ of V onto V^ℓ:

$$\mathcal{P}^\ell \varphi = \sum_{i=1}^{\ell} \langle \varphi, \psi_i \rangle_V \psi_i \quad \text{for } \varphi \in V. \tag{2.38}$$

From (2.38) and $\mathcal{P}^\ell \psi = \psi$ for all $\psi \in V^\ell$ it follows that

$$\langle \mathcal{P}^\ell \varphi, \psi \rangle_V = \langle \varphi, \psi \rangle_V \tag{2.39}$$

for all $\varphi \in V$ and all $\psi \in V^\ell$. Since the ψ_i's are orthonormal in V, we have $\|\mathcal{P}^\ell\|_{L(V)} = 1$, where $L(V)$ denotes the Banach space of all bounded linear operators from V into itself endowed with the common norm.

The POD Galerkin scheme for (1.80) leads to the following linear problem: for given $\mu \in \mathcal{I}$ determine a function $u^\ell \in V^\ell$ such that

$$a(u^\ell, \psi; \mu) = \langle f, \psi \rangle_{V',V} \quad \text{for all } \psi \in V^\ell \tag{2.40}$$

The proof of the existence of a unique solution u^ℓ to (2.40) follows by the Lax-Milgram theorem [**5**].

4.2. POD error estimates. The goal of this section is to present error estimates for the difference between the solution $u = u(\mu)$ to (1.80) and the POD solution $u^\ell(\mu)$ to (2.40) for $\mu \in \mathcal{I}$ in terms of the sum $\sum_{i=\ell+1}^{\infty} \lambda_i$, i.e., in terms of the sum over the eigenvalues corresponding to the not-modelled eigenmodes. Fir the proofs we refer to [**15**].

THEOREM 4.1. *Suppose that* (1.79) *holds. For $\mu \in \mathcal{I} = [\mu_a, \mu_b]$ we denote by $u(\mu)$ and $u^\ell(\mu)$ the solutions to* (1.80) *and* (2.40)*, respectively. Then there exists a constant $C > 0$ depending on μ_a, μ_b, c_V such that*

$$\int_{\mathcal{I}} \|u^\ell(\mu) - u(\mu)\|_V^2 \, \mathrm{d}\mu \le C \sum_{i=\ell+1}^{\infty} \lambda_i. \tag{2.41}$$

REMARK 4.2. Let us introduce for given $\ell \in \mathbb{N}$ the mapping

$$\tilde{\mathfrak{J}} : \underbrace{H \times \ldots \times H}_{\ell-\text{times}} \to \mathbb{R}$$

by

$$\tilde{\mathfrak{J}}(\psi_1, \dots, \psi_\ell) = \int_{\mathcal{I}} \Big\| u(\mu) - \sum_{i=1}^{\ell} \langle u(\mu), \psi_i \rangle_H \psi_i \Big\|_H^2 \, \mathrm{d}\mu.$$

Analogous to Section 3.1 we can compute the POD basis by solving the minimization problem

$$\min \tilde{\mathfrak{J}}(\tilde{\psi}_1, \dots, \tilde{\psi}_\ell) \quad \text{s.t.} \quad \langle \tilde{\psi}_j, \tilde{\psi}_i \rangle_V = \delta_{ij} \quad \text{for } 1 \le i, j \le \ell. \tag{2.42}$$

It turns out that the constant C in (2.41) depends on the spectral norm of S^ℓ, where $S^\ell \in \mathbb{R}^{\ell \times \ell}$ denotes the stiffness matrix with the elements $\langle \psi_i, \psi_j \rangle_V$, $1 \le i, j \le \ell$. ◊

Suppose that the weak solution to (1.80) is not known for all parameters $\mu \in \mathcal{I}$, but for the parameter grid $\{\mu_i\}_{i=1}^n$ introduced in (1.85). Let $u_i = u(\mu_i)$, $1 \le i \le n$, denote the corresponding solutions to (1.79) for the grid points μ_i. We define the snapshot set $\mathcal{V}^n = \operatorname{span}\{u_1, \dots, u_n\} \subset V$ and determine a POD basis of rank $\ell \le n$ for $\mathcal{V}^n$ by solving (2.26).

PROPOSITION 4.3. *Suppose that* (1.79) *holds and that* $\{\mu_j\}_{j=1}^n$ *is a grid in the interval* $\mathcal{I}$ *satisfying* (1.85). *For* μ_j, $1 \le j \le n$, *we denote by* $u(\mu_j)$ *and* $u^\ell(\mu_j)$ *the solutions to* (1.80) *and* (2.40), *respectively. Then there exists a constant* $C > 0$ *depending on* μ_a, μ_b, c_V, *but independent on the grid* $\{\mu_j\}_{j=1}^n$ *such that*

$$\sum_{j=1}^{n} \alpha_j \left\| u^\ell(\mu_j) - u(\mu_j) \right\|_V^2 \le C \sum_{i=\ell+1}^{d^n} \lambda_i^n.$$

Next we suppose that we are given two different grids $\{\mu_j\}_{j=1}^n$ and $\{\bar{\mu}_k\}_{k=1}^m$ in $\mathcal{I}$ satisfying

$$\mu_a = \mu_1 < \mu_2 < \ldots < \mu_n = \mu_b, \quad \mu_a = \bar{\mu}_1 < \bar{\mu}_2 < \ldots < \bar{\mu}_m = \mu_b. \tag{2.43}$$

We set

$$\delta\mu_j = \mu_j - \mu_{j-1},\ j = 2, \ldots, n, \quad \delta\mu = \min_{2 \le j \le n} \delta\mu_j, \quad \Delta\mu = \max_{2 \le j \le n} \delta\mu_j,$$
$$\delta\bar{\mu}_k = \bar{\mu}_k - \bar{\mu}_{k-1},\ k = 2, \ldots, m, \quad \delta\bar{\mu} = \min_{2 \le k \le m} \delta\bar{\mu}_k, \quad \Delta\bar{\mu} = \max_{2 \le k \le m} \delta\bar{\mu}_k.$$

Moreover, let

$$\alpha_1 = \frac{\delta\mu_2}{2}, \quad \alpha_j = \frac{\delta\mu_j + \delta\mu_{j+1}}{2} \text{ for } 2 \le j \le n-1, \quad \alpha_n = \frac{\delta\mu_n}{2},$$
$$\beta_1 = \frac{\delta\bar{\mu}_2}{2}, \quad \beta_k = \frac{\delta\bar{\mu}_k + \delta\bar{\mu}_{k+1}}{2} \text{ for } 2 \le k \le m-1, \quad \beta_m = \frac{\delta\bar{\mu}_m}{2}.$$

Next we present an error estimate for the term

$$\sum_{k=1}^{m} \beta_k \| u(\bar{\mu}_k) - u^\ell(\bar{\mu}_k) \|_V^2,$$

whereas the POD basis of rank ℓ is computed by using the snapshots ensemble $\{u(\mu_j)\}_{j=1}^n$ depending on the grid $\{\mu_j\}_{j=1}^n$. Let $\bar{\mu}_k \in \mathcal{I}$, $k \in \{1, \ldots, m\}$, be given. Then there exists an index $j_k \in \{1, \ldots, n-1\}$ such that

$$\mu_{j_k} \le \bar{\mu}_k \le \mu_{j_k+1}.$$

Let us define $\sigma_m \in \{1, \ldots, m\}$ as the maximum of the occurrence of the same value j_k as k ranges over $1 \le k \le m$. Notice that

$$\max\left\{ |\bar{\mu}_k - \mu_{j_k+1}|, |\bar{\mu}_k - \mu_{j_k}| \right\} \le \delta\mu_{j_k+1} \le \Delta\mu.$$

THEOREM 4.4. *Suppose that* (1.79) *holds, that* $\{\mu_j\}_{j=1}^n$ *and* $\{\bar{\mu}_k\}_{k=1}^m$ *are two grids in the interval* $\mathcal{I}$ *satisfying* (2.43). *For* $\bar{\mu}_k$, $1 \le k \le m$, *we denote by* $u(\bar{\mu}_k)$ *and* $u^\ell(\bar{\mu}_k)$ *the solutions to* (1.80) *and* (2.40), *respectively. Then there exists a constant* $C > 0$ *depending on* μ_a, μ_b, c_V, *but independent on the grids such that*

$$\sum_{k=1}^{m} \beta_k \left\| u^\ell(\bar{\mu}_k) - u(\bar{\mu}_k) \right\|_V^2 \le C \left(\frac{\sigma_m \Delta\bar{\mu}}{\delta\mu} \sum_{i=\ell+1}^{d^n} \lambda_i^n + \Delta\mu^2 \right).$$

In Theorem 4.4 the eigenvalues $\{\lambda_i\}_{i=1}^{d^n}$, the eigenfunctions $\{\psi_i\}_{i=1}^{d^n}$ and σ_m depend on the discretization of $\mathcal{I}$ for the snapshots as well as for the numerical integration. We address this dependence next. For a proof we refer to [**15**].

THEOREM 4.5. *Suppose that* (1.79) *holds, that* $\{\mu_j\}_{j=1}^n$ *and* $\{\bar{\mu}_k\}_{k=1}^m$ *are two grids in the interval* $\mathcal{I}$ *satisfying* (2.43). *Moreover, for both grids we have* $\Delta\mu = O(\delta\bar{\mu})$ *and* $\Delta\bar{\mu} = O(\delta\mu)$. *For* $\bar{\mu}_k$, $1 \le k \le m$, *we denote by* $u(\bar{\mu}_k)$ *and* $u^\ell(\bar{\mu}_k)$ *the solutions to* (1.80) *and* (2.40), *respectively. If* $\lambda_\ell \ne \lambda_{\ell+1}$ *holds, then there exists a constant* $C > 0$ *depending on* μ_a, μ_b, c_V, *but independent on the grids such that*

$$\sum_{k=1}^{m} \beta_k \left\| u^\ell(\bar{\mu}_k) - u(\bar{\mu}_k) \right\|_V^2 \le C \left(\sum_{i=\ell+1}^{\infty} \lambda_i + \Delta\mu^2 \right).$$

4.3. Continuous POD for semi-linear problem. Let us turn to a certain non-linear problem. Suppose that $F : V \to V'$ is a non-linear, locally Lipschitz-continuous mapping satisfying

$$\langle F(\phi) - F(\varphi), \phi - \varphi \rangle_{V',V} \geq 0 \quad \text{for all } \varphi, \psi \in V, \tag{2.44}$$

i.e., F is monotone. Instead of (1.80) we consider

$$a(u, \varphi; \mu) + \langle F(u), \varphi \rangle_{V',V} = \langle f, \varphi \rangle_{V',V} \quad \text{for all } \varphi \in V. \tag{2.45}$$

Example 4.6. Let us give an example for a semi-linear problem satisfying (2.44). We consider

$$-\Delta u + u^3 + \mu u = g \text{ in } \Omega \quad \text{and} \quad \frac{\partial u}{\partial n} + u = g_R \text{ on } \Gamma. \tag{2.46}$$

A *weak solution to* (2.46) satisfies $u \in V$ and

$$\int_\Omega \nabla u \cdot \nabla \varphi + \big(u^3 + \mu u\big)\varphi \,\mathrm{d}x + \int_\Gamma u\varphi \,\mathrm{d}s = \int_\Omega g\varphi \,\mathrm{d}x + \int_\Gamma g_R \varphi \,\mathrm{d}s \tag{2.47}$$

for all $\varphi \in V$. We utilize the parametrized bilinear form $a(\cdot\,, \cdot\,; \mu) : V \times V \to \mathbb{R}$ given by

$$a(\varphi, \phi; \mu) = \int_\Omega \nabla\varphi \cdot \nabla\phi \,\mathrm{d}x + \int_\Gamma \varphi\phi \,\mathrm{d}s + \mu \int_\Omega \varphi\phi \,\mathrm{d}x = \langle \varphi, \phi \rangle_V + \mu \langle \varphi, \phi \rangle_H$$

for all $\varphi, \phi \in V$, $\mu \in \mathcal{I}$ and the linear and continuous functional $f : V \to \mathbb{R}$ defined as

$$\langle f, \varphi \rangle_{V',V} = \int_\Omega g\varphi \,\mathrm{d}x + \int_\Gamma g_R \varphi \,\mathrm{d}s$$

for all $\varphi \in V$. Moreover, we define the non-linear operator $F : V \to V'$ by

$$\langle F(\phi), \varphi \rangle_{V',V} = \int_\Omega \phi^3 \varphi \,\mathrm{d}x \quad \text{for } \phi, \varphi \in V.$$

Then, a weak solution to (2.46) satisfies the variational formulation (2.45). Recall that $\varphi \in V$ implies $\varphi \in L^6(\Omega)$. Consequently, $F(\varphi) \in H \subset V'$. Let $\phi, \varphi \in V$ and $\chi = \phi - \varphi \in V$. From

$$\langle F(\phi) - F(\varphi), \chi \rangle_{V',V} = 3 \int_0^1 \int_\Omega (\phi + s\chi)^2 \chi^2 \,\mathrm{d}x\mathrm{d}s \geq 0$$

it follows that (2.44) holds. Existence of a solution to (2.47) is proved in [**7**]. Suppose that $u, v \in V$ are two solutions to (2.47). Then we have

$$a(u - v, \varphi; \mu) + \langle F(u) - F(v), \varphi \rangle_{V',V} = 0 \quad \text{for all } \varphi \in V \text{ and } \mu \in \mathcal{I}.$$

Choosing $\varphi = u - v$, using (1.79) and (2.44) we derive that $u = v$ in V. Thus, (2.47) has a unique solution.$\Diamond$

Suppose that we have computed a POD basis $\{\psi_i\}_{i=1}^\ell$ of rank ℓ by utilizing the solution $u(\mu)$ to (2.45) for all $\mu \in \mathcal{I}$. The POD Galerkin scheme for (2.46) is as follows: Find $u^\ell = u^\ell(\mu)$, $\mu \in \mathcal{I}$, such that

$$a(u^\ell, \psi; \mu) + \langle F(u^\ell), \psi \rangle_{V',V} = \langle f, \psi \rangle_{V',V} \quad \text{for all } \psi \in V^\ell. \tag{2.48}$$

In the following theorem an error estimate is presented. The proof is ananlogous to the proof of Theorem 4.1.

Theorem 4.7. *Let $F : V \to V'$ be a locally Lipschitz-continuous mapping satisfying* (2.44). *Suppose that for every $\mu \in \mathcal{I} = [\mu_a, \mu_b]$ there exist unique solutions to* (2.45) *and* (2.48) *denoted by $u(\mu)$ and $u^\ell(\mu)$, respectively. Then there exists a constant $C > 0$ depending on μ_a, μ_b, c_V and a Lipschitz constant for F such that*

$$\int_\mathcal{I} \|u^\ell(\mu) - u(\mu)\|_V^2 \,\mathrm{d}\mu \leq C \sum_{i=\ell+1}^\infty \lambda_i.$$

Remark 4.8. If the POD basis is computed by the strategy in Section 3.2, POD error estimates can also be derived combining the techniques in Section 4.2 and the arguments in the proof of Theorem 4.7. $\Diamond$

CHAPTER 3

Suboptimal control using POD

Optimal control problems for nonlinear partial differential equations are often hard to tackle numerically so that the need for developing novel techniques emerges. One such technique is given by reduced order methods. Recently the application of reduced-order models to optimal control problems for partial differential equations 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, where the elements of the subspaces are uncorrelated to the physical properties of the system that they approximate. The reduced basis method as developed, e.g., in [**6, 13, 31, 35**] is one such reduced-order method with the basis elements corresponding to the dynamics of expected control regimes.

In our application we apply POD to derive a Galerkin approximation in the spatial variable, with basis functions corresponding to the solution of the physical system at pre-specified time instances. This leads to a drastic reduction of the degrees of freedom and allows for a fast solution of the optimal control problem. This chapter is organized in the following manner. In Section 1 we recall basic facts from finite-dimensional optimal control theory necessary for the investigation of PDE constrained optimization problems. In Section 2 we focus on the choice of the POD ansatz functions in the context of optimal control problems. Finally, in Section 3 we discuss a-posteriori error estimates that can be used to determine the cardinality of the POD basis functions in order to guarantee a given tollerance for the difference between the (unknown) optimal control and its computed suboptimal POD control.

1. The finite-dimensional case

Before we investigate PDE-constrained optimization problems we start with optimization problems in finite dimensions. This emphasizes the optimization aspects, whereas in PDE-constrained optimization we also have to deal with functional analysis and PDE theory. The presentation follows parts of the book [**40**].

Suppose that $J : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a given cost functional. Then we consider the minimization problem

$$\min J(y,u) \quad \text{s.t.} \quad (y,u) \in \mathbb{R}^n \times U_{ad} \text{ with } Ay = Bu, \tag{3.1}$$

where U_{ad} is a non-empty subset of $\mathbb{R}^m$ and $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ hold. We suppose that A is non-singular and define the *admissible set*

$$\mathcal{F}(3.1) = \big\{(y,u) \in \mathbb{R}^n \times \mathbb{R}^m \,\big|\, Ay = Bu \text{ and } u \in U_{ad}\big\}.$$

Then, each pair $(y,u) \in \mathcal{F}(3.1)$ satisfies $y = A^{-1}Bu$. Hence, for any $u \in U_{ad}$ there exists a unique $y \in \mathbb{R}^n$ so that $(y,u) \in \mathcal{F}(3.1)$ holds. For that reason we call u the *control* and y the associated (unique) *state*. Let us introduce the matrix $S = A^{-1}Bu \in \mathbb{R}^{n\times m}$. Then, for any control $u \in U_{ad}$ the associated state is given by $y = Su$. Moreover, we define the so-called reduced cost functional $\hat{J} : \mathbb{R}^m \to \mathbb{R}$ as

$$\hat{J}(u) = J(Su,u) \quad \text{for } u \in \mathbb{R}^m.$$

Now, (3.1) can be equivalently expressed ny

$$\min \hat{J}(u) \quad \text{s.t.} \quad u \in U_{ad}. \tag{3.2}$$

Definition 1.1. *A vector $u^* \in U_{ad}$ is called an* optimal control *for* (3.2) *if $\hat{J}(u^*) \le \hat{J}(u)$ holds for all $u \in U_{ad}$. The associated* optimal state *is $y^* = Su^*$.*

Now we can give sufficient conditions so that (3.1) possesses at least one optimal control.

THEOREM 1.2. *If $\hat{J}$ is continuous, U_{ad} non-empty, bounded, closed and A be invertible, there exists at least one optimal control u^* solving* (3.2).

An optimal solution to (3.2) can be characterized by *optimality conditions*.

THEOREM 1.3. *If $\hat{J}$ is continuously differentiable and u^* an optimal control for* (3.2) *(or* (3.1)*), then*

$$\hat{J}'(u^*)(u-u^*) \geq 0 \quad \text{for all } u \in U_{ad}, \tag{3.3}$$

where $\hat{J}'(u^*)h$ denotes the directional derivative of $\hat{J}$ at u^* in direction $h \in \mathbb{R}^m$.

Note that (3.3) is a *first-order necessary optimality condition* for (3.2). Moreover,

$$\begin{aligned}\hat{J}(u^*)h &= \nabla_y J(Su^*,u^*)^T(Sh) + \nabla_u J(Su^*,u^*)^T h \\ &= \nabla_y J(Su^*,u^*)^T(A^{-1}Bu^*) + \nabla_u J(Su^*,u^*)^T h \\ &= \big(B^T A^{-T}\nabla_y J(Su^*,u^*) + \nabla_u J(Su^*,u^*)\big)^T h.\end{aligned}$$

Therefore, (3.3) is equivalent with

$$\langle B^T A^{-T}\nabla_y J(Su^*,u^*) + \nabla_u J(Su^*,u^*), u-u^*\rangle_{\mathbb{R}^m} \geq 0 \quad \text{for all } u \in U_{ad}. \tag{3.4}$$

To avoid the numerical realization of A^{-T} we introduce the *adjoint state*

$$p^* = -A^{-T}\nabla_y J(Su^*,u^*)$$

i.e., p^* solves the linear system

$$A^T p^* = -\nabla_y J(Su^*,u^*) \tag{3.5}$$

that is called the *adjoint equation*. Then,

$$\hat{J}'(u^*) = \nabla_u J(Su^*,u^*) - B^T p^*$$

Inserting p^* into (3.4) we find

$$\langle \nabla_u J(Su^*,u^*) - B^T p^*, u-u^*\rangle_{\mathbb{R}^m} \geq 0 \quad \text{for all } u \in U_{ad}.$$

Summarizing we obtain

$$Ay^* = Bu^*, \qquad u^* \in U_{ad} \tag{3.6a}$$

$$A^T p^* = -\nabla_y J(Su^*,u^*), \tag{3.6b}$$

$$\langle \nabla_u J(Su^*,u^*) - B^T p^*, u-u^*\rangle_{\mathbb{R}^m} \geq 0 \quad \text{for all } u \in U_{ad} \tag{3.6c}$$

If $U_{ad} = \mathbb{R}^m$ holds, we have instead of (3.6c)

$$\nabla_u J(Su^*,u^*) - B^T p^* = 0.$$

A different approach in deriving optimality conditions is based on the *Lagrange functional* $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$L(y,u,p) = J(y,u) + \big(Ay - Bu\big)^T p \quad \text{for } y,\, p \in \mathbb{R}^n \text{ and } u \in \mathbb{R}^m.$$

We find that (3.6a) and (3.6b) are equivalent with

$$\nabla_p L(y^*,u^*,p^*) = 0 \quad \text{and} \quad \nabla_y L(y^*,u^*,p^*) = 0,$$

respetively. Furthermore, (3.6b) can be written equivalently as

$$\langle \nabla_u L(y^*,u^*,p^*), u-u^*\rangle_{\mathbb{R}^m} \geq 0 \quad \text{for all } u \in U_{ad}.$$

Consequently, the adjoint equation can be derived from the partial derivative of the Lagrangian with respect to y. Therefore, its solution p^* is also called the *Lagrange multiplier* associated with the optimal solution pair (y^*,u^*).

Note that (y^*, u^*) is a solution to the first-order necessary optimality conditions for the minimization problem

$$\min L(y, u, p^*) \quad \text{s.t.} \quad y \in \mathbb{R}^n,\ u \in U_{ad}.$$

A specific situation is the admissible set

$$U_{ad} = \left\{ u \in \mathbb{R}^m \,\middle|\, u_a \le u \le u_b \right\}$$

where '$\le$' is interpreted componentwise, i.e, $u_{a,i} \le u_i \le u_{b,i}$ for $1 \le i \le m$, and $u_a \le u_b$ holds. Then, (3.6c) implies

$$\langle \nabla_u J(Su^*, u^*) - B^T p^*, u^* \rangle_{\mathbb{R}^m} \le \langle \nabla_u J(Su^*, u^*) - B^T p^*, u \rangle_{\mathbb{R}^m} \quad \text{for all } u \in U_{ad}$$

Thus, u^* solves

$$\min_{u \in U_{ad}} \langle \nabla_u J(Su^*, u^*) - B^T p^*, u \rangle_{\mathbb{R}^m}.$$

Since $u \in U_{ad}$ holds, the components u_i, $1 \le i \le m$, of u are not coupled. Thus, u can be computed componentwise by solving

$$\min_{u_{a,i} \le u_i \le u_{b,i}} \left(\nabla_u J(Su^*, u^*) - B^T p^* \right)_i u_i, \quad 1 \le i \le m.$$

We obtain

$$u_i^* = \begin{cases} u_{b,i} & \text{if } \left(\nabla_u J(Su^*, u^*) - B^T p^* \right)_i < 0, \\ u_{a,i} & \text{if } \left(\nabla_u J(Su^*, u^*) - B^T p^* \right)_i > 0 \end{cases} \tag{3.7}$$

In case of $\left(\nabla_u J(Su^*, u^*) - B^T p^* \right)_i = 0$ we do not get any information from the variational inequatlity (3.6c).

Define

$$\mu_a = \left[B^T p + \nabla_u J(\bar{y}, \bar{u})) \right]_+ \quad \text{and} \quad \mu_b = \left[B^T p + \nabla_u J(\bar{y}, \bar{u})) \right]_- \tag{3.8}$$

where $[s]_+ = \max(s, 0)$ stands for the positive part function and $[s]_- = -\min(s, 0)$ denotes for the negative part function. From (3.7) we derive the conditions

$$\mu_a \ge 0,\ u_a - \bar{u} \le 0,\ \left(u_a - \bar{u} \right)^T \mu_a = 0, \mu_b \ge 0,\ \bar{u} - u_b \le 0,\ \left(\bar{u} - u_b \right)^T \mu_b = 0. \tag{3.9}$$

The system (3.9) is called the *complementarity system*. Utilizing (3.8) we find $\mu_a - \mu_b = \nabla_u J(\bar{y}, \bar{u}) + B^T p$, i.e.,

$$\nabla_u J(\bar{y}, \bar{u}) + B^T p + \mu_b - \mu_a = 0. \tag{3.10}$$

Next we extend the Lagrange function by

$$\mathcal{L}(y, u, p, \mu_a, \mu_b) = J(y, u) + \left(Ay - Bu \right)^T p + \left(u_a - u \right)^T \mu_a + \left(u - u_b \right)^T \mu_b.$$

Then, (3.10) can be written as

$$\nabla_u \mathcal{L}(\bar{y}, \bar{u}, p, \mu_a, \mu_b) = 0.$$

Moreover, the equation

$$\nabla_y \mathcal{L}(\bar{y}, \bar{u}, p, \mu_a, \mu_b) = 0$$

is equivalent to the adjoint equation (since $\nabla_y \mathcal{L} \equiv \nabla_y L$). The vectors μ_a and μ_b are the *Lagrange multipliers* for the inequality constraints $u_a - u \le 0$ and $u - u_b \le 0$. The first-order necessary optimality conditions (*Karush-Kuhn-Tucker conditions*) are given as follows:

$$\begin{cases} \nabla_y \mathcal{L}(\bar{y}, \bar{u}, p, \mu_a, \mu_b) = 0, \\ \nabla_u \mathcal{L}(\bar{y}, \bar{u}, p, \mu_a, \mu_b) = 0, \\ A\bar{y} = B\bar{u},\ \bar{u} \in U_{ad}, \\ \mu_a \ge 0,\ \mu_b \ge 0,\ (u_a - u)^T \mu_a = 0,\ (u - u_b)^T \mu_b = 0. \end{cases}$$

2. Proper orthogonal decomposition for optimality systems

The POD method is based on a Galerkin type discretization with basis elements created from the dynamical system itself. In the context of optimal control this approach may suffer from the fact that the basis elements are computed from a reference trajectory containing features which are quite different from those of the optimally controlled trajectory. A method is proposed which avoids this problem of unmodelled dynamics in the proper orthogonal decomposition approach to optimal control. It is referred to as *optimality system proper orthogonal decomposition* (OS-POD). For more details we refer the reader to [**24**].

2.1. The augmented optimal control problem. The dynamical system under consider is of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\langle y(t),\varphi\rangle_H + a(y(t),\varphi) + \langle \mathcal{N}(y(t)),\varphi\rangle_{V^*,V} = \sum_{k=1}^{m} u_k(t)\,\langle b_k,\varphi\rangle_H \tag{3.11a}$$

for allmost all $t \in (0,T]$ and

$$\langle y(0),\varphi\rangle_H = \langle y_0,\,\varphi\rangle_H \quad \text{for all } \varphi \in V, \tag{3.11b}$$

with the following specifications holding throughout

- $T > 0$, V and H are separable real Hilbert spaces, with V dense and compact in H, and $V \subset H = H^* \subset V^*$ a Gelfand triple,
- $a : V \times V \to \mathbb{R}$ is a symmetric bilinear form satisfying $a(\varphi,\varphi) \geq \alpha\,\|\varphi\|_V^2$ for some $\alpha > 0$ independent of φ,
- $\mathcal{N} : V \to V^*$ is a twice continuously Fréchet-differentiable operator,
- the control shape functions b_k are chosen in H with control intensities $u \in L^2(0,T;\mathbb{R}^m)$,
- $y_0 \in V$.

We associate with a the isomorphism $\mathcal{A} : V \to V^*$, which can alternatively be considered as linear unbounded selfadjoint operator in H with domain $D(\mathcal{A}) = \{\varphi \in V : \mathcal{A}\varphi \in H\}$. Defining $\mathcal{B} : \mathbb{R}^m \to H$ by $\mathcal{B}(\upsilon) = \sum_{k=1}^m \upsilon_k\, b_k$ we can express (3.11) in operator form as

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}\,y(t) + \mathcal{A}y(t) + \mathcal{N}(y(t)) = \mathcal{B}(u(t)) \quad \text{for } t \in (0,T], \\ y(0) = y_0. \end{cases} \tag{3.12}$$

Further assumptions are necessary for the nonlinearity $\mathcal{N}$. We choose conditions which are satisfied for nonlinearities of Navier-Stokes type, see [**39**, Chapter III], for example.

$$\textbf{(H1)} \qquad \begin{cases} \text{For every } u \in L^2(\mathbb{R}^m) \text{ there exists a unique solution} \\ y = y(u) \in L^2(D(\mathcal{A})) \cap H^1(V) \text{ and moreover} \\ \text{there exists a continuous function } c_1 : \mathbb{R} \to \mathbb{R} \text{ such that} \\ \|y(u)\|_{L^2(D(\mathcal{A}))\cap H^1(V)} \leq c_1\big(\|u\|_{L^2(\mathbb{R}^m)}\big) \text{ for all } u \in L^2(\mathbb{R}^m). \end{cases}$$

Here and throughout we shall abbreviate $L^2(0,T;Y)$ by $L^2(Y)$, and analogously for $H^1(0,T;Y)$ and $C([0,T];Y)$. We further require the assumptions

$$\textbf{(H2)} \qquad \begin{cases} \text{there exist real constants } c_2 \text{ and } c_3 \text{ such that} \\ -\langle \mathcal{N}(\psi),\psi\rangle_{V^*,V} \leq \frac{\alpha}{2}\,\|\psi\|_V^2 + c_2\,\|\psi\|_H^2 + c_3 \text{ for all } \psi \in V \text{ and} \\ \mathcal{N} \text{ maps bounded sets in } V \text{ to bounded sets in } V^*, \end{cases}$$

and

$$\textbf{(H3)} \qquad D(\mathcal{A}) \text{ embeds compactly into } V.$$

For the Navier-Stokes nonlinearity, (**H2**) is satisfied with $c_2 = c_3 = 0$. We consider an optimal control problem of tracking type. Different cost functionals could be treated quite analogously.

$$\begin{cases} \min J(y,u) = \min \dfrac{\beta}{2}\displaystyle\int_0^T \|y(t)-z(t)\|_H^2 \,\mathrm{d}t + \dfrac{1}{2}\int_0^T u(t)^T \mathbf{R} u(t)\,\mathrm{d}t \\ \text{subject to } u \in L^2(\mathbb{R}^m) \text{ and (3.11)}, \end{cases} \tag{3.13}$$

where $\beta > 0$, $z \in L^2(H)$, and $\mathbf{R} \in \mathbb{R}^{m\times m}$ is positive definite and symmetric. To denote the reduced cost functional we write $\hat{J}(u) = J(y(u),u)$, with $y(u)$ the solution to (3.11) for given u. With (**H1**) holding it is standard to argue existence of a solution $(y^*,u^*) = (y(u^*),u^*)$ to (3.13).

2.2. The POD approximation. Next we introduce a POD-Galerkin model for (3.13). To define the POD reduction with basis $\{\psi_i\}_{i=1}^{\ell}$ let

$$X = H \quad \text{or} \quad X = V$$

and for $y \in L^2(X)$ let $\mathcal{R} : X \to X$ be given by

$$\mathcal{R}\psi = \int_0^T \langle y(t),\psi\rangle_X \, y(t)\,\mathrm{d}t \quad \text{for } \psi \in X.$$

Clearly $\mathcal{R}$ is a bounded, nonnegative, selfadjoint operator which can be expressed as

$$\mathcal{R} = \mathcal{Y}\mathcal{Y}^*,$$

where $\mathcal{Y} : L^2(\mathbb{R}) \to X$ is defined by

$$\mathcal{Y}v = \int_0^T v(t)\, y(t,\cdot)\,\mathrm{d}t \quad \text{for } v \in L^2(\mathbb{R}),$$

and the adjoint $\mathcal{Y}^* : X \to L^2(\mathbb{R})$ is given by

$$\mathcal{Y}^*\psi = \langle y(t,\cdot),\psi(\cdot)\rangle_X \quad \text{for } \psi \in X.$$

We shall also utilize the operator $\mathcal{K} : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ defined by

$$\mathcal{K} = \mathcal{Y}^*\mathcal{Y}$$

or explicitly

$$(\mathcal{K}v)(t) = \int_0^T \langle y(t,\cdot), y(s,\cdot)\rangle_X \, v(s)\,\mathrm{d}s \quad \text{for } v \in L^2(\mathbb{R}).$$

For $x \in L^2(X)$ it follows that the operator $\mathcal{K}$ is compact. Moreover, except for possibly 0, $\mathcal{K}$ and $\mathcal{R}$ possess the same eigenvalues which are positive with identical multiplicities and ψ is eigenvector of $\mathcal{R}$ if and only if $\mathcal{Y}^*\psi = \langle y(t,\cdot),\psi\rangle_X$ is an eigenvector of $\mathcal{K}$.

We shall utilize POD bases $\{\psi_i(y)\}_{i=1}^{\ell}$ with respect to $X = H$ or $X = V$ satisfying $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_\ell > 0$, and

$$\mathcal{R}(y)\psi_i = \int_0^T \langle y(t,\cdot),\psi_i\rangle_X \, y(t)\,\mathrm{d}t = \lambda_i\psi_i \qquad \text{for } i = 1,\ldots,\ell, \tag{3.14a}$$

$$\langle \psi_i,\psi_j\rangle_X = \delta_{ij} \qquad \text{for } i,j = 1,\ldots,\ell. \tag{3.14b}$$

The POD-subspaces are denoted by

$$V^\ell = \operatorname{span}\{\psi_1,\ldots,\psi_\ell\}.$$

Note that V^ℓ depends on y. In this paper the POD-subspaces are generated by trajectories y which arise as controlled trajectories of (3.11). We shall require the following condition

$$\min\{\lambda_\ell(\mathcal{R}(y)) \,|\, y \text{ solves (3.11) with } u \in L^2(\mathbb{R}^m)\} > 0. \tag{**H4**}$$

Note that $\psi_i \in V$ also for $X = H$. This follows from (3.14a) using that $y \in L^2(V)$. Moreover, $\psi_i \in D(\mathcal{A})$ for $y \in L^2(D(\mathcal{A}))$.

To obtain the POD-Galerkin approximation to (3.11) we make the ansatz

$$y^\ell(t) = \sum_{j=1}^{\ell} x_j(t)\,\psi_j,$$

replace y by y^ℓ in (3.11), take inner products in H with respect to $\{\psi_i\}_{i=1}^\ell$ and obtain the system of ordinary differential equations in $\mathbb{R}^\ell$

$$\begin{cases} E(\psi)\,\dot{x}(t) + A(\psi)\,x(t) + \mathfrak{N}(x(t),\psi) = B(\psi)u(t) \quad \text{for} \quad t \in (0,T] \\ E(\psi)\,x(0) = x_0(\psi). \end{cases}$$

Here $E : X^\ell \times X^\ell \to \mathbb{R}^{\ell\times\ell}$ with $X^\ell = \bigotimes_{i=1}^\ell X$ is defined by

$$E_{ij}(\varphi,\phi) = \langle \varphi_i, \phi_j \rangle_H \text{ and } E(\varphi) = E(\varphi,\varphi),$$

$A : X^\ell \times X^\ell \to \mathbb{R}^{\ell\times\ell}$ is defined by

$$A_{ij}(\varphi,\phi) = a(\varphi_i,\phi_j) \text{ and } A(\varphi) = A(\varphi,\varphi),$$

$B : X^\ell \to \mathbb{R}^{\ell\times m}$ and $x_0 : X^\ell \to \mathbb{R}^\ell$ are given by

$$B_{ij}(\varphi) = \langle \varphi_i, b_j \rangle_H, \quad x_{0,i}(\varphi) = \langle y_0, \varphi_i \rangle_H,$$

and the nonlinearity $\mathfrak{N} : \mathbb{R}^\ell \times X^\ell \times X^\ell \to \mathbb{R}^\ell$ by

$$\mathfrak{N}_i(x,\psi,\varphi) = \left\langle \mathcal{N}\Big(\sum_{j=1}^{\ell} x_j\,\psi_j\Big), \varphi_i \right\rangle_{V^*,V} \text{ with } \mathfrak{N}(x,\varphi) = \mathfrak{N}(x,\varphi,\varphi).$$

Discretizing the cost function in the same manner we obtain

$$\begin{aligned} J^\ell(x,\psi,u) = \frac{\beta}{2}\int_0^T \big(x(t)^T(E(\psi)\,x(t) - 2z^\ell(t,\psi)) + \|z(t)\|_H^2\big)\,\mathrm{d}t \\ + \frac{1}{2}\int_0^T u^T(t)\,\mathbf{R}u(t)\,\mathrm{d}t, \end{aligned}$$

where $z^\ell : (0,T)\times X^\ell \to \mathbb{R}^\ell$ is given by

$$z^\ell(t,\varphi)_i = \langle z(t), \varphi_i \rangle_H,$$

and $J^\ell : L^2(\mathbb{R}^\ell) \times X^\ell \times L^2(\mathbb{R}^m) \to \mathbb{R}^+$.

We are now prepared to specify the POD-Galerkin reduced optimal control problem augmented with the POD-generation criteria:

$$\begin{cases} \min J^\ell(x,\psi,u) \quad \text{over} \quad (x,\psi,u) \in L^2(\mathbb{R}^\ell) \times X^\ell \times L^2(\mathbb{R}^m), \\ \text{subject to} \\ E(\psi)\,\dot{x}(t) + A(\psi)x(t) + \mathfrak{N}(x(t),\psi) = B(\psi)u(t) \quad \text{for } t \in (0,T], \\ E(\psi)\,x(0) = x_0(\psi), \\ \frac{\mathrm{d}}{\mathrm{d}t}\,y(t) + \mathcal{A}y(t) + \mathcal{N}(y(t)) = \mathcal{B}(u(t)) \quad \text{for } t \in (0,T], \\ y(0) = y_0, \\ \mathcal{R}(y)\psi_i = \lambda_i\psi_i \quad \text{for } i = 1,\dots,\ell, \\ \langle \psi_i, \psi_j \rangle_X = \delta_{ij} \quad \text{for } i,j = 1,\dots,\ell. \end{cases} \tag{3.15}$$

If the POD-eigenvalue problem is solved at a reference trajectory $y(\bar{u})$ corresponding to a fixed reference control $\bar{u}$, this results in the last four equations from (3.15). The remaining optimization is the standard one in the POD-Galerkin optimal control approach. In [**24**] te next result is proved.

THEOREM 2.1. *If* (**H1**)–(**H4**) *is satisfied, then* (3.15) *admits a (global) solution* $(x^*,\psi^*,u^*) \in W^{1,2}(\mathbb{R}^\ell)\times X^\ell \times L^2(\mathbb{R}^m)$ *with* $(\lambda^*,y^*) \in \mathbb{R}^m \times (L^2(D(\mathcal{A})) \cap W^{1,2}(V))$ *and* $y^* = y(u^*)$.

2.3. The optimality system. Suppose that $(x^*, \psi^*, u^*) \in W^{1,2}(\mathbb{R}^\ell) \times X^\ell \times L^2(\mathbb{R}^m)$ is a local solution to (3.15). We proceed by deriving an optimality system. For this purpose we assume that the eigenvalues of $\mathcal{R}(y^*)$ with $y^* = y(u^*)$ are distinct. If this is not the case then in the following results we have to keep the orthonormality condition on the subspace corresponding to a multiple eigenvalue as explicit constraints. For $\lambda_i \neq \lambda_j$ we have $\langle \psi_i, \psi_j \rangle_X = 0$ since $\mathcal{R}$ is selfadjoint. Therefore (3.14b) will be replaced by $\|\psi_i^*\|_X = 1$ for $i = 1, \dots, \ell$.

Henceforth the state and the control variables are considered in the space

$$Z = H^1(\mathbb{R}^\ell) \times W(0,T) \times X^\ell \times \mathbb{R}^\ell \times L^2(\mathbb{R}^m),$$

where $W(0,T) = L^2(V) \cap H^1(V^*)$ and the generic element of Z is denoted by $\mathsf{z} = (x, y, \psi, \lambda, u)$. We utilize adjoint variables form the space

$$\Xi = L^2(\mathbb{R}^\ell) \times \mathbb{R}^\ell \times L^2(V) \times H \times X^\ell \times \mathbb{R}^\ell$$

with generic element $\xi = (q, q_0, p, p_0, \mu, \eta)$. For $i = 1, \dots, \ell$ we introduce $\mathcal{G}_i : H^1(\mathbb{R}^\ell) \times X^\ell \times L^2(\mathbb{R}^m) \times H^1(\mathbb{R}^\ell) \to X'$ by

$$\begin{aligned}
\mathcal{G}_i(x, \psi, u, q) = &\int_0^T \left(x_i \Big(\sum_{j=1}^\ell x_j \psi_j - z \Big) + q_i \sum_{j=1}^\ell \dot{x}_j \, \psi_j + \dot{x}_j \sum_{j=1}^\ell q_j \, \psi_j \right) \mathrm{d}t \\
&+ \int_0^T \left(q_i \sum_{j=1}^\ell x_j \, \mathcal{A}\psi_j + x_i \sum_{j=1}^\ell q_j \, \mathcal{A} \, \psi_j - q_i \sum_{k=1}^m b_k \, u_k \right) \mathrm{d}t \\
&+ \sum_{j=1}^\ell (x_j(0) \, \psi_j \, q_i(0) + x_i(0) \psi_j \, q_j(0)) - y_0 \, q_i(0) \\
&+ \mathcal{N}\Big(\sum_{k=1}^\ell x_k \, \psi_k \Big) q_i + x_i \sum_{j=1}^\ell \mathcal{N}' \Big(\sum_{k=1}^\ell x_k \, \psi_k \Big)^* q_j \psi_j .
\end{aligned}$$

THEOREM 2.2. *Let* (**H1**)–(**H4**) *hold and let*

$$\mathsf{z} = (x, y, \psi, \lambda, u) \in W^{1,2}(\mathbb{R}^\ell) \times W^{1,2}(V) \times X^\ell \times \mathbb{R}^m \times L^2(\mathbb{R}^m)$$

denote a solution to (3.15). *Assume that the eigenvalues of* $\mathcal{R}(y)$ *are distinct and that*

$$\frac{\mathrm{d}}{\mathrm{d}t} v + \mathcal{A}v + \mathcal{N}'(y(t))v - \mathcal{B}\tilde{u} = w \text{ for } t \in (0,T] \quad \text{and} \quad v(0) = v_0$$

admits a solution $(v, \tilde{u}) \in W(0,T) \times L^2(\mathbb{R}^m)$ *for every* $(w, v_0) \in L^2(V^*) \times H$. *Then there exist* $(q, p, \mu, \eta) \in L^2(\mathbb{R}^\ell) \times L^2(V) \times X^\ell \times \mathbb{R}^\ell$ *such that the following optimality system holds:*

$$\begin{cases}
-E(\psi) \, \dot{q}(t) + \big(A(\psi) + \mathfrak{N}_x^T(x(t), \psi) \big) \, q(t) = -\beta \big(E(\psi) x(t) - z(t, \psi) \big), \\
q(T) = 0,
\end{cases} \tag{3.16}$$

$$\begin{cases}
-\dot{p}(t) + \mathcal{A}p(t) + \mathcal{N}'(y(t))^* \, p(t) \\
\qquad = \sum_{i=1}^\ell \langle y(t), \mu_i \rangle_X \mathcal{I}^{-1} \psi_i + \langle y(t), \psi_i \rangle_X \mathcal{I}^{-1} \mu_i, \\
p(T) = 0,
\end{cases}$$

$$\begin{cases}
\eta_i = -\frac{1}{2} \langle \mathcal{G}_i(x, \psi, u, q), \psi_i \rangle_{X^*, X} \\
\mu_i = -(\mathcal{R} - \lambda_i \, I)^{-1} \left[2 \, \eta_i \, \psi_i + \mathcal{I} \, \mathcal{G}_i(x, \psi, u, q) \right] \quad \text{for } i = 1, \dots, \ell,
\end{cases}$$

$$\mathbf{R}u(t) = B^T(\psi) \, q(t) + \mathcal{B}^* p(t).$$

The optimality conditions can also be formulated in terms of the operator $\mathcal{K}$. This is helpful if $\mathcal{K}$ is of smaller dimension than $\mathcal{R}$.

COROLLARY 2.3. *Let* $\mathsf{z} = (x, y, \psi, \lambda, u) \in Z$ *denote a solution of* (3.15) *and let the assumptions of Theorem* 2.2 *hold. Then there exists* $(q, p^{\mathcal{K}}, \mu^{\mathcal{K}}, \eta) \in L^2(\mathbb{R}^\ell) \times L^2(V) \times X^\ell \times \mathbb{R}^\ell$ *satisfying the optimality system consisting of* (3.16) *and*

$$\begin{cases} -\dot{p}^{\mathcal{K}}(t) + \mathcal{A}p^{\mathcal{K}}(t) + \mathcal{N}'(y(t))^* p^{\mathcal{K}}(t) \\ \qquad = \sum_{i=1}^{\ell} \int_0^T 2\mathcal{I}^{-1} y(s,\cdot)\mu_i^{\mathcal{K}}(s)\,\mathrm{d}s\,\varphi_i(t) \\ \qquad + \sum_{i=1}^{\ell} \int_0^T \mathcal{I}^{-1}\, y(s,\cdot)\varphi_i(s)\,\mathrm{d}s\,\mu_i^{\mathcal{K}}(t) \\ p^{\mathcal{K}}(T) = 0, \end{cases}$$

$$\begin{cases} \eta_i = -\frac{1}{2}\,\langle \mathcal{G}_i(x, \psi_i, u, q), \psi_i\rangle_{X^*,X} \\ \mu_i^{\mathcal{K}} = -(\mathcal{K} - \lambda_i\, I)^{-1}(\mathcal{K}^{-1}\,\tilde{\mathcal{G}}_i + 2\eta_i\,\varphi_i), \end{cases}$$

$$\mathbf{R}\,u(t) = B^T(\psi)\,q(t) + \mathcal{B}^* p^{\mathcal{K}}(t),$$

where $\sqrt{\lambda_i}\,\psi_i = \mathcal{Y}^*\,\psi_i$ *and* $\tilde{\mathcal{G}}_i = \sqrt{\lambda_i}\,\mathcal{Y}^*\mathcal{I}\,\mathcal{G}_i(x, \psi_i, u, q)$.

3. POD a-posteriori error estimates

The main focus of this section is on an a-posteriori analysis for the method of proper orthogonal decomposition (POD) applied to optimal control problems governed by parabolic and elliptic PDEs. Based on a perturbation method it is deduced how far the suboptimal control, computed on the basis of the POD model, is from the (unknown) exact one. For more details and for the proofs we refer the reader to [**41**].

3.1. The linear-quadratic parabolic optimal control problem. Let V and H be real, separable Hilbert spaces and suppose that V is dense in H with compact embedding. By $\langle\cdot\,,\cdot\rangle_H$ we denote the inner product in H. The inner product in V is given by a symmetric bounded, coercive, bilinear form $a : V \times V \to \mathbb{R}$:

$$\langle\varphi, \psi\rangle_V = a(\varphi, \psi) \quad \text{for all } \varphi, \psi \in V$$

with associated norm $\|\cdot\|_V = \sqrt{a(\cdot\,,\cdot)}$. By identifying H and its dual H' it follows that $V \hookrightarrow H = H' \hookrightarrow V'$, each embedding being continuous and dense.

Recall that for $T > 0$ the space $W(0,T)$

$$W(0,T) = \big\{\varphi \in L^2(0,T;V) : \varphi_t \in L^2(0,T;V')\big\}$$

is a Hilbert space endowed with the common inner product (see, for example, [**3**, p. 473]). It is well-known that $W(0,T)$ is continuously embedded into $C([0,T];H)$, the space of continuous functions from $[0,T]$ to H.

Let $\mathfrak{I}$ be an open and bounded subset of $\mathbb{R}^d$ with $d \in \mathbb{N}$. By $U_{\mathsf{ad}} \subset L^2(\mathfrak{I})$ we define the closed, convex and bounded subset

$$U_{\mathsf{ad}} = \big\{u \in L^2(\mathfrak{I}) \,|\, u_a(s) \le u(s) \le u_b(s) \text{ for almost all (f.a.a.) } s \in \mathfrak{I}\big\}$$

with $u_a, u_b \in L^2(\mathfrak{I})$ satisfying $u_a \le u_b$ almost everywhere (a.e.) in $\mathfrak{I}$. For $y_0 \in H$, $r \in L^2(0,T;V')$ and $u \in U_{\mathsf{ad}}$ we consider the linear evolution problem

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\langle y(t), \varphi\rangle_H + a(y(t), \varphi) = \langle (r + \mathcal{B}u)(t), \varphi\rangle_{V',V} \qquad \text{f.a.a. } t \in [0,T],\ \forall \varphi \in V, \tag{3.17a}$$

$$\langle y(0), \varphi\rangle_H = \langle y_0, \varphi\rangle_H \qquad \forall \varphi \in V, \tag{3.17b}$$

where $\mathcal{B} : L^2(\mathfrak{I}) \to L^2(0,T;V')$ is a continuous, linear operator.

It is well-known (see, e.g., [**3**]) that for every $r \in L^2(0,T;V')$, $u \in L^2(\mathfrak{I})$ and $y_0 \in H$ there exists a unique weak solution $y \in W(0,T)$ satisfying (3.17) and

$$\|y\|_{W(0,T)} \le C\big(\|u\|_{L^2(\mathfrak{I})} + \|y_0\|_H + \|r\|_{L^2(0,T;V')}\big)$$

with a constant $C > 0$ independent of y.

REMARK 3.1. Let $\hat{y}_0 \in W(0,T)$ be the unique solution to

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\left\langle \hat{y}_0(t),\varphi\right\rangle_H + a(\hat{y}_0(t),\varphi) &= \langle r(t),\varphi\rangle_{V',V} && \text{f.a.a. } t\in[0,T],\ \forall\varphi\in V,\\ \left\langle \hat{y}_0(0),\varphi\right\rangle_H &= \langle y_0,\varphi\rangle_H && \forall\varphi\in V. \end{aligned}$$

Moreover, we introduce the linear and bounded operator $\mathcal{S}: L^2(\mathcal{I}) \to W(0,T)$ as follows: $\tilde{y} = \mathcal{S}u \in W(0,T)$ is the unique solution to

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\left\langle \tilde{y}(t),\varphi\right\rangle_H + a(\tilde{y}(t),\varphi) &= \langle (\mathcal{B}u)(t),\varphi\rangle_{V',V} && \text{f.a.a. } t\in[0,T],\ \forall\varphi\in V,\\ \left\langle \tilde{y}(0),\varphi\right\rangle_H &= 0 && \forall\varphi\in V. \end{aligned}$$

Then, $y = \hat{y}_0 + \mathcal{S}u$ is the weak solution to (3.17). ◊

Next we introduce the cost functional $J: W(0,T)\times L^2(\mathcal{I}) \to \mathbb{R}$ by

$$J(y,u) = \frac{\alpha_1}{2}\,\|\mathcal{C}y - z_1\|^2_{W_1} + \frac{\alpha_2}{2}\,\|\mathcal{D}y(T) - z_2\|^2_{W_2} + \frac{\sigma}{2}\,\|u\|^2_{L^2(\mathcal{I})},$$

where W_1, W_2 are Hilbert spaces, $\mathcal{C}: L^2(0,T;H)\to W_1$ and $\mathcal{D}: H\to W_2$ are bounded linear operators, and $(z_1,z_2)\in W_1\times W_2$ holds. Furthermore, α_1,α_2 are nonnegative parameters and $\sigma>0$.

The optimal control problem is given by

$$\min J(y,u) \quad \text{s.t.} \quad (y,u)\in W(0,T)\times U_{\mathsf{ad}} \text{ solves (3.17).} \tag{3.18}$$

Applying standard arguments (see [**28**], for instance) one can prove that there exists a unique optimal solution $\bar{x} = (\bar{y},\bar{u})$ to (3.18).

Suppose that $\bar{x} = (\bar{y},\bar{u})$ is the optimal solution to (3.18) (in the paper, a bar indicates optimality). Then there exists a unique Lagrange-multiplier $\bar{p}\in W(0,T)$ satisfying together with $\bar{x}$ the *first-order necessary optimality conditions*, which consist of the *state equations* (3.17), the *adjoint equations* in $[0,T]$

$$-\frac{\mathrm{d}}{\mathrm{d}t}\langle \bar{p}(t),\varphi\rangle_H + a(\bar{p}(t),\varphi) = \alpha_1\left\langle z_1 - \mathcal{C}\bar{y}, \mathcal{C}\varphi\right\rangle_{W_1} \text{ f.a.a. } t\in[0,T],\ \forall\varphi\in V, \tag{3.19a}$$

$$\langle \bar{p}(T),\varphi\rangle_H = \alpha_2\langle z_2 - \mathcal{D}\bar{y}(T), \mathcal{D}\varphi\rangle_{W_2} \quad \forall\varphi\in V, \tag{3.19b}$$

and of the *variational inequality*

$$\langle \sigma\bar{u} - \mathcal{B}^\star\bar{p}, u-\bar{u}\rangle_{L^2(\mathcal{I})} \geq 0 \quad \forall u\in\mathcal{U}_{\mathsf{ad}}. \tag{3.20}$$

Here, the linear and bounded operator $\mathcal{B}^\star: L^2(0,T;V)\to L^2(\mathcal{I})' \sim L^2(\mathcal{I})$ stands for the dual operator of $\mathcal{B}$.

Utilizing the solution operator $\mathcal{S}$ (see Remark 3.1) we introduce the so-called reduced cost functional as

$$\hat{J}(u) = J(\hat{y}_0 + \mathcal{S}u, u).$$

Then, we can express (3.18) as the reduced problem

$$\min \hat{J}(u) \quad \text{s.t.} \quad u\in U_{\mathsf{ad}}. \tag{3.21}$$

It follows that $\hat{J}'(\bar{u}) = \sigma\bar{u} - \mathcal{B}^\star\bar{p}\in L^2(\mathcal{I})$ is the gradient of $\hat{J}$ at $\bar{u}$, where $\bar{p}$ solves the dual sytem (3.19) for $\bar{y} = \hat{y}_0 + \mathcal{S}\bar{u}$. Moreover, the variational inequality (3.20) is equivalent to

$$\bar{u}(s) = \mathcal{P}_{[u_a(s),u_b(s)]}\left(\frac{1}{\sigma}\big(\mathcal{B}^\star\bar{p}\big)(s)\right) \quad \text{f.a.a. } s\in\mathcal{I}, \tag{3.22}$$

where $\mathcal{P}_{[a,b]}: \mathbb{R}\to[a,b]$ denotes the projection operator onto the convex interval $[a,b]\subset\mathbb{R}$.

3.2. A-posteriori error analysis. In principle, this section contains the main idea underlying our a-posteriori error analysis. Suppose that u_p is an arbitrary control of U_{ad}. Our goal is to estimate the difference

$$\|\bar{u} - u_p\|_{L^2(\mathcal{I})}$$

without the knowledge of the optimal solution $\bar{u}$.

If $u_p \neq \bar{u}$ then u_p does not satisfy the necessary (and by convexity sufficient) optimality conditions (3.20) respectively (3.22). However, there exists a function $\zeta \in L^2(\mathcal{I})$ such that

$$\langle \sigma u_p - \mathcal{B}^\star p_p + \zeta, u - u_p \rangle_{L^2(\mathcal{I})} \geq 0 \quad \forall u \in U_{ad}, \tag{3.23}$$

where $p_p \in W(0,T)$ solves the adjoint equation associated with u_p

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\langle p_p(t), \varphi \rangle_H + a(p_p(t), \varphi) &= \alpha_1 \langle z_1 - \mathcal{C}y_p, \mathcal{C}\varphi \rangle_{W_1} \text{ f.a.a. } t \in [0,T], \forall \varphi \in V, \\ \langle p_p(T), \varphi \rangle_H &= \alpha_2 \langle z_2 - \mathcal{D}y_p(T), \mathcal{D}\varphi \rangle_{W_2} \quad \forall \varphi \in V, \end{aligned} \tag{3.24}$$

and $y_p = \hat{y} + \mathcal{S}u_p$ is the state corresponding to u_p. Therefore, u_p satisfies the optimality condition of a perturbed parabolic optimal control problem with "perturbation" ζ. The smaller ζ is, the closer u_p is to $\bar{u}$. The computation of ζ is possible on the basis of the known data u_p, y_p, and p_p.

THEOREM 3.2. *Let $\bar{u}$ be the optimal solution to* (3.18)*, $\bar{y}$ the associated optimal state, and $\bar{p}$ the associated Lagrange multiplier. Suppose that $u_p \in U_{\mathsf{ad}}$ is chosen arbitrarily, $y_p = \hat{y} + \mathcal{S}u_p$, and p_p is the solution to* (3.24)*. Then it follows that*

$$\|\bar{u} - u_p\|_{L^2(\mathcal{I})} \leq \frac{1}{\sigma} \|\zeta\|_{L^2(\mathcal{I})},$$

where ζ is chosen such that (3.23) *holds.*

The function ζ satisfying (3.23) can be constructed from knowledge of u_p and the associated adjoint state p_p solving to (3.24).

PROPOSITION 3.3. *Suppose that the hypotheses of Theorem* 3.2 *are satisfied. Define $\zeta \in L^2(\mathcal{I})$ as follows:*

$$\zeta(s) = \begin{cases} \big[(\sigma u_p - \mathcal{B}^\star p_p)(s)\big]_- & \text{on } \mathcal{A}_- = \big\{s \in \mathcal{I} \,\big|\, u_p(s) = u_a(s)\big\}, \\ \big[(\sigma u_p - \mathcal{B}^\star p_p)(s)\big]_+ & \text{on } \mathcal{A}_+ = \big\{s \in \mathcal{I} \,\big|\, u_p(s) = u_b(s)\big\}, \\ -(\sigma u_p - \mathcal{B}^\star p_p)(s) & \text{on } \mathcal{J} = \mathcal{I} \setminus \big(\mathcal{A}_- \cup \mathcal{A}_+\big). \end{cases} \tag{3.25}$$

Then, the estimate

$$\|\bar{u} - u_p\|_{L^2(\mathcal{I})} \leq \frac{1}{\sigma} \|\zeta\|_{L^2(\mathcal{I})} \tag{3.26}$$

is satisfied.

We will call (3.26) an a-posteriori error estimate, since, in the next section, we shall apply it to suboptimal controls u_p that have already been computed from a POD model. After having computed u_p, we determine the associated state y_p and adjoint state (Lagrange multiplier) p_p. Then we can determine ζ and its L^2-norm and (3.26) gives an upper bound for the distance of u_p to $\bar{u}$. In this way, the error caused by the POD method can be estimated a-posteriorily. If the error is too large, then we have to include more POD basis functions in our Galerkin ansatz. This approach compensates the lack of a-priori error estimates for the POD method.

Next we turn to the POD approximation for (3.18). Let an arbitrary $u \in L^2(\mathcal{I})$ be chosen such that the corresponding state variable $y = \hat{y}_0 + \mathcal{S}u \in W(0,T)$ belongs to $C([0,T];V)$. Then,

$$\mathcal{V} = \operatorname{span}\big\{y(t) \,\big|\, t \in [0,T]\big\} \subseteq V. \tag{3.27}$$

If $y_0 \neq 0$ holds, then $\operatorname{span}\{y_0\} \subset \mathcal{V}$ and $d = \dim \mathcal{V} \geq 1$, but $\mathcal{V}$ may have infinite dimension. We define a bounded linear operator $\mathcal{Y} : L^2(0,T) \to V$ by

$$\mathcal{Y}\varphi = \int_0^T \varphi(t) y(t)\, \mathrm{d}t \quad \text{for } \varphi \in L^2(0,T).$$

Its Hilbert space adjoint $\mathcal{Y}^\star : V \to L^2(0,T)$ satisfying

$$\langle \mathcal{Y}\varphi, z\rangle_V = \langle \varphi, \mathcal{Y}^\star z\rangle_{L^2(0,T)} \quad \text{for } (\varphi, z) \in L^2(0,T) \times V$$

is given by

$$\big(\mathcal{Y}^\star z\big)(t) = \langle z, y(t)\rangle_V \quad \text{for } z \in V \text{ and f.a.a. } t \in [0,T].$$

The bounded linear operator $\mathcal{R} = \mathcal{Y}\mathcal{Y}^\star : V \to \mathcal{V} \subseteq V$ has the form

$$\mathcal{R}z = \int_0^T \langle z, y(t)\rangle_V \, y(t)\,\mathrm{d}t \quad \text{for } z \in V. \tag{3.28}$$

Moreover, let $\mathcal{K} = \mathcal{Y}^\star\mathcal{Y} : L^2(0,T) \to L^2(0,T)$ be defined by

$$\big(\mathcal{K}\varphi\big)(t) = \int_0^T \langle y(s), y(t)\rangle_V \, \varphi(s)\,\mathrm{d}s \quad \text{for } \varphi \in L^2(0,T).$$

The operator $\mathcal{K}$ is linear, bounded, self-adjoint, and compact. This implies that $\mathcal{R}$ is compact as well. Moreover, $\mathcal{R}$ is non-negative. From the Hilbert-Schmidt theorem [**36**, p. 203] it follows that there exists a complete orthonormal basis $\{\psi_i\}_{i=1}^d$ for $\mathcal{V} = \operatorname{range}(\mathcal{R})$ and a sequence $\{\lambda_i\}_{i=1}^d$ of real numbers such that

$$\mathcal{R}\psi_i = \lambda_i\psi_i \text{ for } i = 1,\dots,d \quad \text{and} \quad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0. \tag{3.29}$$

Remark 3.4. 1) To obtain a complete orthonormal basis in the separable Hilbert space V we need an orthonormal basis for $(\operatorname{range}(\mathcal{R}))^\perp$. This can be done by the Gram-Schmidt procedure. Hence, we suppose in the following that $\{\psi_i\}_{i=1}^\infty$ is a complete orthonormal basis for V.

2) Analogously to the theory of singular value decompositions for matrices, we find that the linear, bounded, compact and self-adjoint operator $\mathcal{K}$ has the same eigenvalues $\{\lambda_i\}_{i\in\mathbb{N}}$ as the operator $\mathcal{R}$. For all $\lambda_i > 0$ the corresponding eigenfunctions of $\mathcal{K}$ are given by

$$v_i(t) = \frac{1}{\sqrt{\lambda_i}}\,\big(\mathcal{Y}^*\psi_i\big)(t) = \frac{1}{\sqrt{\lambda_i}}\,\langle \psi_i, y(t)\rangle_V \text{ f.a.a. } t \in [0,T] \text{ and } 1 \leq i \leq \ell.$$

$\lozenge$

In the following proposition we formulate properties of the eigenvalues and eigenfunctions of $\mathcal{R}$. Therefore, for given $\ell \in \mathbb{N}$ we introduce the mapping

$$\mathfrak{J} : \underbrace{V \times \ldots \times V}_{\ell-\text{times}} \to \mathbb{R}, \quad \mathfrak{J}(\psi_1,\dots,\psi_\ell) := \int_0^T \Big\| y(t) - \sum_{i=1}^{\ell} \langle y(t), \psi_i\rangle_V \, \psi_i \Big\|_V^2 \,\mathrm{d}t.$$

Note that

$$\mathfrak{J}(\psi_1,\dots,\psi_\ell) = \int_0^T \Big\| y(t) - \mathcal{P}^\ell y(t) \Big\|_V^2 \,\mathrm{d}t. \tag{3.30}$$

Proposition 3.5. *Suppose that V is a separable Hilbert space, $y \in C([0,T];V)$ holds and $\mathcal{V}$ is given as in* (3.27)*. Let the linear operator $\mathcal{R} : V \to V$ be defined as in* (3.28)*. Then, $\mathcal{R}$ is bounded, self-adjoint, compact and non-negative, and there exists $\{\lambda_i\}_{i\in\mathbb{N}}$ and $\{\psi_i\}_{i\in\mathbb{N}}$ satisfying* (3.29)*. Moreover, for any $\ell \leq d = \dim\mathcal{V}$ the elements $\{\psi_i\}_{i=1}^\ell$ solve the minimization problem*

$$\min \mathfrak{J}(\tilde\psi_1,\dots,\tilde\psi_\ell) \quad \text{s.t.} \quad \langle \tilde\psi_j, \tilde\psi_i\rangle_V = \delta_{ij} \quad \text{for } 1 \leq i,j \leq \ell \tag{3.31}$$

and

$$\mathfrak{J}(\psi_1,\dots,\psi_\ell) = \sum_{i=\ell+1}^{\infty} \lambda_i.$$

For a proof we refer to [**12**, Section 3] and [**36**, Sections II and VI], for instance.

In real computations, we do not have the whole trajectory $y(t)$ for all $t \in [0,T]$. For that purpose let $0 = t_1 < t_2 < \ldots < t_n = T$ be a given grid in $[0,T]$ and let $y_j = y(t_j)$ denote approximations for y at time instance t_j, $j = 1, \ldots, n$. We set $\mathcal{V}^n = \operatorname{span}\{y_1, \ldots, y_n\}$ with $d^n = \dim \mathcal{V}^n \leq n$. Then, for given $\ell \leq n$ we consider the minimization problem

$$\min \sum_{j=1}^{n} \alpha_j \left\| y_j - \sum_{i=1}^{\ell} \langle y_j, \psi_i^n \rangle_V \, \psi_i^n \right\|_V^2 \text{ s.t. } \langle \psi_i^n, \psi_j^n \rangle_V = \delta_{ij} \text{ for } 1 \leq i,j \leq \ell \tag{3.32}$$

instead of (3.31). In (3.32) the α_j's stand for the trapezoidal weights

$$\alpha_1 = \frac{t_2 - t_1}{2}, \quad \alpha_j = \frac{t_{j+1} - t_{j-1}}{2} \text{ for } 2 \leq j \leq n-1, \quad \alpha_n = \frac{t_n - t_{n-1}}{2}.$$

The solution to (3.32) is given by the solution to the eigenvalue problem

$$\mathcal{R}^n \psi_i^n = \sum_{j=1}^{n} \alpha_j \, \langle y_j, \psi_i^n \rangle_V \, y_j = \lambda_i^n \psi_i^n, \quad i = 1, \ldots, \ell,$$

where $\mathcal{R}^n : V \to \mathcal{V}^n \subset V$ is a linear, bounded, compact, self-adjoint and non-negative operator. Thus, there exists an orthonormal set $\{\psi_i^n\}_{i=1}^{d^n}$ of eigenfunctions and corresponding non-negative eigenvalues $\{\lambda_i^n\}_{i=1}^{d^n}$ satisfying

$$\mathcal{R}^n \psi_i^n = \lambda_i^n \psi_i^n, \quad \lambda_1^n \geq \lambda_2^n \geq \ldots \geq \lambda_{d^n}^n > 0. \tag{3.33}$$

Let $y = \hat{y}_0 + \mathcal{S}u$ be the state associated with some control $u \in L^2(\mathcal{I})$, and let $\mathcal{V}$ be given as in (3.27). We fix ℓ with $\ell \leq \dim \mathcal{V}$ and compute the first ℓ POD basis functions $\psi_1, \ldots, \psi_\ell \in V$ by solving either $\mathcal{R}\psi_i = \lambda_i \psi_i$ or $\mathcal{K} v_i = \lambda v_i$ for $i = 1, \ldots, \ell$ (see Remark 3.4). Then we define the finite dimensional linear space

$$V^\ell = \operatorname{span}\{\psi_1, \ldots, \psi_\ell\} \subset V.$$

Endowed with the topology in V it follows that V^ℓ is a Hilbert space. Let $\mathcal{P}^\ell$ denote the orthogonal projection $\mathcal{P}^\ell$ of V onto V^ℓ defined by

$$\mathcal{P}^\ell \varphi = \sum_{i=1}^{\ell} \langle \varphi, \psi_i \rangle_V \, \psi_i \quad \text{for } \varphi \in V. \tag{3.34}$$

Combining (3.30) and (3.31) we obtain that

$$\mathfrak{J}(\psi_1, \ldots, \psi_\ell) = \int_0^T \left\| y(t) - \mathcal{P}^\ell y(t) \right\|_V^2 \mathrm{d}t = \left\| y - \mathcal{P}^\ell y \right\|_{L^2(0,T;V)}^2 = \sum_{i=\ell+1}^{\infty} \lambda_i.$$

The POD Galerkin scheme for the state equation (3.17) leads to the following linear problem: determine a function $y^\ell = \sum_{i=1}^{\ell} y_i(t) \psi_i$ such that

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle y^\ell(t), \psi \rangle_H + a(y^\ell(t), \psi) = \langle (r + \mathcal{B}u)(t), \psi \rangle_{V',V} \text{ f.a.a. } t \in [0,T], \forall \psi \in V^\ell, \tag{3.35a}$$

$$\langle y^\ell(0), \psi \rangle_H = \langle y_0, \psi \rangle_H \quad \forall \psi \in V^\ell. \tag{3.35b}$$

For every $r \in L^2(0,T;V')$, $u \in L^2(\mathcal{I})$, $y_0 \in H$ and for every $\ell \in \mathbb{N}$ problem (3.35) admits a unique solution $y^\ell \in H^1(0,T;V^\ell)$; see [**11**, Proposition 3.4]. From $V^\ell \hookrightarrow V$ it follows that $y^\ell \in W(0,T)$ holds.

Let $\hat{y}_0^\ell \in H^1(0,T;V^\ell)$ be the solution to (3.35) for $u \equiv 0$. Analogously to Remark 3.1 we introduce the linear operator $\mathcal{S}^\ell : L^2(\mathcal{I}) \to H^1(0,T;V^\ell)$ for fixed ℓ: For given $u \in L^2(\mathcal{I})$ the element $\tilde{y}^\ell = \mathcal{S}^\ell u$ solves (3.35) with $r \equiv 0$ and $y_0 \equiv 0$. Thus, y^ℓ is given by $y^\ell = \hat{y}_0^\ell + \tilde{y}^\ell$. It follows from [**11**, Proposition 3.4] that the operator $\mathcal{S}^\ell$ is bounded independently of ℓ.

PROPOSITION 3.6. *For given $r \in L^2(0,T;V')$, $u \in L^2(\mathcal{I})$, and $y_0 \in H$ we suppose that $y = \hat{y} + \mathcal{S}u$ belongs to $y \in C([0,T];V)$. Suppose that, for $\ell \leq \dim \mathcal{V}$, the elements $\{\psi_i\}_{i=1}^{\ell}$ solve* (3.31). *Then, there exists a constant $C > 0$ such that*

$$\|y - y^\ell\|^2_{W(0,T)} \leq C\Big(\big\|y^\ell(0) - \mathcal{P}^\ell y_0\big\|^2_H + \big\|y_t - \mathcal{P}^\ell y_t\big\|^2_{L^2(0,T;V')} + \sum_{i=\ell+1}^{\infty} \lambda_i \Big),$$

where the linear projector $\mathcal{P}^\ell : V \to V^\ell$ is given by (3.34) *and $y^\ell = \hat{y}_0^\ell + \mathcal{S}^\ell u$ denotes the unique solution to* (3.35).

Proposition 3.6 permits to show that the POD approximations y^ℓ converge to y in the $W(0,T)$-norm:

PROPOSITION 3.7. *For given $r \in L^2(0,T;V')$, $u \in L^2(\mathcal{I})$, and $y_0 \in V$ we suppose that $y = \hat{y} + \mathcal{S}u$ belongs to $y \in H^1(0,T;V)$. Suppose that, for $\ell \leq \dim \mathcal{V}$, the elements $\{\psi_i\}_{i=1}^{\ell}$ solve* (3.31). *Then, it follows that*

$$\lim_{\ell \to \infty} \big\|y - y^\ell\big\|_{W(0,T)} = 0,$$

where $y^\ell = \hat{y}_0^\ell + \mathcal{S}^\ell u$ denotes the unique solution to (3.35).

REMARK 3.8.
1) Due to the continuous embedding of $W(0,T)$ into the space $C([0,T];H)$, Proposition 3.7 implies $y^\ell \to y$ in $C([0,T];H)$ as $\ell \to \infty$. In particular, $y^\ell(T)$ converges to $y(T)$ in H as ℓ tends to ∞.
2) Let us mention that the convergence result in Proposition 3.7 is true for any fixed u provided that the system $\{\psi_i\}_{i=1}^{\infty}$ computed from the snapshots associated with u is complete. ◊

We turn to the POD Galerkin scheme for the adjoint system (3.19a). For that purpose let $u \in L^2(\mathcal{I})$ be arbitrarily given, $\{\psi_1, \ldots, \psi_\ell\}$ the associated POD basis of rank ℓ, and let $y^\ell \in H^1(0,T;V^\ell)$ denote the unique solution to (3.35). Then, $p^\ell = \sum_{i=1}^{\ell} p_i(t)\psi_i$ satisfies the linear system

$$-\frac{\mathrm{d}}{\mathrm{d}t}\langle p^\ell(t), \psi\rangle_H + a(p^\ell(t), \psi) = \alpha_1 \langle z_1 - \mathcal{C}y^\ell, \mathcal{C}\psi\rangle_{W_1} \text{ f.a.a. } t \in [0,T], \forall \psi \in V^\ell, \tag{3.36a}$$

$$\langle p^\ell(T), \psi\rangle_H = \alpha_2 \langle z_2 - \mathcal{D}y^\ell(T), \mathcal{D}\psi\rangle_{W_2} \quad \forall \psi \in V^\ell. \tag{3.36b}$$

PROPOSITION 3.9. *For given $r \in L^2(0,T;V')$, $u \in L^2(\mathcal{I})$, $y_0 \in H$ suppose that $y = \hat{y} + \mathcal{S}u$ belongs to $H^1(0,T;V)$. Suppose that for $\ell \leq \dim \mathcal{V}$ the elements $\{\psi_i\}_{i=1}^{\ell}$ solve* (3.31). *Let $y^\ell = \hat{y}_0^\ell + \mathcal{S}^\ell u$, p, and p^ℓ be the solutions to* (3.35), (3.19) *and* (3.36), *respectively. Then there exists a constant $C > 0$ depending on α_1, α_2, $\mathcal{C}$, and $\mathcal{D}$*

$$\begin{aligned}\big\|p - p^\ell\big\|_{L^2(0,T;V)} \leq\ & C\Big(\big\|p(T) - \mathcal{P}^\ell p(T)\big\|_H + \|p - \mathcal{P}^\ell p\|_{W(0,T)} \Big) \\ & + C\Big(\big\|y(T) - y^\ell(T)\big\|_H + \big\|y - y^\ell\big\|_{L^2(0,T;H)} \Big).\end{aligned}$$

where the linear projector $\mathcal{P}^\ell : V \to V^\ell$ is given by (3.34). *If, in addition, $y_0 \in V$ and $p \in H^1(0,T;V)$ hold, then $\lim_{\ell \to \infty} \|p - p^\ell\|_{L^2(0,T;V)} = 0$ holds.*

REMARK 3.10. Arguing as in Remark 3.8-2) we derive that the convergence result of Proposition 3.9 remains true if the POD basis is computed using an input $\tilde{u} \in L^2(\mathcal{I})$ that differs from u. Of course, the convergence rate of p^ℓ to p as $\ell \to \infty$ depends on the approximation properties of the POD basis for the adjoint variable p; see [**4, 11**]. ◊

The Galerkin projection of (3.21) leads to the discretized optimal control problem

$$\min \hat{J}^\ell(u) \quad \text{s.t.} \quad u \in U_{\mathsf{ad}}, \tag{3.37}$$

where $\hat{J}^\ell(u) = J(y^\ell(u), u)$ is the reduced objective function and $y^\ell(u)$ denotes the solution to (3.35) associated with $u \in U_{\mathsf{ad}}$. We call (3.37) a reduced-order model for (3.21).

Problem (3.37) admits a unique optimal solution $\bar{u}^\ell$ that is interpreted as a suboptimal solution to (3.21). First-order necessary optimality conditions for (3.37) are given by

$$\langle \sigma \bar{u}^\ell - \mathcal{B}^\star \bar{p}^\ell, u - \bar{u}^\ell \rangle_{L^2(\mathfrak{I})} \geq 0 \quad \text{for all } u \in U_{\mathsf{ad}},$$

where, $\bar{y}^\ell \in H^1(0,T;V^\ell)$ denotes the optimal state solving (3.35) with $u = \bar{u}$ and $\bar{p}^\ell \in H^1(0,T;V^\ell)$ is the adjoint state for the POD model.

We proceed similarly as in [**11**, Section 4]. However, an essential difference is that we derive convergence results utilizing a POD basis of rank ℓ that is not necessarily related to the optimal control $\bar{u}$ as an input function for the generation of the snapshots.

PROPOSITION 3.11. *Suppose that the POD basis of rank ℓ is computed using an arbitrarily chosen $u \in L^2(\mathfrak{I})$. Let $\bar{u}$ and $\bar{u}^\ell$ be the optimal solutions to* (3.21) *and* (3.37)*, respectively. Moreover, $\bar{p} \in W(0,T)$ denotes the adjoint state associated with $\bar{u}$. Then,*

$$\|\bar{u} - \bar{u}^\ell\|_{L^2(\mathfrak{I})} \leq c \, \|\bar{p} - \hat{p}^\ell\|_{L^2(0,T;V)},$$

where $\hat{p}^\ell$ solves

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\langle \hat{p}^\ell(t), \psi \rangle_H + a(\hat{p}^\ell(t), \psi) &= \alpha_1 \langle z_1 - \mathcal{C}\hat{y}^\ell, \mathcal{C}\psi \rangle_{W_1} \text{ f.a.a. } t \in [0,T], \forall \psi \in V^\ell, \\ \langle \hat{p}^\ell(T), \psi \rangle_H &= \alpha_2 \langle z_2 - \mathcal{D}\hat{y}^\ell(T), \mathcal{D}\psi \rangle_{W_2} \quad \forall \psi \in V^\ell \end{aligned}$$

and $\hat{y}^\ell$ is the solution to

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\langle \hat{y}^\ell(t), \psi \rangle_H + a(\hat{y}^\ell(t), \psi) &= \langle (r + \mathcal{B}\bar{u})(t), \psi \rangle_{V',V} \text{ f.a.a. } t \in [0,T], \forall \psi \in V^\ell, \\ \langle \hat{y}^\ell(0), \psi \rangle_H &= \langle y_0, \psi \rangle_H \quad \forall \psi \in V^\ell. \end{aligned}$$

Notice that $\hat{p}^\ell$ is the POD-approximate associated with $\hat{y}^\ell$ and $\hat{y}^\ell = \hat{y}_0^\ell + \mathcal{S}^\ell \bar{u}$. Therefore, both $\hat{y}^\ell$ and $\hat{p}^\ell$ are associated with the same optimal control $\bar{u}$ so that we can apply Proposition 3.6 and Proposition 3.9 to estimate the difference $\bar{y} - \hat{y}^\ell$ and $\bar{p} - \hat{p}^\ell$, respectively. In contrast to this, $\bar{y}^\ell = \hat{y}_0^\ell + \mathcal{S}^\ell \bar{u}^\ell$ corresponds to the suboptimal control $\bar{u}^\ell$, which we estimate in the next theorem.

THEOREM 3.12. *Suppose that the POD basis of rank ℓ is computed using an arbitrarily chosen $u \in L^2(\mathfrak{I})$. Let $\bar{u}$ and $\bar{u}^\ell$ be the optimal solutions to* (3.21) *and* (3.37)*, respectively. Moreover, let $\bar{y}$ and $\bar{p}$ denote the optimal state and adjoint, respectively, associated with $\bar{u}$. Then there exists a constant $C > 0$ not depending on ℓ such that*

$$\begin{aligned} &\left\|\bar{u} - \bar{u}^\ell\right\|_{L^2(\mathfrak{I})} \\ &\qquad \leq C\Big(\left\|\bar{y} - \mathcal{P}^\ell \bar{y}\right\|_{W(0,T)} + \left\|\bar{y}^\ell(0) - \mathcal{P}^\ell y_0\right\|_H + \left\|\bar{p} - \mathcal{P}^\ell \bar{p}\right\|_{W(0,T)} \Big), \end{aligned} \tag{3.38}$$

where the linear projector $\mathcal{P}^\ell : V \to V^\ell$ is given in (3.34).

If, in addition, $y_\circ \in V$ and $\bar{y}$, $\bar{p} \in H^1(0,T;V)$ hold and $\{\psi_i\}_{i=1}^\infty$ is a complete orthonormal basis for V, then

$$\lim_{\ell \to \infty} \left\|\bar{u} - \bar{u}^\ell\right\|_{L^2(\mathfrak{I})} = 0.$$

REMARK 3.13. Let us consider the following idealized situation [**11**]: Let $\bar{u}$ be the optimal solution to (3.21). Moreover, let $\bar{y}$, $\bar{p} \in H^1(0,T;V)$ denote the optimal state and adjoint state, respectively, associated with $\bar{u}$ and let $y_0 \in V$. Then we consider the minimization problem

$$\min_{\psi_1,\ldots,\psi_\ell} \|\bar{y} - \mathcal{P}^\ell \bar{y}\|^2_{H^1(0,T;V)} + \|\bar{p} - \mathcal{P}^\ell \bar{p}\|^2_{H^1(0,T;V)} \text{ s.t. } \langle \psi_i, \psi_j \rangle_V = \delta_{ij}, \ 1 \leq i,j \leq \ell.$$

Its solution $\{\bar{\psi}_i\}_{i=1}^\ell$ of rank ℓ satisfies the eigenvalue problem

$$\bar{\mathcal{R}}\bar{\psi}_i = \bar{\lambda}_i \bar{\psi}_i, \quad 1 \leq i \leq \ell,$$

where the linear, bounded, non-negative and self-adjoint operator $\bar{\mathcal{R}}$ is defined as

$$\bar{\mathcal{R}}z = \int_0^T \langle \bar{y}(t), z\rangle_V \, y(t) + \langle \bar{y}_t(t), z\rangle_V \, \bar{y}_t(t) + \langle \bar{p}(t), z\rangle_V \, \bar{p}(t) + \langle \bar{p}_t(t), z\rangle_V \, \bar{p}_t(t)\,\mathrm{d}t$$

for $z \in V$. Then, (3.38) can be replaced by

$$\left\|\bar{u} - \bar{u}^\ell\right\|^2_{L^2(\mathcal{I})} \le \bar{C}\left(\left\|\bar{y}^\ell(0) - \mathcal{P}^\ell y_0\right\|^2_H + \sum_{i=\ell+1}^{\infty} \bar{\lambda}_i\right)$$

with a constant $\bar{C} > 0$. Now we can estimate the decay of the norms $\|\bar{y} - \mathcal{P}^\ell \bar{y}\|_{W(0,T)}$ and $\|\bar{p} - \mathcal{P}^\ell \bar{p}\|_{W(0,T)}$ in (3.38) in terms of the eigenvalues $\bar{\lambda}_i$ and obtain an error estimate with respect to the remainder $\sum_{i=\ell+1}^{\infty} \bar{\lambda}_i$. In contrast to this, the decay of the eigenvalues λ_i can only be used to bound $\|\bar{y} - \mathcal{P}^\ell \bar{y}\|_{L^2(0,T;V)}$ from above, but not the expression $\|\bar{y}_t - \mathcal{P}^\ell \bar{y}_t\|_{L^2(0,T;V')} + \|\bar{p} - \mathcal{P}^\ell \bar{p}\|_{W(0,T)}$. ◊

Now we complete the discussion of the a-posteriori estimate by combining Theorem 3.12 and Proposition 3.3. The proposition permits to estimate $\|\bar{u} - \bar{u}^\ell\|$ by the norm of an appropriate ζ, while Theorem 3.12 will be used to show that ζ tends to zero as $\ell \to \infty$, since it ensures convergence of $\bar{u}^\ell$ to the optimal solution $\bar{u}$ of (3.21).

For any ℓ let $\bar{u}^\ell \in U_{\mathsf{ad}}$ be the optimal solution to (3.37). This optimal $\bar{u}^\ell$ is taken as a suboptimal u_p for (3.21), i.e. in Proposition 3.3 we take $u_p := \bar{u}^\ell$.

THEOREM 3.14. 1) *Let $\ell \le d$ be arbitrarily given and $\bar{u}^\ell \in U_{\mathsf{ad}}$ be the optimal solution to* (3.37). *Denote by $\tilde{y} = \tilde{y}(\bar{u}^\ell) = \hat{y}_0 + \mathcal{S}\bar{u}^\ell$ the solution to* (3.17) *with $u = \bar{u}^\ell$ and let $\tilde{p} = \tilde{p}(\bar{u}^\ell)$ solve the associated adjoint equation*

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\langle \tilde{p}(t), \varphi\rangle_H + a(\tilde{p}(t), \varphi) &= \alpha_1 \langle z_1 - \mathcal{C}\tilde{y}, \mathcal{C}\varphi\rangle_{W_1} \text{ f.a.a. } t \in [0,T],\ \forall \varphi \in V, \\ \langle \tilde{p}(T), \varphi\rangle_H &= \alpha_2 \langle z_2 - \mathcal{D}\tilde{y}(T), \mathcal{D}\varphi\rangle_{W_2} \quad \forall \varphi \in V. \end{aligned} \tag{3.39}$$

Define, according to (3.25), *the function $\zeta^\ell \in L^2(\mathcal{I})$ by*

$$\zeta^\ell(s) = \begin{cases} \left[(\sigma\bar{u}^\ell - \mathcal{B}^\star\tilde{p}(\bar{u}^\ell))(s)\right]_- & \text{on } \mathcal{A}^\ell_- = \left\{s \in \mathcal{I} \,\middle|\, \bar{u}^\ell(s) = u_a(s)\right\}, \\ \left[(\sigma\bar{u}^\ell - \mathcal{B}^\star\tilde{p}(\bar{u}^\ell))(s)\right]_+ & \text{on } \mathcal{A}^\ell_+ = \left\{s \in \mathcal{I} \,\middle|\, \bar{u}^\ell(s) = u_b(s)\right\}, \\ -(\sigma\bar{u}^\ell - \mathcal{B}^\star\tilde{p}^\ell(\bar{u}^\ell))(s) & \text{on } \mathcal{I}^\ell = \mathcal{I} \setminus (\mathcal{A}^\ell_- \cup \mathcal{A}^\ell_+). \end{cases} \tag{3.40}$$

Then

$$\|\bar{u} - \bar{u}^\ell\|_{L^2(\mathcal{I})} \le \frac{1}{\sigma}\,\|\zeta^\ell\|_{L^2(\mathcal{I})}.$$

2) *If all hypotheses of Proposition* 3.9 *and Theorem* 3.12 *are satisfied, in particular $\{\psi_i\}_{i=1}^\infty$ is a complete orthonormal basis for V, then the sequences $\{\bar{u}^\ell\}_{\ell\in\mathbb{N}}$ and $\{\mathcal{B}^\star\hat{p}^\ell\}_{\ell\in\mathbb{N}}$ converge to $\bar{u}$ respectively $\mathcal{B}^\star\bar{p}$ in $L^2(\mathcal{I})$ as $\ell \to \infty$ and*

$$\|\zeta^\ell\|_{L^2(\mathcal{I})} \to 0.$$

REMARK 3.15. 1) Notice that $\tilde{y}$ and $\tilde{p}$ must be taken as the solutions to the (full) state and adjoint equation, respectively, not of their POD-approximations.

2) Part 2) of Theorem 3.14 shows that $\|\zeta^\ell\|_{L^2(\mathcal{I})}$ can be expected smaller than any $\varepsilon > 0$ provided that ℓ is taken sufficiently large. Motivated by this result, we set up the Algorithm 1. ◊

REMARK 3.16. In the numerical realization of Algorithm 1, Step 6 requires the solution of the state as well as of the adjoint equation by, e.g., a finite element or finite differerence scheme. ◊

Algorithm 1 POD reduced-order method with a-posteriori estimator.

Choose an input $u \in U_{\mathsf{ad}}$, an initial number ℓ for POD ansatz functions, a maximal number $\ell^{\max} > \ell$ of POD ansatz functions, and a stopping tolerance $\varepsilon > 0$; compute $y = \hat{y}_0 + \mathcal{S}u$.
Determine a POD basis of rank ℓ utilizing the state $y = \hat{y}_0 + \mathcal{S}u$ and derive the reduced-order model (3.37).
repeat
  Establish the discretized optimal control problem (3.37).
  Calculate the optimal solution $\bar{u}^\ell$ of (3.37).
  Evaluate $\tilde{y}(\bar{u}^\ell) = \hat{y}_0 + \mathcal{S}\bar{u}^\ell$ and compute the solution $\tilde{p}(\bar{u}^\ell)$ to (3.39) as well as ζ^ℓ from (3.40).
  if $\|\zeta^\ell\|_{L^2(\mathcal{J})} < \varepsilon$ **or** $\ell = \ell^{\max}$ **then**
    Return ℓ and suboptimal control $\bar{u}^\ell$ and STOP.
  else
    Set $\ell = \ell + 1$.
  end if
until $\ell > \ell^{\max}$

CHAPTER 4

Further topics

1. Parameter identification

1.1. Galerkin proper orthogonal decomposition methods for parameter dependent elliptic systems. In [**16**] estimates for Galerkin POD methods for linear elliptic, parameter-dependent systems are proved. The resulting error bounds depend on the number of POD basis functions and on the parameter grid that is used to generate the snapshots and to compute the POD basis. The error estimates also holds for semi-linear elliptic problems with monotone nonlinearity. Numerical examples are included.

1.2. Impedance identification. The acoustical impedance of a component or trim part is one of its most important characteristics. The trim and its absorption behavior contributes significantly to the comfort inside the car. Therefore, correct impedance values are needed when acoustical simulations of car interior noise are carried out.

A generally used methodology to determine the acoustical impedance is to use cut-out round samples of the material in question and measure the acoustic characteristic in the impedance tube. As a result values for the normal impedance and absorption coefficients can be obtained for this material. Disadvantages of this method are that the measurement considers normal acoustic waves, only, that some materials are inappropriate for the impedance tube and that the effects of the shape of the whole part have to be neglected. Therefore efforts have been made to develop methods for impedance mearurements of entire trim parts, such as carpets, dashboards or seats.

In [**44**] we formulate the identification problem as an optimal control problem, where the cost functional contains a regularization term as well as a least-squares term for the difference of the mearurements and the sound pressure p computed by solving the Helmholtz equation. In contrast to [**10**] we identify the admittance $A \in \mathbb{C}$ instead of the impedance $Z = 1/A$. Due to the the term Ap in the Robin boundary conditions for the Helmholtz equation (normal impedance boundary) the obtained optimal control problem has a bilinear structure, whereas in [**10**] the non-linearity is of the form p/Z. If the admittance A has been estimated, then $Z = 1/A$ is an estimate for the impedance. The optimal control problem is solved by a globalized quasi-Newton method with BFGS update of the Hessian [**34**]. Furthermore, a discretization based on proper orthogonal decomposition (POD) is utilized for the solution of the Helmholtz equation. POD is a powerful technique for model reduction of nonlinear systems.

Let us mention that in [**10**] a standard finite element discretization for the Helmholtz equation is applied. Alternatively, the wave based technique (WBT) is used in [**9**]. A-posteriori analysis in utilized in [**41**] to determine the number of POD ansatz functions in the POD Galerkin projection for an optimal control problem governed by the Helmholtz equation.

1.3. Estimation of regularization parameters in elliptic optimal control problems. In [**17**] parameter estimation problems for a non-linear elliptic problem are considered. Using a Tikhonov regularization techniques the identification problems are formulated in terms of optimal control problems which are solved numerically by an augmented Lagrangian method combined with a globalized sequential quadratic programming algorithm. For the discretization of the partial differential equations a Galerkin scheme based on proper orthogonal decomposition (POD) is utilized, which leads to a fast optimization solver. This method is utilized in a bilevel optimization problem to determine the parameters for the Tikhonov regularization. Numerical examples illustrate the efficiency of the proposed approach.

1.4. Estimation of diffusion coefficients in a scalar Ginzburg-Landau equation. In [**18**] work POD is applied to estimate scalar parameters in a scalar non-linear Ginzburg-Landau equation. The parameter estimation is formulated in terms of an optimal control problem that is solved by an augmented Lagrangian method combined with a sequential quadratic programming algorithm. A numerical example illustrates the efficiency of the proposed solution method.

2. Feedback strategies

2.1. Reduced order output feedback control design for PDE systems. The design of an optimal (output feedback) reduced order control (ROC) law for a dynamic control system is an important example of a difficult and in general non–convex (nonlinear) optimal control problem. In [**27**] we present a novel numerical strategy to the solution of the ROC design problem if the control system is described by partial differential equations (PDE). The discretization of the ROC problem with PDE constraints leads to a large scale (non–convex) nonlinear semidefinite program (NSDP). For reducing the size of the high dimensional control system, first, we apply a POD method to the discretized PDE. The POD approach leads to a low dimensional model of the control system. Thereafter, we solve the corresponding small–sized NSDP by a fully iterative interior point constraint trust region (IPCTR) algorithm. IPCTR is designed to take advantage of the special structure of the NSDP. Finally, the solution is a ROC for the low dimensional approximation of the control system. In our numerical examples we demonstrate that the reduced order controller computed from the small scaled problem can be used to control the large scale approximation of the PDE system.

2.2. HJB-POD based feedback design for the optimal control of evolution problems. The numerical realization of closed loop control for distributed parameter systems is still a significant challenge and in fact infeasible unless specific structural techniques are employed. In [**25**] we propose the combination of model reduction techniques based on POD with the numerical treatment of the Hamilton-Jacobi-Bellman (HJB) equation for infinite horizon optimal control problems by a modification of an algorithm originated by Gonzales-Rofman and further developed by Falcone-Ferretti. The feasibility of the proposed methodology is demonstrated numerically by means of optimal boundary feedback-control for the Burgers equation with noise in the initial condition and in the forcing function.

Bibliography

[1] H. W. Alt. *Lineare Funktionalanalysis. Eine anwendungsorientierte Einführung*. Springer-Verlag, Berlin, 1992.

[2] M. Barrault, Y. Maday, N.C. Nguyen, and A.T. Patera. An empirical interpolation method: application to efficient reduced-basis discretization of partial differential equations. *Comptes Rendus de'l Académie des Sciences Paris*, Ser. I 339:667-672, 2004.

[3] R. Dautray and J.-L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology. Volume 5: Evolution Problems I*. Springer-Verlag, Berlin, 1992.

[4] F. Diwoky and S. Volkwein. Nonlinear boundary control for the heat equation utilizing proper orthogonal decomposition. *International Series of Numerical Mathematics*, 138:73-87, 2001.

[5] L.C. Evans. *Partial Differential Equations*. Graduate Studies in Mathematics, vol. 19, American Mathematical Society, Providence, Rhode Island, (1998).

[6] M.A. Grepl, Y. Maday, N.C. Nguyen, and A.T. Paterra. Efficient reduced-basis treatment of non-affine and nonlinear partial differential equations. *Mat. Mod. Numer. Anal.*, to appear.

[7] M. D. Gunzburger, L. Hou, and T. P. Svobodny. Finite element approximations of an optimal control problem associated with the scalar Ginzburg-Landau equation. *Computers Math. Applic.*, 21:123-131, 1991.

[8] T. Henri and M. Yvon. Convergence estimates of POD Galerkin methods for parabolic problems. Technical Report No. 02-48, Institute of Mathematical Research of Rennes, 2002.

[9] A. Hepberger. *Mathematical methods for the prediction of the interior car noise in the middle frequency range*. PhD tesis, TU Graz, Institute for Mathematics, Austria, 2002.

[10] A. Hepberger, S. Volkwein, F. Diwoky, and H.-H. Priebsch. Impedance identification out of pressure datas with a hybrid measurement-simulation methodology up to 1kHz. In *Proceedings of International Conference on Noise and Vibration Engineering*, Leuven, Belgium, 2006.

[11] M. Hinze and S. Volkwein. Error estimates for abstract linear-quadratic optimal control problems using proper orthogonal decomposition. *Computational Optimization and Applications*, to appear.

[12] P. Holmes, J.L. Lumley, and G. Berkooz. *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge Monographs on Mechanics, Cambridge University Press, 1996.

[13] K. Ito and S.S. Ravindran. A reduced basis method for control problems governed by PDEs. In: Desch, W., Kappel, F., Kunisch, K. (ed), *Control and Estimation of Distributed Parameter Systems*. Proceedings of the International Conference in Vorau, 1996, 153-168, 1998.

[14] M. Kahlbacher. *POD for parameter estimation of bilinear elliptic problems*. Diploma thesis, Institute for Mathematics and Scientific Computing, University of Graz, October 2006.

[15] M. Kahlbacher and S. Volkwein. Galerkin proper orthogonal decomposition methods for parameter dependent elliptic systems. *Discussiones Mathematicae: Differential Inclusions, Control and Optimization*, 27:95-117, 2007.

[16] M. Kahlbacher and S. Volkwein. Model reduction by proper orthogonal decomposition for estimation of scalar parameters in elliptic PDEs. In Proceedings of *ECCOMAS CFD*, P. Wesseling, E. Onate, and J. Periaux (eds.), Egmont aan Zee, 2006.

[17] M. Kahlbacher and S. Volkwein. Estimation of regularization parameters in elliptic optimal control problems by POD model reduction. Submitted, 2008.

[18] M. Kahlbacher and S. Volkwein. Estimation of diffusion coefficients in a scalar Ginzburg-Landau equation by using model reduction. Submitted, 2007.

[19] T. Kato. *Perturbation Theory for Linear Operators*. Springer-Verlag, Berlin, 1980.

[20] 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, 345–371, 1999.

[21] K. Kunisch and S. Volkwein. Galerkin proper orthogonal decomposition methods for parabolic problems. *Numerische Mathematik*, 90:117-148, 2001.

[22] K. Kunisch and S. Volkwein. *Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics. SIAM J. Numer. Anal.*, 40:492-515, 2002.

[23] K. Kunisch and S. Volkwein. Crank-Nicolson Galerkin proper orthogonal decomposition approximations for a general equation in fluid dynamics. Proceedings of the 18th GAMM Seminar on *Multigrid and related methods for optimization problems*, Leipzig, 97–114, 2002.

[24] K. Kunisch and S. Volkwein. Proper orthogonal decomposition for optimality systems. *ESAIM: Mathemematical Modelling and Numerical Analysis*, 42:1-23, 2008.

[25] K. Kunisch, S. Volkwein, and X. Lie. HJB-POD based feedback design for the optimal control of evolution problems. *SIAM J. on Applied Dynamical Systems*, 3:701-722, 2004.

[26] S. Lall, J.E. Marsden and S. Glavaski. Empirical model reduction of controlled nonlinear systems. In: *Proceedings of the IFAC Congress*, vol. F, 473-478, 1999.

[27] 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*, 415:542-757, 2006.

[28] J.L. Lions. *Optimal Control of Systems Governed by Partial Differential Equations*. Springer, Berlin, 1971.

[29] H.V. Ly and H.T. Tran. Modelling and control of physical processes using proper orthogonal decomposition. *Mathematical and Computer Modeling*, 33:223-236, 2001.

[30] L. Machiels, Y. Maday, and A.T. Patera. Output bounds for reduced-order approximations of elliptic partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 190:3413-3426, 2001.

[31] Y. Maday and E.M. Rønquist. A reduced-basis element method. *Journal of Scientific Computing*, 17, 1-4, 2002.

[32] B.C. Moore. Principal component analyisin linear systems: controllability, observability, and model reduction. *IEEE Trans. Automat. Contr.*, 26:17-32, 1981, Issue 1, Feb 1981.

[33] B. Noble. *Applied Linear Algebra*. Englewood Cliffs, NJ : Prentice-Hall, 1969.

[34] J. Nocedal and S.J. Wright. *Numerical Optimization*, Springer Series in Operation Research, Second Edition, Springer Verlag, New York, 2006.

[35] A.T. Patera and G. Rozza *A Posteriori Error Estimation for Parametrized Partial Differential Equations*. MIT Pappalardo Graduate Monographs in Mechanical Engineering, 2007.

[36] M. Reed and B. Simon. *Methods of Modern Mathematical Physics I: Functional Analysis*. Academic Press, New York, 1980.

[37] C.W. Rowley. Model reduction for fluids, using balanced proper orthogonal decomposition. *International Journal of Bifurcation and Chaos*, 15:997-1013, 2005.

[38] L. Sirovich. Turbulence and the dynamics of coherent structures, parts I-III. *Quarterly of Applied Mathematics*, XLV:561-590, 1987.

[39] R. Temam. *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, volume 68 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1988.

[40] F. Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen*. Vieweg Verlag, Wiesbaden, 2005.

[41] F. Tröltzsch and S. Volkwein. POD a-posteriori error estimates for linear-quadratic optimal control problems. *Computational Optimization and Applications*, to appear.

[42] S. Volkwein. *Model Reduction using Proper Orthogonal Decomposition*. Lecture Notes, Institute of Mathematics and Scientific Computing, University of Graz.
see `http://www.uni-graz.at/imawww/volkwein/POD.pdf`

[43] S. Volkwein. Optimal control of a phase-field model using the proper orthogonal decomposition. *Zeitschrift für Angewandte Mathematik und Mechanik*, 81(2001), 83–97.

[44] S. Volkwein and A. Hepberger. Impedance identification by POD model reduction techniques. Submitted, 2007.

[45] K. Willcox and J. Peraire. Balanced model reduction via the proper orthogonal decomposition. *American Institute of Aeronautics and Astronautics (AIAA)*, 40, 2323-2330, 2002.

[46] K. Zhou, J.C. Doyle, and K. Glover. *Robust and Optimal Control*. Prentice Hall, Upper Saddle River, New Jersey, 07458, 1996.