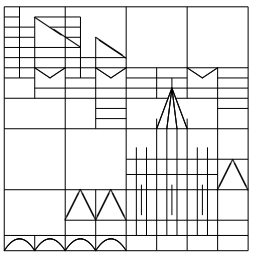
Proper Orthogonal Decomposition: Theory and Reduced-Order Modelling

S. Volkwein



Lecture Notes (August 27, 2013)

University of Konstanz
Department of Mathematics and Statistics

Contents

CHAPTER 1

The POD Method in $\mathbb{R}^m$

In this chapter we introduce the POD method in the Euclidean space $\mathbb{R}^m$. For an extension to the complex space $\mathbb{C}^m$ we refer the reader to [**22**], for instance. The goal is to find a proper orthonormal basis, the *POD basis* $\{\psi_i\}_{i=1}^{\ell}$ *of rank* ℓ, for the *snapshot set* spanned by n given vectors (the so-called *snapshots*) $y_1, \ldots, y_n \in \mathbb{R}^m$. We assume that $\ell \leq \min\{m, n\}$ holds true. The POD method is formulated as a constrained optimization problem that is solved by a Lagrangian frame work in Section 1. It turns out that the associated first-order necessary optimality conditions are strongly related to the singular value decomposition (SVD) of the rectangular matrix $Y \in \mathbb{R}^{m\times n}$ whose columns are given by the snapshots y_j, $1 \leq j \leq n$. In Section 2 we present properties of the POD basis. Section 3 is devoted to the extension of the POD method for the Euclidean space $\mathbb{R}^m$ supplied with a weighted inner product. This is used later in the formulation of the POD method for discretized partial differential equations; see Section 1.3 on Chapter 2. In Section 4 we focus on m-dimensional systems of ordinary differential equations. We consider two different variants of the POD method: one variant utilizes the whole solution trajectory $y(t)$, $t \in [0, T]$, the other one makes use of the solution y at certain time instances $0 \leq t_1 < \ldots < t_n \leq T$. The relationship of both variants is investigated.

1. POD and Singular Value Decomposition (SVD)

Let $Y = [y_1, \ldots, y_n]$ 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,

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j \tag{1.1.1}$$

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

Singular value decomposition (SVD) [**17**] guarantees the existence of real numbers $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d > 0$ and orthogonal matrices $\Psi \in \mathbb{R}^{m\times m}$ with columns $\{\psi_i\}_{i=1}^m$ and $\Phi \in \mathbb{R}^{n\times n}$ with columns $\{\phi_i\}_{i=1}^n$ such that

$$\Psi^\top Y \Phi = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} =: \Sigma \in \mathbb{R}^{m\times n}, \tag{1.1.2}$$

where $D = \operatorname{diag}(\sigma_1, \ldots, \sigma_d) \in \mathbb{R}^{d\times d}$, the zeros in (1.1.2) denote matrices of appropriate dimensions and '$\top$' stands for the transpose of a matrix (or vector). Moreover the vectors $\{\psi_i\}_{i=1}^d$ and $\{\phi_i\}_{i=1}^d$ satisfy

$$Y\phi_i = \sigma_i \psi_i \quad \text{and} \quad Y^\top \psi_i = \sigma_i \phi_i \quad \text{for } i = 1, \ldots, d. \tag{1.1.3}$$

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

From (1.1.2) we deduce that

$$Y = \Psi\Sigma\Phi^\top.$$

It follows that Y can also be expressed as

$$Y = \Psi^d D(\Phi^d)^\top, \tag{1.1.4}$$

where the matrices $\Psi^d \in \mathbb{R}^{m\times d}$ and $\Phi^d \in \mathbb{R}^{n\times d}$ are given by

$$\Psi^d_{ij} = \Psi_{ij} \quad \text{for } 1 \le i \le m,\ 1 \le j \le d,$$
$$\Phi^d_{ij} = \Phi_{ij} \quad \text{for } 1 \le i \le n,\ 1 \le j \le d.$$

Setting $B^d = D(\Phi^d)^\top \in \mathbb{R}^{d\times n}$ we can write (1.1.4) in the form

$$Y = \Psi^d B^d \quad \text{with } B^d = D(\Phi^d)^\top \in \mathbb{R}^{d\times n}.$$

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

$$\begin{aligned} y_j = \sum_{i=1}^d B^d_{ij}\Psi^d_{\cdot,i} = \sum_{i=1}^d \big(D(\Phi^d)^\top\big)_{ij}\psi_i = \sum_{i=1}^d \big(\underbrace{(\Psi^d)^\top\Psi^d}_{=I_d} D(\Phi^d)^\top\big)_{ij}\psi_i \\ \overset{(1.1.4)}{=} \sum_{i=1}^d \big((\Psi^d)^\top Y\big)_{ij}\psi_i = \sum_{i=1}^d \Big(\underbrace{\sum_{k=1}^m \Psi^d_{ki}Y_{kj}}_{=\psi_i^\top y_j}\Big)\psi_i = \sum_{i=1}^d \langle\psi_i, y_j\rangle_{\mathbb{R}^m}\,\psi_i, \end{aligned}$$

where $I_d \in \mathbb{R}^{d\times d}$ stands for the identity matrix and $\langle\cdot\,,\cdot\rangle_{\mathbb{R}^m}$ denotes the canonical inner product in $\mathbb{R}^m$. Thus,

$$y_j = \sum_{i=1}^d \langle y_j, \psi_i\rangle_{\mathbb{R}^m}\,\psi_i \quad \text{for } j = 1, \dots, n. \tag{1.1.5}$$

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

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

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

Note that $(\mathbf{P}^1)$ is a constrained optimization problem that can be solved by considering first-order necessary optimality conditions; see Appendix D. For that purpose we want to write $(\mathbf{P}^1)$ in the standard form (P) on page 78. We introduce the function $e : \mathbb{R}^m \to \mathbb{R}$ by $e(\psi) = 1 - \|\psi\|^2_{\mathbb{R}^m}$ for $\psi \in \mathbb{R}^m$. Then, the equality constraint in $(\mathbf{P}^1)$ can be expressed as $e(\psi) = 0$. To ensure the existence of Lagrange multipliers a constraint qualification is needed. Notice that $\nabla e(\psi) = 2\psi^\top$ is linear independent if $\psi \neq 0$ holds. In particular, a solution to $(\mathbf{P}^1)$ satisfies $\psi \neq 0$. Thus,

any solution to $(\mathbf{P}^1)$ is a *regular point*; see Definition D.2. Let $\mathcal{L} : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ be the Lagrange functional associated with $(\mathbf{P}^1)$, i.e.,

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

Suppose that $\psi \in \mathbb{R}^m$ is a solution to $(\mathbf{P}^1)$. Since ψ is a regular point, we infer from Theorem D.4 that there exists a unique Lagrange multiplier $\lambda \in \mathbb{R}^m$ satisfying the first-order necessary optimality condition

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

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

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

Thus,

$$\nabla_{\psi}\mathcal{L}(\psi, \lambda) = 2\big(YY^{\top}\psi - \lambda\psi\big) \overset{!}{=} 0 \quad \text{in } \mathbb{R}^m. \tag{1.1.6}$$

Equation (1.1.6) yields the eigenvalue problem

$$YY^{\top}\psi = \lambda\psi \quad \text{in } \mathbb{R}^m. \tag{1.1.7a}$$

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

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

Thus, $YY^{\top}$ is positive semi-definite. It follows that $YY^{\top}$ possesses m nonnegative 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 $\frac{\partial \mathcal{L}}{\partial \lambda}(\psi, \lambda) \overset{!}{=} 0$ in $\mathbb{R}$ we infer the constraint

$$\|\psi\|_{\mathbb{R}^m} = 1. \tag{1.1.7b}$$

Due to SVD the vector ψ_1 solves (1.1.7) and

$$\begin{aligned}
\sum_{j=1}^{n}\left|\langle y_j, \psi_1\rangle_{\mathbb{R}^m}\right|^2 &= \sum_{j=1}^{n}\langle y_j, \psi_1\rangle_{\mathbb{R}^m}\langle y_j, \psi_1\rangle_{\mathbb{R}^m} = \sum_{j=1}^{n}\big\langle\langle y_j, \psi_1\rangle_{\mathbb{R}^m} y_j, \psi_1\big\rangle_{\mathbb{R}^m} \\
&= \left\langle\sum_{j=1}^{n}\langle y_j, \psi_1\rangle_{\mathbb{R}^m} y_j, \psi_1\right\rangle_{\mathbb{R}^m} = \left\langle\sum_{j=1}^{n}\left(\sum_{k=1}^{m} Y_{kj}(\psi_1)_k\right)y_j, \psi_1\right\rangle_{\mathbb{R}^m} \\
&= \left\langle\sum_{k=1}^{m}\left(\sum_{j=1}^{n} Y_{\cdot,j}Y_{jk}^{\top}(\psi_1)_k\right), \psi_1\right\rangle_{\mathbb{R}^m} = \big\langle YY^{\top}\psi_1, \psi_1\big\rangle_{\mathbb{R}^m} \\
&= \lambda_1\big\langle\psi_1, \psi_1\big\rangle_{\mathbb{R}^m} = \lambda_1\|\psi_1\|_{\mathbb{R}^m}^2 = \lambda_1,
\end{aligned}$$

where $(\psi_1)_k$ denotes the k-th component of the vector ψ_1.

Note that $\nabla_{\psi\psi}\mathcal{L}(\psi,\lambda) = 2(YY^\top - \lambda I_m) \in \mathbb{R}^{m\times m}$ holds. Let $\psi \in \mathbb{R}^m$ be chosen arbitrary. Since $YY^\top$ is symmetric, there exist m orthonomal eigenvectors $\psi_1,\dots,\psi_m \in \mathbb{R}^m$ of $YY^\top$ satisfying $YY^\top\psi_i = \lambda_i\psi_i$ for $1 \le i \le m$. Then, we can write ψ in the form

$$\psi = \sum_{i=1}^m \langle \psi, \psi_i\rangle_{\mathbb{R}^m}\,\psi_i.$$

At (ψ_1,λ_1) we conlude from $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_m \ge 0$ that

$$\begin{aligned}
&\langle \psi, \nabla_{\psi\psi}\mathcal{L}(\psi_1,\lambda_1)\psi\rangle_{\mathbb{R}^m} = 2\,\langle \psi, \big(YY^\top - \lambda_1 I_m\big)\psi\rangle_{\mathbb{R}^m}\\
&= 2\sum_{i=1}^m\sum_{j=1}^m \langle \psi,\psi_i\rangle_{\mathbb{R}^m}\langle \psi,\psi_j\rangle_{\mathbb{R}^m}\langle \psi_i, \big(YY^\top - \lambda_1 I_m\big)\psi_j\rangle_{\mathbb{R}^m}\\
&= 2\sum_{i=1}^m\sum_{j=1}^m \big(\lambda_j - \lambda_1\big)\,\langle \psi,\psi_i\rangle_{\mathbb{R}^m}\langle \psi,\psi_j\rangle_{\mathbb{R}^m}\langle \psi_i,\psi_j\rangle_{\mathbb{R}^m}\\
&= 2\sum_{i=1}^m \big(\lambda_j - \lambda_1\big)\big|\langle \psi,\psi_i\rangle_{\mathbb{R}^m}\big|^2 \le 0.
\end{aligned}$$

Thus, (ψ_1,λ_1) satisfies the second-order necessary optimality conditions for a maximum, but not the sufficient ones; compare Theorems D.5 and D.6. We next prove that ψ_1 actually solves $(\mathbf{P}^1)$. Suppose that $\tilde\psi \in \mathbb{R}^m$ is an arbitrary vector with $\|\tilde\psi\|_{\mathbb{R}^m} = 1$. Since $\{\psi_i\}_{i=1}^m$ is an orthonormal basis in $\mathbb{R}^m$, we have

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

Thus,

$$\begin{aligned}
\sum_{j=1}^n \big|\langle y_j,\tilde\psi\rangle_{\mathbb{R}^m}\big|^2 &= \sum_{j=1}^n \left|\Big\langle y_j, \sum_{i=1}^m \langle\tilde\psi,\psi_i\rangle_{\mathbb{R}^m}\,\psi_i\Big\rangle_{\mathbb{R}^m}\right|^2\\
&= \sum_{j=1}^n\sum_{i=1}^m\sum_{k=1}^m \Big(\big\langle y_j, \langle\tilde\psi,\psi_i\rangle_{\mathbb{R}^m}\,\psi_i\big\rangle_{\mathbb{R}^m}\big\langle y_j, \langle\tilde\psi,\psi_k\rangle_{\mathbb{R}^m}\,\psi_k\big\rangle_{\mathbb{R}^m}\Big)\\
&= \sum_{j=1}^n\sum_{i=1}^m\sum_{k=1}^m \Big(\langle y_j,\psi_i\rangle_{\mathbb{R}^m}\langle y_j,\psi_k\rangle_{\mathbb{R}^m}\langle\tilde\psi,\psi_i\rangle_{\mathbb{R}^m}\langle\tilde\psi,\psi_k\rangle_{\mathbb{R}^m}\Big)\\
&= \sum_{i=1}^m\sum_{k=1}^m \Big(\Big\langle \underbrace{\sum_{j=1}^n \langle y_j,\psi_i\rangle_{\mathbb{R}^m}\,y_j}_{=\lambda_i\psi_i}, \psi_k\Big\rangle_{\mathbb{R}^m}\langle\tilde\psi,\psi_i\rangle_{\mathbb{R}^m}\langle\tilde\psi,\psi_k\rangle_{\mathbb{R}^m}\Big)\\
&= \sum_{i=1}^m\sum_{k=1}^m \Big(\underbrace{\langle\lambda_i\psi_i,\psi_k\rangle_{\mathbb{R}^m}}_{=\lambda_i\delta_{ik}}\langle\tilde\psi,\psi_i\rangle_{\mathbb{R}^m}\langle\tilde\psi,\psi_k\rangle_{\mathbb{R}^m}\Big)\\
&= \sum_{i=1}^m \lambda_i\big|\langle\tilde\psi,\psi_i\rangle_{\mathbb{R}^m}\big|^2 \le \lambda_1\sum_{i=1}^m\big|\langle\tilde\psi,\psi_i\rangle_{\mathbb{R}^m}\big|^2 = \lambda_1\,\|\tilde\psi\|_{\mathbb{R}^m}^2 = \lambda_1\\
&= \sum_{j=1}^n \big|\langle y_j,\psi_1\rangle_{\mathbb{R}^m}\big|^2.
\end{aligned}$$

Consequently, ψ_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 ψ_1 that again describes the data set $\{y_i\}_{i=1}^n$ as well as possible then we need to solve

$$(\mathbf{P}^2) \qquad \max_{\tilde{\psi}\in\mathbb{R}^m} \sum_{j=1}^{n} \left|\langle y_j, \tilde{\psi}\rangle_{\mathbb{R}^m}\right|^2 \quad \text{s.t.} \quad \|\tilde{\psi}\|_{\mathbb{R}^m} = 1 \text{ and } \langle \tilde{\psi}, \psi_1\rangle_{\mathbb{R}^m} = 0.$$

SVD implies that ψ_2 is a solution to $(\mathbf{P}^2)$ and $\operatorname{argmax}(\mathbf{P}^2) = \sigma_2^2 = \lambda_2$. In fact, ψ_2 solves the first-order necessary optimality conditions (1.1.7) and for

$$\tilde{\psi} = \sum_{i=2}^{m} \langle \tilde{\psi}, \psi_i\rangle_{\mathbb{R}^m}\, \psi_i \in \operatorname{span}\{\psi_1\}^{\perp}$$

we have

$$\sum_{j=1}^{n} \left|\langle y_j, \tilde{\psi}\rangle_{\mathbb{R}^m}\right|^2 \le \lambda_2 = \sum_{j=1}^{n} \left|\langle y_j, \psi_2\rangle_{\mathbb{R}^m}\right|^2.$$

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

THEOREM 1.1.1. *Let $Y = [y_1, \dots, y_n] \in \mathbb{R}^{m\times n}$ be a given matrix with rank $d \le \min\{m, n\}$. Further, let $Y = \Psi\Sigma\Phi^T$ be the singular value decomposition of Y, where $\Psi = [\psi_1, \dots, \psi_m] \in \mathbb{R}^{m\times m}$, $\Phi = [\phi_1, \dots, \phi_n] \in \mathbb{R}^{n\times n}$ are orthogonal matrices and the matrix $\Sigma \in \mathbb{R}^{m\times n}$ has the form as* (1.1.2). *Then, for any $\ell \in \{1, \dots, d\}$ the solution to*

$$(\mathbf{P}^\ell) \qquad \max_{\tilde{\psi}_1,\dots,\tilde{\psi}_\ell\in\mathbb{R}^m} \sum_{i=1}^{\ell}\sum_{j=1}^{n} \left|\langle y_j, \tilde{\psi}_i\rangle_{\mathbb{R}^m}\right|^2 \quad \textit{s.t.} \quad \langle \tilde{\psi}_i, \tilde{\psi}_j\rangle_{\mathbb{R}^m} = \delta_{ij} \textit{ for } 1 \le i, j \le \ell$$

is given by the singular vectors $\{\psi_i\}_{i=1}^{\ell}$, i.e., by the first ℓ columns of Ψ. In $(\mathbf{P}^\ell)$ we denote by δ_{ij} the Kronecker symbol satisfying $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ otherwise. Moreover,

$$\operatorname{argmax}(\mathbf{P}^\ell) = \sum_{i=1}^{\ell} \sigma_i^2 = \sum_{i=1}^{\ell} \lambda_i. \tag{1.1.8}$$

PROOF. Since $(\mathbf{P}^\ell)$ is an equality constrained optimization problem, we introduce the Lagrangian (see Appendix D)

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

by

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

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

$$\frac{\partial \mathcal{L}}{\partial \psi_k}(\psi_1, \dots, \psi_\ell, \Lambda)\delta\psi_k = 0 \quad \text{for all } \delta\psi_k \in \mathbb{R}^m \text{ and } k \in \{1, \dots, \ell\}. \tag{1.1.9}$$

From

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

and (1.1.9) we infer that

$$\sum_{j=1}^{n}\langle y_j,\psi_k\rangle_{\mathbb{R}^m}\,y_j=\frac{1}{2}\sum_{i=1}^{\ell}(\lambda_{ik}+\lambda_{ki})\,\psi_i \text{ in } \mathbb{R}^m \text{ for all } k\in\{1,\dots,\ell\}. \tag{1.1.10}$$

Note that

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

Thus, condition (1.1.10) can be expressed as

$$YY^\top\psi_k=\frac{1}{2}\sum_{i=1}^{\ell}(\lambda_{ik}+\lambda_{ki})\,\psi_i\quad\text{in } \mathbb{R}^m \text{ for all } k\in\{1,\dots,\ell\}. \tag{1.1.11}$$

Now we proceed by induction. For $\ell=1$ we have $k=1$. It follows from (1.1.11) that

$$YY^\top\psi_1=\lambda_1\psi_1\quad\text{in } \mathbb{R}^m \tag{1.1.12}$$

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

$$YY^\top\psi_k=\lambda_k\psi_k\quad\text{in } \mathbb{R}^m \text{ for all } k\in\{1,\dots,\ell\}. \tag{1.1.13}$$

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

$$YY^\top\psi_k=\lambda_k\psi_k\quad\text{in } \mathbb{R}^m \text{ for all } k\in\{1,\dots,\ell+1\}. \tag{1.1.14}$$

By assumption we have (1.1.13). Thus, we only have to prove that

$$YY^\top\psi_{\ell+1}=\lambda_{\ell+1}\psi_{\ell+1}\quad\text{in } \mathbb{R}^m. \tag{1.1.15}$$

Due to (1.1.11) we have

$$YY^\top\psi_{\ell+1}=\frac{1}{2}\sum_{i=1}^{\ell+1}(\lambda_{i,\ell+1}+\lambda_{\ell+1,i})\,\psi_i\quad\text{in } \mathbb{R}^m. \tag{1.1.16}$$

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 $1 \le j \le \ell$. Using (1.1.13) and the symmetry of $YY^\top$ we have for any $j \in \{1, \dots, \ell\}$

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

This gives

$$\lambda_{\ell+1,i} = -\lambda_{i,\ell+1} \quad \text{for any } i \in \{1, \dots, \ell\}. \tag{1.1.17}$$

Inserting (1.1.17) into (1.1.16) we obtain

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

Setting $\lambda_{\ell+1} = \lambda_{\ell+1,\ell+1}$ we obtain (1.1.15).
Summarizing, the necessary optimality conditions for $(\mathbf{P}^\ell)$ are given by the symmetric $m \times m$ eigenvalue problem

$$YY^\top \psi_i = \lambda_i \psi_i \quad \text{for } i = 1, \dots, \ell. \tag{1.1.18}$$

It follows from SVD that $\{\psi_i\}_{i=1}^{\ell}$ solves (1.1.18). The proof that $\{\psi_i\}_{i=1}^{\ell}$ is a solution to $(\mathbf{P}^\ell)$ and that $\operatorname{argmax}(\mathbf{P}^\ell) = \sum_{i=1}^{\ell} \sigma_i^2$ holds is analogous to the proof for $(\mathbf{P}^1)$; see Exercise 1.5.5. □

Motivated by the previous theorem we give the next definition. Moreover, in Algorithm 1 the computation of a POD basis of rank ℓ is summarized.

DEFINITION 1.1.2. *For $\ell \in \{1, \dots, d\}$ the vectors $\{\psi_i\}_{i=1}^{\ell}$ are called* POD basis of rank ℓ.

Algorithm 1 (POD basis of rank ℓ)

Require: Snapshots $\{y_j\}_{j=1}^{n} \subset \mathbb{R}^m$, POD rank $\ell \le d$ and `flag` for the solver;
1: Set $Y = [y_1, \dots, y_n] \in \mathbb{R}^{m \times n}$;
2: **if** `flag` $= 0$ **then**
3: Compute singular value decomposition $[\Psi, \Sigma, \Phi] = \operatorname{svd}(Y)$;
4: Set $\psi_i = \Psi_{\cdot,i} \in \mathbb{R}^m$ and $\lambda_i = \Sigma_{ii}^2$ for $i = 1, \dots, \ell$;
5: **else if** `flag` $= 1$ **then**
6: Determine $R = YY^\top \in \mathbb{R}^{m \times m}$;
7: Compute eigenvalue decomposition $[\Psi, \Lambda] = \operatorname{eig}(R)$;
8: Set $\psi_i = \Psi_{\cdot,i} \in \mathbb{R}^m$ and $\lambda_i = \Lambda_{ii}$ for $i = 1, \dots, \ell$;
9: **end if**
10: **return** POD basis $\{\psi_i\}_{i=1}^{\ell}$ and eigenvalues $\{\lambda_i\}_{i=1}^{\ell}$;

2. Properties of the POD Basis

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

COROLLARY 1.2.1 (Optimality of the POD basis). *Let all hypotheses of Theorem* 1.1.1 *be satisfied. Suppose that $\hat{\Psi}^d \in \mathbb{R}^{m\times d}$ denotes a matrix with pairwise orthonormal vectors $\hat{\psi}_i$ and that the expansion of the columns of Y in the basis $\{\hat{\psi}_i\}_{i=1}^{d}$ be given by*

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

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

$$\|Y - \Psi^\ell B^\ell\|_F \leq \|Y - \hat{\Psi}^\ell C^\ell\|_F. \tag{1.2.1}$$

In (1.2.1)*, $\|\cdot\|_F$ denotes the Frobenius norm given by*

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

the matrix Ψ^ℓ denotes the first $\ell \leq d$ columns of Ψ, B^ℓ the first ℓ rows of B and similarly for $\hat{\Psi}^\ell$ and C^ℓ. Moreover, $\operatorname{trace}(A)$ *denotes the sum over the diagonal elements of a given matrix A.*

PROOF OF COROLLARY 1.2.1. From Exercise 1.4.6 it follows that

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

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,

$$\begin{aligned}
\|Y - \Psi^\ell B^\ell\|_F^2 = \|\Psi^k (B^d - B_0^\ell)\|_F^2 = \|B^d - B_0^\ell\|_F^2 &= \sum_{i=\ell+1}^{d}\sum_{j=1}^{n} |B^d_{ij}|^2 \\
&= \sum_{i=\ell+1}^{d}\sum_{j=1}^{n} |\langle y_j, \psi_i \rangle_{\mathbb{R}^m}|^2 \\
&= \sum_{i=\ell+1}^{d}\sum_{j=1}^{n} \langle \langle y_j, \psi_i \rangle_{\mathbb{R}^m} y_j, \psi_i \rangle_{\mathbb{R}^m} = \sum_{i=\ell+1}^{d} \langle YY^\top \psi_i, \psi_i \rangle_{\mathbb{R}^m} \\
&= \sum_{i=\ell+1}^{d} \sigma_i^2,
\end{aligned} \tag{1.2.2}$$

By Theorem 1.1.1 the vectors $\psi_1, \ldots, \psi_\ell$ solve $(\mathbf{P}^\ell)$. From (1.2.2),

$$\|Y\|_F^2 = \|\hat{\Psi}^d C^d\|_F^2 = \|C^d\|_F^2 = \sum_{i=1}^{d}\sum_{j=1}^{n} |C^d_{ij}|^2$$

and

$$\|Y\|_F^2 = \|\Psi^d B^d\|_F^2 = \|B^d\|_F^2 = \sum_{i=1}^{d}\sum_{j=1}^{n} |B^d_{ij}|^2 = \sum_{i=1}^{d} \sigma_i^2$$

we infer that

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

which gives (1.2.1). □

Notice that

$$\begin{aligned}\|Y-\hat\Psi^\ell C^\ell\|_F^2 &= \sum_{i=1}^{m}\sum_{j=1}^{n}\Big|Y_{ij}-\sum_{k=1}^{\ell}\hat\Psi_{ik}^\ell C_{kj}\Big|^2 = \sum_{j=1}^{n}\sum_{i=1}^{m}\Big|Y_{ij}-\sum_{k=1}^{\ell}\langle\hat\psi_k,y_j\rangle_{\mathbb{R}^m}\hat\Psi_{ik}^\ell\Big|^2\\ &= \sum_{j=1}^{n}\Big\|y_j-\sum_{k=1}^{\ell}\langle y_j,\hat\psi_k\rangle_{\mathbb{R}^m}\hat\psi_k\Big\|_{\mathbb{R}^m}^2.\end{aligned}$$

Analogously,

$$\|Y-\Psi^\ell B^\ell\|_F^2 = \sum_{j=1}^{n}\Big\|y_j-\sum_{k=1}^{\ell}\langle y_j,\psi_k\rangle_{\mathbb{R}^m}\psi_k\Big\|_{\mathbb{R}^m}^2.$$

Thus, (1.2.1) implies that

$$\sum_{j=1}^{n}\Big\|y_j-\sum_{k=1}^{\ell}\langle y_j,\psi_k\rangle_{\mathbb{R}^m}\psi_k\Big\|_{\mathbb{R}^m}^2 \le \sum_{j=1}^{n}\Big\|y_j-\sum_{k=1}^{\ell}\langle y_j,\hat\psi_k\rangle_{\mathbb{R}^m}\hat\psi_k\Big\|_{\mathbb{R}^m}^2$$

for any other set $\{\hat\psi_i\}_{i=1}^\ell$ of ℓ pairwise orthonormal vectors. Hence, it follows from Corollary 1.2.1 that the POD basis of rank ℓ can also be determined by solving

$$\begin{aligned}&\min_{\tilde\psi_1,\dots,\tilde\psi_\ell\in\mathbb{R}^m}\sum_{j=1}^{n}\Big\|y_j-\sum_{i=1}^{\ell}\langle y_j,\tilde\psi_i\rangle_{\mathbb{R}^m}\tilde\psi_i\Big\|_{\mathbb{R}^m}^2\\ &\text{s.t. } \langle\tilde\psi_i,\tilde\psi_j\rangle_{\mathbb{R}^m}=\delta_{ij} \text{ for } 1\le i,j\le\ell.\end{aligned}\tag{1.2.3}$$

Remark 1.2.2. We compare first-order optimality conditions for $(\mathbf{P}^\ell)$ and (1.2.3). Let $\{\psi_i\}_{i=1}^\ell$ be a given set of orthonormal vectors in $\mathbb{R}^m$, i.e.

$$\langle\psi_i,\psi_k\rangle_{\mathbb{R}^m}=\delta_{ik} \quad \text{for } i,k\in\{1,\dots,m\}.\tag{1.2.4}$$

For any index $k\in\{1,\dots,\ell\}$ and any direction $\psi_\delta\in\mathbb{R}^m$ we have

$$\begin{aligned}0 &= \frac{\partial}{\partial\psi_k}\big(\delta_{ik}\big)\psi_\delta = \frac{\partial}{\partial\psi_k}\big(\langle\psi_i,\psi_k\rangle_{\mathbb{R}^m}\big)\psi_\delta\\ &= \begin{cases}\langle\psi_i,\psi_\delta\rangle_{\mathbb{R}^m} & \text{for } i\in\{1,\dots,\ell\}\setminus\{k\},\\ 2\langle\psi_i,\psi_\delta\rangle_{\mathbb{R}^m} & \text{for } i=k.\end{cases}\end{aligned}$$

Hence

$$\langle\psi_i,\psi_\delta\rangle_{\mathbb{R}^m}=0 \quad \text{for } i\in\{1,\dots,\ell\} \text{ and } \psi_\delta\in\mathbb{R}^m.\tag{1.2.5}$$

Suppose that $y_1, \dots, y_n \in \mathbb{R}^m$ are the given snapshots. For $\ell \in \{1, \dots, m\}$ we set

$$z_j = z_j(\psi_1, \dots, \psi_\ell) = y_j - \sum_{i=1}^{\ell} \langle y_j, \psi_i \rangle_{\mathbb{R}^m} \psi_i \in \mathbb{R}^m \quad \text{for } j = 1, \dots, n.$$

Let

$$J(\psi_1, \dots, \psi_\ell) = \sum_{j=1}^{n} \|z_j\|_{\mathbb{R}^m}^2. \tag{1.2.6}$$

Using (1.2.4) we have

$$\begin{aligned} \|z_j\|_2^2 = \langle z_j, z_j \rangle_{\mathbb{R}^m} &= \Big\langle y_j - \sum_{k=1}^{\ell} \langle y_j, \psi_k \rangle_{\mathbb{R}^m} \psi_k, y_j - \sum_{i=1}^{\ell} \langle y_j, \psi_i \rangle_{\mathbb{R}^m} \psi_i \Big\rangle_{\mathbb{R}^m} \\ &= \langle y_j, y_j \rangle_{\mathbb{R}^m} - 2 \sum_{i=1}^{\ell} \langle y_j, \psi_i \rangle_{\mathbb{R}^m} \langle y_j, \psi_i \rangle_{\mathbb{R}^m} \\ &\quad + \sum_{i=1}^{\ell} \sum_{k=1}^{\ell} \langle y_j, \psi_i \rangle_{\mathbb{R}^m} \langle y_j, \psi_k \rangle_{\mathbb{R}^m} \langle \psi_i, \psi_k \rangle_{\mathbb{R}^m} \\ &= \|y_j\|_{\mathbb{R}^m}^2 - 2 \sum_{i=1}^{\ell} \big| \langle y_j, \psi_i \rangle_{\mathbb{R}^m} \big|^2 + \sum_{i=1}^{\ell} \big| \langle y_j, \psi_i \rangle_{\mathbb{R}^m} \big|^2 \\ &= \|y_j\|_{\mathbb{R}^m}^2 - \sum_{i=1}^{\ell} \big| \langle y_j, \psi_i \rangle_{\mathbb{R}^m} \big|^2. \end{aligned} \tag{1.2.7}$$

Combining (1.2.6) and (1.2.7) we derive

$$J(\psi_1, \dots, \psi_\ell) = \sum_{j=1}^{n} \|z_j\|_{\mathbb{R}^m}^2 = \sum_{j=1}^{n} \Big(\|y_j\|_{\mathbb{R}^m}^2 - \sum_{i=1}^{\ell} \big| \langle y_j, \psi_i \rangle_{\mathbb{R}^m} \big|^2 \Big). \tag{1.2.8}$$

For any $k \in \{1, \dots, \ell\}$ we will consider the derivatives

$$\frac{\partial}{\partial \psi_k} \Big(\sum_{j=1}^{n} \Big(\|y_j\|_{\mathbb{R}^m}^2 - \sum_{i=1}^{\ell} \big| \langle y_j, \psi_i \rangle_{\mathbb{R}^m} \big|^2 \Big) \Big) \quad \text{and} \quad \frac{\partial}{\partial \psi_k} \Big(\sum_{j=1}^{n} \|z_j(\psi_1, \dots, \psi_\ell)\|_{\mathbb{R}^m}^2 \Big)$$

Due to (1.2.8) both derivatives must be the same. Notice that

$$\begin{aligned} \frac{\partial J}{\partial \psi_k}(\psi_1, \dots, \psi_\ell)\psi_\delta &= \frac{\partial}{\partial \psi_k} \Big(\sum_{j=1}^{n} \Big(\|y_j\|_{\mathbb{R}^m}^2 - \sum_{i=1}^{\ell} \big| \langle y_j, \psi_i \rangle_{\mathbb{R}^m} \big|^2 \Big) \Big) \psi_\delta \\ &= \sum_{j=1}^{n} \frac{\partial}{\partial \psi_k} \Big(\|y_j\|_{\mathbb{R}^m}^2 - \sum_{i=1}^{\ell} \langle y_j, \psi_i \rangle_{\mathbb{R}^m} \big|^2 \Big) \psi_\delta \\ &= - \sum_{j=1}^{n} 2 \langle y_j, \psi_k \rangle_{\mathbb{R}^m} \langle y_j, \psi_\delta \rangle_{\mathbb{R}^m} \\ &= \Big\langle - 2 \sum_{j=1}^{n} \langle y_j, \psi_k \rangle_{\mathbb{R}^m} y_j, \psi_\delta \Big\rangle_{\mathbb{R}^m} \end{aligned}$$

for any direction $\psi_\delta \in \mathbb{R}^m$ and for $1 \le k \le \ell$. Note that

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

Then, we find that

$$\frac{\partial J}{\partial \psi_k}(\psi_1, \dots, \psi_\ell) = -2YY^\top \psi_k \quad \text{for } 1 \le k \le \ell. \tag{1.2.9}$$

On the other hand we have

$$\frac{\partial z_j}{\partial \psi_k}\psi_\delta = -\langle y_j, \psi_k \rangle_{\mathbb{R}^m}\psi_\delta - \langle y_j, \psi_\delta \rangle_{\mathbb{R}^m}\psi_k = -\langle y_j, \psi_k \rangle_{\mathbb{R}^m}\,\psi_\delta - \langle y_j, \psi_\delta \rangle_{\mathbb{R}^m}\,\psi_k$$

for $1 \le k \le \ell$ and $\psi_\delta \in \mathbb{R}^m$. Using (1.2.4) and (1.2.5) we find that

$$\begin{aligned}
\frac{\partial}{\partial \psi_k}\Big(\|z_j\|^2_{\mathbb{R}^m}\Big)\psi_\delta = \frac{\partial}{\partial \psi_k}\big(\langle z_j, z_j \rangle_{\mathbb{R}^m}\big)\psi_\delta &= 2\left\langle z_j, \frac{\partial z_j}{\partial \psi_k} u_\delta \right\rangle_{\mathbb{R}^m} \\
&= 2\left\langle y_j - \sum_{i=1}^{\ell} \langle y_j, \psi_i \rangle_{\mathbb{R}^m}\,\psi_i, -\langle y_j, \psi_\delta \rangle_{\mathbb{R}^m}\,\psi_k - \langle y_j, \psi_k \rangle_{\mathbb{R}^m}\,\psi_\delta \right\rangle_{\mathbb{R}^m} \\
&= -2\,\langle y_j, \psi_\delta \rangle_{\mathbb{R}^m}\langle y_j, \psi_k \rangle_{\mathbb{R}^m} + 2\sum_{i=1}^{\ell} \langle y_j, \psi_i \rangle_{\mathbb{R}^m}\langle y_j, \psi_\delta \rangle_{\mathbb{R}^m}\langle y_j, \psi_k \rangle_{\mathbb{R}^m} \\
&\quad - 2\,\langle y_j, \psi_k \rangle_{\mathbb{R}^m}\langle y_j, \psi_\delta \rangle_{\mathbb{R}^m} + 2\sum_{i=1}^{\ell} \langle y_j, u_i \rangle_{\mathbb{R}^m}\langle y_j, u_k \rangle_{\mathbb{R}^m}\langle y_j, \psi_\delta \rangle_{\mathbb{R}^m} \\
&= -2\,\langle y_j, \psi_k \rangle_{\mathbb{R}^m}\langle y_j, \psi_\delta \rangle_{\mathbb{R}^m} = \big\langle -2\langle y_j, \psi_k \rangle_{\mathbb{R}^m}\, y_j, \psi_\delta \big\rangle_{\mathbb{R}^m}
\end{aligned}$$

for any direction $\psi_\delta \in \mathbb{R}^m$, for $j = 1, \dots, n$ and for $1 \le k \le \ell$. Summarizing, we have

$$\frac{\partial J}{\partial \psi_k}(\psi_1, \dots, \psi_\ell) = -2\sum_{j=1}^{n} \langle y_j, \psi_k \rangle_{\mathbb{R}^m}\, y_j = -2YY^\top \psi_k$$

which coincides with (1.2.9). ◊

Remark 1.2.3. It follows from Corollary 1.2.1 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:

$$\sum_{i=1}^{\ell}\sum_{j=1}^{n} \big|\langle y_j, \psi_i \rangle_{\mathbb{R}^m}\big|^2 = \sum_{i=1}^{\ell} \sigma_i^2 = \sum_{i=1}^{\ell} \lambda_i \ge \sum_{i=1}^{\ell}\sum_{j=1}^{n} \big|\langle y_j, \hat{\psi}_i \rangle_{\mathbb{R}^m}\big|^2$$

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

The next corollary states that the POD coefficients are uncorrelated.

Corollary 1.2.4 (Uncorrelated POD coefficients). *Let all hypotheses of Theorem* 1.1.1 *hold. Then.*

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

PROOF. The claim follows from (1.1.18) and $\langle \psi_i, \psi_k \rangle_{\mathbb{R}^m} = \delta_{ik}$ for $1 \le i, k \le \ell$:

$$\sum_{j=1}^{n} \langle y_j, \psi_i \rangle_{\mathbb{R}^m} \langle y_j, \psi_k \rangle_{\mathbb{R}^m} = \Big\langle \underbrace{\sum_{j=1}^{n} \langle y_j, \psi_i \rangle_{\mathbb{R}^m} y_j}_{= Y Y^\top \psi_i}, \psi_k \Big\rangle_{\mathbb{R}^m} = \langle \sigma_i^2 \psi_i, \psi_k \rangle_{\mathbb{R}^m} = \sigma_i^2 \delta_{ik}.$$

□

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 $\phi_1, \dots, \phi_\ell \in \mathbb{R}^n$ by solving the symmetric $n \times n$ eigenvalue problem

$$Y^\top Y \phi_i = \lambda_i \phi_i \quad \text{for } i = 1, \dots, \ell \tag{1.2.10}$$

and set, by (1.1.3),

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

For historical reasons [**20**] 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 (1.1.18).

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

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

Notice that we have $\sum_{i=1}^{d} \lambda_i = \operatorname{trace}(YY^\top) = \operatorname{trace}(Y^\top Y)$. Let us mention that POD is also called *Principal Component Analysis* (PCA) and *Karhunen-Loève Decomposition*. In Algorithm 2 we extend Algorithm 1.

Algorithm 2 (POD basis of rank ℓ)

Require: Snapshots $\{y_j\}_{j=1}^n \subset \mathbb{R}^m$, POD rank $\ell \le d$ and `flag` for the solver;
Set $Y = [y_1, \dots, y_n] \in \mathbb{R}^{m \times n}$;
if `flag` $= 0$ **then**
  Compute singular value decomposition $[\Psi, \Sigma, \Phi] = \operatorname{svd}(Y)$;
  Set $\psi_i = \Psi_{\cdot,i} \in \mathbb{R}^m$ and $\lambda_i = \Sigma_{ii}^2$ for $i = 1, \dots, \ell$;
else if `flag` $= 1$ **then**
  Determine $R = YY^\top \in \mathbb{R}^{m \times m}$;
  Compute eigenvalue decomposition $[\Psi, \Lambda] = \operatorname{eig}(R)$;
  Set $\psi_i = \Psi_{\cdot,i} \in \mathbb{R}^m$ and $\lambda_i = \Lambda_{ii}$ for $i = 1, \dots, \ell$;
else if `flag` $= 2$ **then**
  Determine $K = Y^\top Y \in \mathbb{R}^{n \times n}$;
  Compute eigenvalue decomposition $[\Phi, \Lambda] = \operatorname{eig}(K)$;
  Set $\psi_i = Y \Phi_{\cdot,i} / \sqrt{\lambda_i} \in \mathbb{R}^m$ and $\lambda_i = \Lambda_{ii}$ for $i = 1, \dots, \ell$;
end if
Compute $\mathcal{E}(\ell) = \sum_{i=1}^{\ell} \lambda_i / \sum_{i=1}^{d} \lambda_i$;
return POD basis $\{\psi_i\}_{i=1}^{\ell}$, eigenvalues $\{\lambda_i\}_{i=1}^{\ell}$ and ratio $\mathcal{E}(\ell)$;

3. The POD Method with a Weighted Inner Product

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

$$\langle \psi, \tilde{\psi} \rangle_W = \psi^\top W \tilde{\psi} = \langle \psi, W\tilde{\psi} \rangle_{\mathbb{R}^m} = \langle W\psi, \tilde{\psi} \rangle_{\mathbb{R}^m} \quad \text{for } \psi, \tilde{\psi} \in \mathbb{R}^m, \tag{1.3.1}$$

where $W \in \mathbb{R}^{m\times m}$ is a symmetric, positive definite matrix. Furthermore, let $\|\psi\|_W = \sqrt{\langle \psi, \psi \rangle_W}$ for $\psi \in \mathbb{R}^m$ be the associated induced norm. For the choice $W = I_m$, the inner product (1.3.1) coincides the Euclidean inner product.

Example 1.3.1. 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:

$$\langle \varphi, \tilde{\varphi} \rangle_{L^2_h(\Omega)} = h\left(\frac{\varphi_1^h \tilde{\varphi}_1^h}{2} + \sum_{i=2}^{m-1} \left(\varphi_i^h \tilde{\varphi}_i^h \right) + \frac{\varphi_m^h \tilde{\varphi}_m^h}{2} \right), \tag{1.3.2}$$

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

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

Thus, the discrete L^2-inner product can be written as a weighted inner product of the form (1.3.1). Let us also refer to Exercise 1.5.7, where an extension to a two-dimensional domain Ω is investigated. ◊

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

$$\max_{\tilde{\psi} \in \mathbb{R}^m} \sum_{j=1}^n \left| \langle y_j, \tilde{\psi} \rangle_W \right|^2 \quad \text{s.t.} \quad \|\tilde{\psi}\|_W = 1. \tag{$\mathbf{P}^1_W$}$$

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

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

We introduce the function $e:\mathbb{R}^m\to\mathbb{R}$ by $e(\psi)=1-\|\psi\|_{\mathbb{R}^m}^2=1-\psi^\top W\psi$ for $\psi\in\mathbb{R}^m$. Then, the equality constraint in $(\mathbf{P}_W^1)$ can be expressed as $e(\tilde\psi)=0$. Notice that $\nabla e(\psi)=2\psi^\top W$ is linear independent if $\psi\neq 0$ holds. Suppose that $\psi\in\mathbb{R}^m$ is a solution to $(\mathbf{P}_W^1)$. Then, $\psi\neq 0$ is true, so that any solution ψ is a regular point for $(\mathbf{P}_W^1)$; compare Definition D.2. Consequently, there exists a Lagrange multiplier associated with the optimal solution ψ, so that the first-order necessary optimality condition

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

is satisfied; see Theorem D.4. We compute the gradient of $\mathcal{L}$ with respect to ψ: Since W is symmetric, we derive

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

Thus,

$$\nabla_\psi\mathcal{L}(\psi,\lambda)=2\big(WYY^\top W\psi-\lambda W\psi\big)\overset{!}{=}0\quad\text{in }\mathbb{R}^m.\tag{1.3.3}$$

Equation (1.3.3) yields the generalized eigenvalue problem

$$(WY)(WY)^\top\psi=\lambda W\psi.\tag{1.3.4}$$

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

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

Note that $(W^\alpha)^{-1}=W^{-\alpha}$ and $W^{\alpha+\beta}=W^\alpha W^\beta$ for $\alpha,\,\beta\in\mathbb{R}$; see Exercise 1.5.8. Moreover, we have

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

and $\|\psi\|_W=\|W^{1/2}\psi\|_{\mathbb{R}^m}$ for $\psi\in\mathbb{R}^m$. Setting $\bar\psi=W^{1/2}\psi\in\mathbb{R}^m$ and $\bar Y=W^{1/2}Y\in\mathbb{R}^{m\times n}$ and multiplying (1.3.4) by $W^{-1/2}$ from the left we deduce the

symmetric, $m \times m$ eigenvalue problem

$$\bar{Y}\bar{Y}^\top \bar{\psi} = \lambda \bar{\psi} \quad \text{in } \mathbb{R}^m. \tag{1.3.5a}$$

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

$$\|\bar{\psi}\|_{\mathbb{R}^m} = 1. \tag{1.3.5b}$$

Thus, the first-order optimality conditions (1.3.5) for $(\mathbf{P}^1_W)$ are — as for $(\mathbf{P}^1)$ (compare (1.1.7)) — an $m \times m$ eigenvalue problem, but the matrix Y as well as the vector ψ have to be weighted by the matrix $W^{1/2}$.

Notice that $\nabla_{\psi\psi}\mathcal{L}(\psi, \lambda) = 2(WYY^\top W\psi - \lambda W\psi) \in \mathbb{R}^{m\times m}$. Let $\psi \in \mathbb{R}^m$ be chosen arbitrary. Since $\bar{Y}\bar{Y}^\top$ is symmetric, there exist m orthonomal (with respect to the Euclidean inner product) eigenvectors $\bar{\psi}_1, \dots, \bar{\psi}_m \in \mathbb{R}^m$ of $\bar{Y}\bar{Y}^\top$ satisfying $\bar{Y}\bar{Y}^\top \bar{\psi}_i = \lambda_i \bar{\psi}_i$ for $1 \le i \le m$. We set $\psi_i = W^{-1/2}\bar{\psi}_i$, $1 \le i \le m$. Then, $\{\psi_i\}_{i=1}^m$ form an orthonormal (with respect to the weighted inner product) basis in $\mathbb{R}^m$ and $WYY^\top W\psi_i = \lambda_i W\psi_i$ holds true. We write ψ in the form

$$\psi = \sum_{j=1}^m \langle \psi, \psi_i \rangle_{\mathbb{R}^m} \psi_i.$$

At (ψ_1, λ_1) we conlude from $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_m \ge 0$ that

$$\begin{aligned}
&\psi^\top \nabla_{\psi\psi}\mathcal{L}(\psi_1, \lambda_1)\psi = 2\psi^\top \big(WYY^\top W - \lambda_1 W\big)\psi \\
&= 2\sum_{i=1}^m \sum_{j=1}^m \langle \psi, \psi_i \rangle_{\mathbb{R}^m} \langle \psi, \psi_j \rangle_{\mathbb{R}^m} \psi_i^\top \big(WYY^\top W - \lambda_1 W\big)\psi_j \\
&= 2\sum_{i=1}^m \sum_{j=1}^m \big(\lambda_j - \lambda_1\big) \langle \psi, \psi_i \rangle_{\mathbb{R}^m} \langle \psi, \psi_j \rangle_{\mathbb{R}^m} \psi_i^\top W\psi_j \\
&= 2\sum_{i=1}^m \big(\lambda_j - \lambda_1\big) \big|\langle \psi, \psi_i \rangle_{\mathbb{R}^m}\big|^2 \le 0.
\end{aligned}$$

Thus, the matrix $\nabla_{\psi\psi}\mathcal{L}(\psi_1, \lambda_1)$ is negative semi-definite, which is the second-order necessary optimality condition; compare Theorem D.5. However, analogously to Section 1 it can be shown (see Exercise 1.4.1)) that

$$\psi_1 = W^{-1/2}\bar{\psi}_1$$

solves $(\mathbf{P}^1_W)$, where $\bar{u}_1$ is an eigenvector of $\bar{Y}\bar{Y}^\top$ corresponding to the largest eigenvalue λ_1 with $\|\bar{\psi}_1\|_{\mathbb{R}^m} = 1$. Due to SVD the vector ψ_1 can be also determined by solving the symmetric $n \times n$ eigenvalue problem

$$\bar{Y}^\top \bar{Y}\bar{\phi}_1 = \lambda_1 \bar{\phi}_1$$

where $\bar{Y}^\top \bar{Y} = Y^\top WY$, and setting

$$\psi_1 = W^{-1/2}\bar{\psi}_1 = \frac{1}{\sqrt{\lambda_1}} W^{-1/2}\bar{Y}\bar{\phi}_1 = \frac{1}{\sqrt{\lambda_1}} Y\bar{\phi}_1. \tag{1.3.6}$$

As in Section 1 we can continue by looking at a second vector $\psi \in \mathbb{R}^m$ with $\langle \psi, \psi_1 \rangle_W = 0$ that maximizes $\sum_{j=1}^n |\langle y_j, \psi \rangle_W|^2$. Let us generalize Theorem 1.1.1 as follows; see Exercise 1.5.9.

THEOREM 1.3.2. *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 $\bar{Y} = \bar{\Psi}\Sigma\bar{\Phi}^\top$ be the singular value decomposition of $\bar{Y}$, where $\bar{\Psi} = [\bar{\psi}_1,\dots,\bar{\psi}_m] \in \mathbb{R}^{m\times m}$, $\bar{\Phi} = [\bar{\phi}_1,\dots,\bar{\phi}_n] \in \mathbb{R}^{n\times n}$ are orthogonal matrices and the matrix Σ has the form*

$$\bar{\Psi}^\top \bar{Y} \bar{\Phi} = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} = \Sigma \in \mathbb{R}^{m\times n}.$$

Then the solution to

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

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

$$\operatorname{argmax}(\mathbf{P}_W^\ell) = \sum_{i=1}^{\ell} \sigma_i^2 = \sum_{i=1}^{\ell} \lambda_i. \tag{1.3.7}$$

PROOF. Using similar arguments as in the proof of Theorem 1.1.1 one can prove that $\{\psi_i\}_{i=1}^\ell$ solves $(\mathbf{P}_W^\ell)$; see Exercise 1.4.8. □

REMARK 1.3.3. Due to SVD and $\bar{Y}^\top\bar{Y} = Y^\top W Y$ the POD basis $\{\psi_i\}_{i=1}^\ell$ of rank ℓ can be determined by the method of snapshots as follows: Solve the symmetric $n \times n$ eigenvalue problem

$$Y^\top W Y \bar{\phi}_i = \lambda_i \bar{\phi}_i \quad \text{for } i = 1,\dots,\ell,$$

and set

$$\psi_i = W^{-1/2}\bar{\psi}_i = \frac{1}{\sqrt{\lambda_i}} W^{-1/2}\big(\bar{Y}\bar{\phi}_i\big) = \frac{1}{\sqrt{\lambda_i}} W^{-1/2} W^{1/2} Y \bar{\phi}_i = \frac{1}{\sqrt{\lambda_i}} Y \bar{\phi}_i$$

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

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

For $m \gg n$ the method of snapshots turns out to be faster than computing the POD basis via (1.3.5). Notice that the matrix $W^{1/2}$ is also not required for the method of snapshots. ◊

In Algorithm 3 we extend Algorithm 2.

4. POD for Time-Dependent Systems

For $T > 0$ we consider the semi-linear initial value problem

$$\dot{y}(t) = Ay(t) + f(t, y(t)) \quad \text{for } t \in (0,T], \tag{1.4.1a}$$

$$y(0) = y_\circ, \tag{1.4.1b}$$

where $y_\circ \in \mathbb{R}^m$ 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 (1.4.1) has a unique (classical) solution $y \in C^1(0,T_*;\mathbb{R}^m) \cap C([0,T_*];\mathbb{R}^m)$ for some maximal time $T_* \in (0,T]$.

Algorithm 3 (POD basis of rank ℓ with a weighted inner product)

Require: Snapshots $\{y_j\}_{j=1}^n \subset \mathbb{R}^m$, POD rank $\ell \leq d$, symmetric, positive-definite matrix $W \in \mathbb{R}^{m\times m}$ and `flag` for the solver;
Set $Y = [y_1, \ldots, y_n] \in \mathbb{R}^{m\times n}$;
if `flag` $= 0$ **then**
  Determine $\bar{Y} = W^{1/2}Y \in \mathbb{R}^{m\times n}$;
  Compute singular value decomposition $[\bar{\Psi}, \Sigma, \bar{\Phi}] = \mathrm{svd}\,(\bar{Y})$;
  Set $\psi_i = W^{-1/2}\bar{\Psi}_{\cdot,i} \in \mathbb{R}^m$ and $\lambda_i = \Sigma_{ii}^2$ for $i = 1, \ldots, \ell$;
else if `flag` $= 1$ **then**
  Determine $\bar{Y} = W^{1/2}Y \in \mathbb{R}^{m\times n}$;
  Set $R = \bar{Y}\bar{Y}^\top \in \mathbb{R}^{m\times m}$;
  Compute eigenvalue decomposition $[\bar{\Psi}, \Lambda] = \mathrm{eig}\,(R)$;
  Set $\psi_i = W^{-1/2}\bar{\Psi}_{\cdot,i} \in \mathbb{R}^m$ and $\lambda_i = \Lambda_{ii}$ for $i = 1, \ldots, \ell$;
else if `flag` $= 2$ **then**
  Determine $K = Y^\top W Y \in \mathbb{R}^{n\times n}$;
  Compute eigenvalue decomposition $[\bar{\Phi}, \Lambda] = \mathrm{eig}\,(K)$;
  Set $\psi_i = Y\bar{\Phi}_{\cdot,i}/\sqrt{\lambda_i} \in \mathbb{R}^m$ and $\lambda_i = \Lambda_{ii}$ for $i = 1, \ldots, \ell$;
end if
Compute $\mathcal{E}(\ell) = \sum_{i=1}^{\ell} \lambda_i / \sum_{i=1}^{d} \lambda_i$;
return POD basis $\{\psi_i\}_{i=1}^{\ell}$, eigenvalues $\{\lambda_i\}_{i=1}^{\ell}$ and ratio $\mathcal{E}(\ell)$;

Throughout we suppose that we can choose $T_* = T$. Then, the solution y to (1.4.1) is given by the implicit integral representation

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

with $e^{tA} = \sum_{n=0}^{\infty} t^n A^n/(n!)$.

4.1. Application of POD for Time-Dependent Systems. Let $0 \leq t_1 < t_2 < \ldots < t_n \leq T$ 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 (1.4.1) at the given time instances t_j, $j \in \{1, \ldots, n\}$. Our goal is to determine a POD basis of rank $\ell \leq \min\{m, n\}$ that describes the ensemble

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

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

$$(\hat{\mathbf{P}}_W^{n,\ell}) \qquad \min_{\tilde{\psi}_1,\ldots,\tilde{\psi}_\ell \in \mathbb{R}^m} \sum_{j=1}^{n} \alpha_j \left\| y_j - \sum_{i=1}^{\ell} \langle y_j, \tilde{\psi}_i \rangle_W \tilde{\psi}_i \right\|_W^2$$
$$\text{s.t. } \langle \tilde{\psi}_i, \tilde{\psi}_j \rangle_W = \delta_{ij},\ 1 \leq i, j \leq \ell,$$

where the α_j's denote nonnegative weights which will be specified later on. Note that for $\alpha_j = 1$ for $j = 1, \ldots, n$ and $W = I_m$ problem $(\hat{\mathbf{P}}_W^{n,\ell})$ coincides with (1.2.3).

EXAMPLE 1.4.1. Let us consider the following one-dimensional heat equation:

$$\theta_t(t,x) = \theta_{xx}(t,x) \qquad \text{for all } (t,x) \in Q = (0,T) \times \Omega, \tag{1.4.2a}$$

$$\theta_x(t,0) = \theta_x(t,1) = 0 \qquad \text{for all } t \in (0,T), \tag{1.4.2b}$$

$$\theta(0,x) = \theta_\circ(x) \qquad \text{for all } x \in \Omega = (0,1) \subseteq \mathbb{R}, \tag{1.4.2c}$$

where $\theta_\circ \in C(\overline{\Omega})$ is a given initial condition. To solve (1.4.2) numerically we apply a classical finite difference approximation for the spatial variable x. In Example 1.3.1 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 (1.4.2a) and the boundary conditions (1.4.2b) 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:

$$\left\{ \begin{array}{ll} \dot{y}_1(t) & = \dfrac{-2y_1(t) + 2y_2(t)}{h^2}, \\ \dot{y}_i(t) & = \dfrac{y_{i-1}(t) - 2y_i(t) + y_{i+1}(t)}{h^2}, \quad i = 2, \ldots, m-1, \\ \dot{y}_m(t) & = \dfrac{-2y_m(t) + 2y_{m-1}(t)}{h^2} \end{array} \right. \tag{1.4.3a}$$

for $t \in (0,T]$. From (1.4.2c) we infer the initial condion

$$y_i(0) = \theta_\circ(x_i), \quad i = 1, \ldots, m. \tag{1.4.3b}$$

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

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

we can express (1.4.3) in the form

$$\begin{array}{ll} \dot{y}(t) & = Ay(t) \quad \text{for } t \in (0,T], \\ y(0) & = y_\circ \end{array} \tag{1.4.4}$$

Setting $f \equiv 0$ the linear initial-value problem coincides with (1.4.1). 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 (1.3.2); see Example 1.3.1. Next we choose a time grid $\{t_j\}_{j=1}^n$ in the interval $[0,T]$ and define $y_j = y(t_j)$ for $j = 1, \ldots, n$. If we are interested in finding a POD basis of rank $\ell \leq \min\{m,n\}$ that desribes the ensemble $\{y_j\}_{j=1}^n$ as well as possible, we end up with $(\hat{\mathbf{P}}_W^{n,\ell})$. $\Diamond$

To solve $(\hat{\mathbf{P}}_W^{n,\ell})$ we apply the techniques used in Sections 1 and 3, i.e., we use the Lagrangian framework; see Appendix D. 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

$$\mathcal{L}(\psi_1, \ldots, \psi_\ell, \Lambda) = \sum_{j=1}^{n} \alpha_j \left\| y_j - \sum_{i=1}^{\ell} \langle y_j, \psi_i \rangle_W \psi_i \right\|_W^2 + \sum_{i=1}^{\ell}\sum_{j=1}^{\ell} \Lambda_{ij}\big(1 - \langle \psi_i, \psi_j \rangle_W\big)$$

for $\psi_1, \ldots, \psi_\ell \in \mathbb{R}^m$ 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 conditions

$$\nabla_{\psi_i} \mathcal{L}(\psi_1, \ldots, \psi_\ell, \Lambda) \overset{!}{=} 0 \quad \text{in } \mathbb{R}^m,\ 1 \le i \le \ell, \tag{1.4.5a}$$

and

$$\langle \psi_i, \psi_j \rangle_W \overset{!}{=} \delta_{ij}, \quad 1 \le i, j \le \ell; \tag{1.4.5b}$$

compare Theorem D.4. From (1.4.5a) we derive

$$YDY^\top W\psi_i = \lambda_i \psi_i \quad \text{for } i = 1, \ldots, \ell, \tag{1.4.6}$$

where $D = \operatorname{diag}(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^{n\times n}$. Inserting $\psi_i = W^{-1/2}\bar{\psi}_i$ in (1.4.6) and multiplying (1.4.6) by $W^{1/2}$ from the left yield

$$W^{1/2}YDY^\top W^{1/2}\bar{\psi}_i = \lambda_i \bar{\psi}_i. \tag{1.4.7a}$$

From (1.4.5b) we find

$$\langle \bar{\psi}_i, \bar{\psi}_j \rangle_{\mathbb{R}^m} = \bar{\psi}_i^\top \bar{\psi}_j = \psi_i^\top W \psi_j = \langle \psi_i, \psi_j \rangle_W = \delta_{ij}, \quad 1 \le i, j \le \ell. \tag{1.4.7b}$$

Setting $\bar{Y} = W^{1/2}YD^{1/2} \in \mathbb{R}^{m\times n}$ and using $W^\top = W$ as well as $D^\top = D$ we infer from (1.4.7) that the solution $\{\psi_i\}_{i=1}^{\ell}$ to $(\hat{\mathbf{P}}_W^{n,\ell})$ is given by the symmetric $m \times m$ eigenvalue problem

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

Note that

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

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

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

Then we set (by SVD)

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

compare (1.3.6). Note that

$$\langle \psi_i, \psi_j \rangle_W = \psi_i^T W \psi_j = \frac{1}{\sqrt{\lambda_i\lambda_j}} \bar{\psi}_i^\top \underbrace{D^{1/2}Y^\top WYD^{1/2}}_{=\bar{Y}^\top\bar{Y}} \bar{\phi}_j = \frac{\lambda_i}{\sqrt{\lambda_i\lambda_j}} \bar{\phi}_i^\top \bar{\phi}_j = \frac{\lambda_i \delta_{ij}}{\sqrt{\lambda_i\lambda_j}}$$

for $1 \le i, j \le \ell$, i.e., the POD basis vectors $\psi_1, \ldots, \psi_\ell$ are orthonormal in $\mathbb{R}^m$ with respect to the inner product $\langle \cdot, \cdot \rangle_W$.

In Algorithm 4 the computation of a POD basis of rank ℓ is summarized for finite-dimensional dynamical systems.

Algorithm 4 (POD basis of rank ℓ for finite-dimensional dynamical systems)

Require: Snapshots $\{y_j\}_{j=1}^n \subset \mathbb{R}^m$, POD rank $\ell \leq d$, symmetric, positive-definite matrix $W \in \mathbb{R}^{m\times m}$, diagonal matrix $D \in \mathbb{R}^{n\times n}$ containing the temporal quadrature weights and `flag` for the solver;

Set $Y = [y_1, \dots, y_n] \in \mathbb{R}^{m\times n}$;
if `flag` $= 0$ **then**
  Determine $\bar{Y} = W^{1/2} Y D^{1/2} \in \mathbb{R}^{m\times n}$;
  Compute singular value decomposition $[\bar{\Psi}, \Sigma, \bar{\Phi}] = \operatorname{svd}(\bar{Y})$;
  Set $\psi_i = W^{-1/2}\bar{\Psi}_{\cdot,i} \in \mathbb{R}^m$ and $\lambda_i = \Sigma_{ii}^2$ for $i = 1, \dots, \ell$;
else if `flag` $= 1$ **then**
  Determine $\bar{Y} = W^{1/2} Y D^{1/2} \in \mathbb{R}^{m\times n}$;
  Set $R = \bar{Y}\bar{Y}^\top \in \mathbb{R}^{m\times m}$;
  Compute eigenvalue decomposition $[\bar{\Psi}, \Lambda] = \operatorname{eig}(R)$;
  Set $\psi_i = W^{-1/2}\bar{\Psi}_{\cdot,i} \in \mathbb{R}^m$ and $\lambda_i = \Lambda_{ii}$ for $i = 1, \dots, \ell$;
else if `flag` $= 2$ **then**
  Determine $K = D^{1/2} Y^\top W Y D^{1/2} \in \mathbb{R}^{n\times n}$;
  Compute eigenvalue decomposition $[\bar{\Phi}, \Lambda] = \operatorname{eig}(K)$;
  Set $\psi_i = Y D^{1/2} \bar{\Phi}_{\cdot,i} / \sqrt{\lambda_i} \in \mathbb{R}^m$ and $\lambda_i = \Lambda_{ii}$ for $i = 1, \dots, \ell$;
end if
Compute $\mathcal{E}(\ell) = \sum_{i=1}^{\ell} \lambda_i / \sum_{i=1}^{d} \lambda_i$;
return POD basis $\{\psi_i\}_{i=1}^{\ell}$, eigenvalues $\{\lambda_i\}_{i=1}^{\ell}$ and ratio $\mathcal{E}(\ell)$;

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

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

Moreover, we have not discussed so far what is the motivation to introduce the nonnegative 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 nonnegative weights $\{\alpha_j\}_{j=1}^n$?

To address these two questions we will introduce a *continuous version* of POD. Suppose that (1.4.1) has a unique solution $y : [0, T] \to \mathbb{R}^m$. 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

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

To solve $(\hat{\mathbf{P}}_W^{\ell})$ we use similar arguments as in Sections 1 and 3. For $\ell = 1$ we obtain instead of $(\hat{\mathbf{P}}_W^{\ell})$ the minimization problem

$$\min_{\tilde{\psi} \in \mathbb{R}^m} \int_0^T \big\| y(t) - \langle y(t), \tilde{\psi} \rangle_W \, \tilde{\psi} \big\|_W^2 \, \mathrm{d}t \quad \text{s.t.} \quad \|\tilde{\psi}\|_W^2 = 1, \tag{1.4.8}$$

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

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

Thus,

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

we conclude that (1.4.8) is equivalent with the following maximization problem

$$\max_{\tilde{\psi} \in \mathbb{R}^m} \int_0^T \big| \langle y(t), \tilde{\psi} \rangle_W \big|^2 \, \mathrm{d}t \quad \text{s.t.} \quad \|\tilde{\psi}\|_W^2 = 1. \tag{1.4.9}$$

The Lagrange functional $\mathcal{L} : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ associated with (1.4.9) is given by

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

Arguing as in Sections 1 and 3 any optimal solution to (1.4.9) is a regular point; see Exercise 1.5.10. Consequently, first-order necessary optimality conditions are given by

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

Therefore, we compute the partial derivative of $\mathcal{L}$ with respect to the i-th component ψ_i of the vector ψ:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \psi_i}(u, \lambda) &= \frac{\partial}{\partial \psi_i} \bigg(\int_0^T \Big| \sum_{k=1}^{m} \sum_{\nu=1}^{m} y_k(t) W_{k\nu} \psi_\nu \Big|^2 \, \mathrm{d}t + \lambda \Big(1 - \sum_{k=1}^{m} \sum_{\nu=1}^{m} \psi_k W_{k\nu} \psi_\nu \Big) \bigg) \\ &= 2 \int_0^T \Big(\sum_{k=1}^{m} \sum_{\nu=1}^{m} y_k(t) W_{k\nu} \psi_\nu \Big) \sum_{\mu=1}^{m} y_\mu(t) W_{\mu i} \, \mathrm{d}t - 2\lambda \sum_{k=1}^{m} W_{ik} \psi_k \\ &= 2 \bigg(\int_0^T \langle y(t), \psi \rangle_W W y(t) \, \mathrm{d}t - \lambda W \psi \bigg)_i \end{aligned}$$

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

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

which gives

$$\int_0^T \langle y(t), \psi \rangle_W \, W y(t) \, \mathrm{d}t = \lambda W \psi \quad \text{in } \mathbb{R}^m. \tag{1.4.10}$$

Multiplying (1.4.10) by W^{-1} from the left yields

$$\int_0^T \langle y(t), \psi \rangle_W \, y(t) \, \mathrm{d}t = \lambda \psi \quad \text{in } \mathbb{R}^m. \tag{1.4.11}$$

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

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

LEMMA 1.4.2. *The operator $\mathcal{R}$ is linear and bounded (i.e., continuous). Moreover,*

1) *$\mathcal{R}$ is nonnegative:*

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

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

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

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

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

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

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

for an arbitrary $\psi \in \mathbb{R}^m$. 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

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

for all $\psi \in \mathbb{R}^m$ holds, $\mathcal{R}$ is nonnegative. Finally, we infer from

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

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

Utilizing the operator $\mathcal{R}$ we can write (1.4.11) as the eigenvalue problem

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

It follows from Lemma 1.4.2 that $\mathcal{R}$ possesses eigenvectors $\{\psi_i\}_{i=1}^m$ and associated real eigenvalues $\{\lambda_i\}_{i=1}^m$ such that

$$\mathcal{R}\psi_i = \lambda_i\psi_i \text{ for } 1 \leq i \leq m \quad \text{and} \quad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0. \tag{1.4.12}$$

Note that

$$\int_0^T \left|\langle y(t),\psi_i\rangle_W\right|^2 \mathrm{d}t = \int_0^T \left\langle \langle y(t),\psi_i\rangle_W\, y(t),\psi_i\right\rangle_W \mathrm{d}t = \langle \mathcal{R}\psi_i,\psi_i\rangle_W = \lambda_i \left\|\psi_i\right\|_W^2 = \lambda_i \quad \text{for } i\in\{1,\dots,m\}$$

so that ψ_1 solves (1.4.8). Proceeding as in Sections 1 and 3 we obtain the following result; see Exercise 1.5.11.

THEOREM 1.4.3. *Suppose that* (1.4.1) *has a unique solution* $y:[0,T]\to\mathbb{R}^m$. *Then the POD basis of rank* ℓ *solving the minimization problem* $(\hat{\mathbf{P}}_W^\ell)$ *is given by the eigenvectors* $\{\psi_i\}_{i=1}^\ell$ *of* $\mathcal{R}$ *corresponding to the* ℓ *largest eigenvalues* $\lambda_1\geq\ldots\geq\lambda_\ell$.

REMARK 1.4.4 (Methods of snapshots). Let us define the linear and bounded operator $\mathcal{Y}:L^2(0,T)\to\mathbb{R}^m$ by

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

The (Hilbert space) adjoint $\mathcal{Y}^\star:\mathbb{R}^m\to L^2(0,T)$ satisfying (see Definition A.5)

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

is given as

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

Then we have

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

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

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

for all $\phi\in L^2(0,T)$ 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, nonnegative and self-adjoint. Moreover, $\mathcal{K}$ is compact. Therefore, the POD basis can also be computed as follows: Solve

$$\mathcal{K}\phi_i=\lambda_i\phi_i \text{ for } 1\leq i\leq\ell,\quad \lambda_1\geq\ldots\geq\lambda_\ell>0,\quad \int_0^T \phi_i(t)\phi_j(t)\,\mathrm{d}t=\delta_{ij} \tag{1.4.13}$$

and set

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

Note that (1.4.13) is a symmetric eigenvalue problem in the infinite-dimensional function space $L^2(0,T)$. ◊

In Algorithm 5 the computation of a POD basis of rank ℓ is summarized in the context of the continuous version of the POD method.

Algorithm 5 (POD basis of rank ℓ for dynamical systems [continuous version])

Require: Snapshots $\{y(t)\,|\,t \in [0,T]\} \subset \mathbb{R}^m$, POD rank $\ell \leq d$, symmetric, positive-definite matrix $W \in \mathbb{R}^{m\times m}$ and `flag` for the solver;

if `flag` $= 1$ **then**
  Set $\mathcal{R} = \int_0^T \langle y(t), \bullet\rangle_W \, y(t)\,\mathrm{d}t \in L(\mathbb{R}^m)$;
  Solve the eigenvalue problem $\mathcal{R}\psi_i = \lambda_i\psi_i$, $1 \leq i \leq \ell$, with $\langle \psi_i, \psi_j\rangle_W = \delta_{ij}$;
else if `flag` $= 2$ **then**
  Set $\mathcal{K} = \int_0^T \langle y(s), y(\cdot)\rangle_W \bullet \,\mathrm{d}s \in L(L^2(0,T))$;
  Solve the problem $\mathcal{K}\phi_i = \lambda_i\phi_i$, $1 \leq i \leq \ell$, with $\langle \phi_i, \phi_j\rangle_{L^2(0,T)} = \delta_{ij}$;
  Set $\psi_i = \int_0^T y(t)\phi_i(t)\,\mathrm{d}t/\sqrt{\lambda_i} \in \mathbb{R}^m$;
end if
Compute $\mathcal{E}(\ell) = \sum_{i=1}^{\ell}\lambda_i / \sum_{i=1}^{d}\lambda_i$;
return POD basis $\{\psi_i\}_{i=1}^{\ell}$, eigenvalues $\{\lambda_i\}_{i=1}^{\ell}$ and ratio $\mathcal{E}(\ell)$;

Let us turn back to the optimality conditions (1.4.6). For any $\psi \in \mathbb{R}^m$ and $i \in \{1,\ldots,m\}$ we derive

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

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

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

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

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

holds for all $\psi \in \mathbb{R}^m$ so that $\mathcal{R}^n$ is nonnegative. Further,

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

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

$$\mathcal{R}^n\psi_i^n = \lambda_i^n\psi_i^n, \quad \lambda_1^n \geq \ldots\lambda_\ell^n \geq \ldots\lambda_{d(n)}^n > \lambda_{d(n)+1}^n = \ldots = \lambda_m^n = 0, \tag{1.4.14a}$$

$$\mathcal{R}\psi_i = \lambda_i\psi_i, \quad \lambda_1 \geq \ldots\lambda_\ell \geq \ldots\lambda_d > \lambda_{d+1} = \ldots = \lambda_m = 0. \tag{1.4.14b}$$

Let us note that

$$\int_0^T \|y(t)\|_W^2\,\mathrm{d}t = \sum_{i=1}^{d}\lambda_i = \sum_{i=1}^{m}\lambda_i. \tag{1.4.15}$$

In fact,

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

Taking the inner product with u_i, using (1.4.14b) and summing over i we arrive at

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

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

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

and hence

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

which is (1.4.15). Analogously, we obtain

$$\sum_{j=1}^{n} \alpha_j \, \|y(t_j)\|_W^2 = \sum_{i=1}^{d(n)} \lambda_i^n = \sum_{i=1}^{m} \lambda_i^n \quad \text{for every } n \in \mathbb{N}; \tag{1.4.16}$$

see Exercise 1.5.12. 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

$$\sum_{j=1}^{n} \alpha_j \, \|y(t_j)\|_W^2 \to \int_0^T \|y(t)\|_W^2 \, \mathrm{d}t \quad \text{as } \Delta t \to 0 \tag{1.4.17}$$

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

$$\alpha_1 = \frac{\Delta t}{2}, \ \alpha_j = \Delta t \text{ for } 2 \le j \le n-1, \ \alpha_n = \frac{\Delta t}{2}. \tag{1.4.18}$$

Suppose that we have

$$\lim_{n \to \infty} \|\mathcal{R}^n - \mathcal{R}\|_{L(\mathbb{R}^m)} = \lim_{n \to \infty} \sup_{\|\psi\|_W = 1} \|\mathcal{R}^n \psi - \mathcal{R}\psi\|_W = 0 \tag{1.4.19}$$

provided $y \in C^1([0,T]; \mathbb{R}^m)$ is satisfied. In (1.4.19) we denote by $L(\mathbb{R}^m)$ the Banach space of all linear and bounded operators mapping from $\mathbb{R}^m$ into itself; see Appendix A. Combining (1.4.17) with (1.4.15) and (1.4.16) we find

$$\sum_{i=1}^{m} \lambda_i^n \to \sum_{i=1}^{m} \lambda_i \quad \text{as } n \to \infty. \tag{1.4.20}$$

Now choose and fix

$$\ell \quad \text{such that} \quad \lambda_\ell \neq \lambda_{\ell+1}. \tag{1.4.21}$$

Then by spectral analysis of compact operators [**13**, pp. 212-214] and (1.4.19) it follows that

$$\lambda_i^n \to \lambda_i \quad \text{for } 1 \le i \le \ell \text{ as } n \to \infty. \tag{1.4.22}$$

Combining (1.4.20) and (1.4.22) there exists $\bar{n} \in \mathbb{N}$ such that

$$\sum_{i=\ell+1}^{m} \lambda_i^n \le 2 \sum_{i=\ell+1}^{m} \lambda_i \quad \text{for all } n \ge \bar{n}, \tag{1.4.23}$$

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

$$\sum_{i=\ell+1}^{d(n)} \left|\langle y_\circ, \psi_i^n\rangle_W\right|^2 \leq 2 \sum_{i=\ell+1}^{m} \left|\langle y_\circ, \psi_i\rangle_W\right|^2 \quad \text{for all } n \geq \bar{n}, \tag{1.4.24}$$

provided that $\sum_{i=\ell+1}^m |\langle y_\circ, \psi_i\rangle_W|^2 \neq 0$ and (1.4.19) hold. Recall that the vector $y_\circ \in \mathbb{R}^m$ stands for the initial condition in (1.4.1b). Then we have

$$\|y_\circ\|_W^2 = \sum_{i=1}^{m} \left|\langle y_\circ, \psi_i\rangle_W\right|^2. \tag{1.4.25}$$

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

$$\|y_\circ\|_W^2 = \sum_{i=1}^{d(n)} \left|\langle y_\circ, \psi_i^n\rangle_W\right|^2. \tag{1.4.26}$$

Therefore, for $\ell < d(n)$ by (1.4.25) and (1.4.26)

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

As a consequence of (1.4.19) and (1.4.21) we have $\lim_{n\to\infty} \|\psi_i^n - \psi_i\|_W = 0$ for $i = 1, \ldots, \ell$ and hence (1.4.24) follows.

Summarizing we have the following theorem.

THEOREM 1.4.5. *Suppose that* (1.4.1) *has a unique solution* $y : [0, T] \to \mathbb{R}^m$. *Let* $\{(\psi_i^n, \lambda_i^n)\}_{i=1}^m$ *and* $\{(\psi_i, \lambda_i)\}_{i=1}^m$ *be the eigenvector-eigenvalue pairs given by* (1.4.14). *Suppose that* $\ell \in \{1, \ldots, m\}$ *is fixed such that* (1.4.21) *and*

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

hold. Then we have

$$\lim_{n\to\infty} \|\mathcal{R}^n - \mathcal{R}\|_{L(\mathbb{R}^m)} = 0. \tag{1.4.27}$$

This implies

$$\lim_{n\to\infty} \left|\lambda_i^n - \lambda_i\right| = \lim_{n\to\infty} \|\psi_i^n - \psi_i\|_W = 0 \quad \textit{for } 1 \leq i \leq \ell,$$

$$\lim_{n\to\infty} \sum_{i=\ell+1}^{m} \left(\lambda_i^n - \lambda_i\right) = 0 \quad \textit{and} \quad \lim_{n\to\infty} \sum_{i=\ell+1}^{m} \left|\langle y_\circ, \psi_i^n\rangle_W\right|^2 = \sum_{i=\ell+1}^{m} \left|\langle y_\circ, \psi_i\rangle_W\right|^2.$$

PROOF. We only have to verify (1.4.27). For that purpose we choose an arbitrary $\psi \in \mathbb{R}^m$ with $\|\psi\|_W = 1$ and introduce $f_\psi : [0, T] \to \mathbb{R}^m$ by

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

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

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

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

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

Hence,

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

Consequently,

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

which is (1.4.27). □

5. Exercises

Exercise 1.5.1. Let $A \in \mathbb{R}^{m\times n}$, $m > n$, a matrix with rank n. Suppose that $\Psi^\top A \Phi = \Sigma$ is the singular value decomposition of A with the singular values $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_n > 0$. Prove the following claims:

a) $A\phi_i = \sigma_i\psi_i$ and $A^\top\psi_i = \sigma_i\phi_i$ for $i = 1, \ldots, n$, where $\{\psi_i\}_{i=1}^m \subset \mathbb{R}^m$ and $\{\phi_i\}_{i=1}^n \subset \mathbb{R}^n$ denote the columns of $\Psi \in \mathbb{R}^{m\times m}$ and $\Phi \in \mathbb{R}^{n\times n}$, respectively.
b) $\|A\|_2 = \sigma_1$.
c) The matrix $A^\top A$ is symmetric and prositive definite.
d) The set of all positive singular values of A coincides with the set of square roots of all positive eigenvalues of $A^\top A$.

Exercise 1.5.2. Assume that $A \in \mathbb{R}^{n\times n}$ is an invertible matrix and that $A = \Psi\Sigma\Phi^\top$ is a singular value decomposition on A. What is the singular value decomposition of A^{-1}?

Exercise 1.5.3. Compute the singular value decomposition of the matrix

$$A = \begin{pmatrix} -2 & 0 \\ 0 & 1 \\ 0 & -1 \end{pmatrix}.$$

Exercise 1.5.4. Show that any optimal solution to $(\mathbf{P}^\ell)$ is a regular point.

Exercise 1.5.5. Verify the claim in Theorem 1.1.1 that $\operatorname{argmax}(\mathbf{P}^\ell) = \sum_{i=1}^\ell \sigma_i^2$ holds true.

Exercise 1.5.6. Show that the Frobenius norm is a matrix norm and that

$$\|AB\|_F \le \|A\|_F\|B\|_F \quad \text{for any } A,\, B \in \mathbb{R}^{n\times n}$$

is valid. Suppose that $\Psi^d \in \mathbb{R}^{m\times d}$ is a matrix with pairwise orthonormal vectors $\psi_i \in \mathbb{R}^m$, $1 \le i \le d$. Prove that

$$\|\Psi^d A\|_F = \|A\|_F \quad \text{for any matrix } A \in \mathbb{R}^{d\times n}.$$

Exercise 1.5.7. We extend Example 1.3.1 to the two-dimensional domain $\Omega = (0,1)\times(0,1) \subset \mathbb{R}^2$ be given. We choose the trapezoidal quadrature rule with an equidistant grid size $h = 1/(n-1)$ in both dimensions. Determine the weighting matrix $W \in \mathbb{R}^{m\times m}$, where $m = n^2$ holds, so that the trapezoidal approximation can be written as the weighted inner product $\langle\cdot\,,\cdot\rangle_W$.

Exercise 1.5.8. Suppose that $W \in \mathbb{R}^{m\times m}$ is symmetric and positive definite. Let $\eta_1 \ge \ldots \ge \eta_m > 0$ denote the eigenvalues of W and $W^\alpha = Q\operatorname{diag}(\eta_1^\alpha, \ldots, \eta_m^\alpha)Q^\top$ be the eigenvalue decomposition of W. We define

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

Show that $(W^\alpha)^{-1}$ exists and $(W^\alpha)^{-1} = W^{-\alpha}$. Prove that $W^{\alpha+\beta} = W^\alpha W^\beta$ holds for $\alpha,\, \beta \in \mathbb{R}$.

Exercise 1.5.9. Verify the claims of Theorem 1.3.2.

a) Ensure a regular point condition, which guarantees the existence of Lagrange multiplieres.
b) Prove that $\psi_i = W^{-1/2}\bar\psi_i$, $1 \le i \le \ell$, solves $(\mathbf{P}_W^\ell)$, where the matrix W and the vectors $\bar\psi_1, \ldots, \bar\psi_m$ are introduced in Theorem 1.3.2.
c) Show that (1.3.7) holds.

Exercise 1.5.10. Agrue that any optimal solution to (1.4.9) is a regular point.

Exercise 1.5.11. Prove that u_1 given by (1.4.12) is a global solution to (1.4.8). How can this result be extended for $(\hat{\mathbf{P}}_W^\ell)$?

Exercise 1.5.12. Verify (1.4.16).

CHAPTER 2

The POD Method for Partial Differential Equations

In this chapter we formulate the POD method for partial differential equations (PDEs). For that purpose an extension of the approach presented in Chapter 1 to separable Hilbert spaces is needed. Our approach is motivated by the goal to derive reduced-order models for parabolic and elliptic partial differential equations. In Section 1 we focus on parabolic PDEs. The presented approach generalizes the theory presented in Section 4 of Chapter 1. We also discuss the numerical realization as well as the treatment of nonlinearities. Parametrized elliptic problems are analyzed in Section 2. Whereas for parabolic problems the time t serves as the sampling parameter, a variation of the parameter values are used for elliptic problems to build a POD basis.

Throughout this chapter we make use of the following notations and assumptions: Let V and H be real, separable Hilbert spaces and suppose that V is dense in H with compact embedding. By $\langle \cdot\,, \cdot \rangle_V$ and $\langle \cdot\,, \cdot \rangle_H$ we denote the inner products in V and H with associated norm $\|\cdot\|_V = \sqrt{\langle \cdot\,, \cdot \rangle_V}$ and $\|\cdot\|_H = \sqrt{\langle \cdot\,, \cdot \rangle_H}$, respectively.

1. POD for Parabolic Partial Differential Equations

Now we consider the POD method for linear evolution problems. Then, its numerical approximation is dicussed. Moreover, we explain the extension to nonlinear evolution problems by using empirical interpolation.

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

$$\big|a(t; \varphi, \psi)\big| \leq \beta \, \|\varphi\|_V \|\psi\|_V, \tag{2.1.1a}$$

$$a(t; \varphi, \varphi) \geq \kappa \, \|\varphi\|_V^2 - \eta \, \|\varphi\|_H^2 \tag{2.1.1b}$$

for all $\varphi, \psi \in V$ and t, t_1, $t_2 \in [0, T]$, where $\beta, \kappa > 0$ and $\eta \geq 0$ are constants, which do not depend on t. By identifying H with its dual H' it follows that

$$V \hookrightarrow H = H' \hookrightarrow V',$$

each embedding being continuous and dense. In Appendix B we introduce the function space $W(0, T)$, which is a Hilbert space endowed with the common inner product. When the time t is fixed, the expression $\varphi(t)$ stands for the function $\varphi(t, \cdot)$ considered as a function in Ω only.

For $y_\circ \in H$ and $f \in L^2(0,T;V')$ we consider the linear evolution problem

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\,\langle y(t),\varphi\rangle_H + a(t;y(t),\varphi) &= \langle f(t),\varphi\rangle_{V',V} \ \text{ f.a.a. } t\in[0,T],\ \forall \varphi\in V,\\ \langle y(0),\varphi\rangle_H &= \langle y_\circ,\varphi\rangle_H \qquad \forall \varphi\in V. \end{aligned} \tag{2.1.2}$$

Throughout we write 'f.a.a.' for 'for almost all'.

EXAMPLE 2.1.1. Suppose that $\Omega \subsetneq \mathbb{R}^{\mathsf{d}}$, $\mathsf{d}\in\{1,2,3\}$, is an open and bounded domain with Lipschitz-continuous boundary $\Gamma = \partial\Omega$. For $T>0$ we set $Q = (0,T)\times\Omega$ and $\Sigma = (0,T)\times\Gamma$. Let $H = L^2(\Omega)$ and $V = H^1(\Omega)$. Then, for given $y_\circ \in H$, $\mathrm{f}\in L^2(0,T;H)$ and $\mathrm{g}\in L^2(0,T;L^2(\Gamma_C))$, we consider the linear heat equation

$$\begin{aligned} y_t(t,\boldsymbol{x}) - \nabla\cdot\big(\mathrm{c}(t,\boldsymbol{x})\nabla y(t,\boldsymbol{x})\big) + \mathrm{a}(t,\boldsymbol{x})y(t,\boldsymbol{x}) &= \mathrm{f}(t,\boldsymbol{x}), \quad (t,\boldsymbol{x})\in Q,\\ \mathrm{c}(t,\boldsymbol{s})\frac{\partial y}{\partial n}(t,\boldsymbol{s}) &= \mathrm{g}(t,\boldsymbol{s}), \quad (t,\boldsymbol{s})\in\Sigma,\\ y(0,\boldsymbol{x}) &= y_\circ(\boldsymbol{x}), \quad \boldsymbol{x}\in\Omega, \end{aligned} \tag{2.1.3}$$

where $\mathrm{c}\in C([0,T];L^\infty(\Omega))$ satisfying $\mathrm{c}(t,\boldsymbol{x})\geq \mathrm{c}_a > 0$ f.a.a. $(t,\boldsymbol{x})\in Q$, $\mathrm{a}\in C([0,T];L^\infty(\Omega))$ and $\mathrm{b}\in L^\infty(0,T;L^\infty(\Gamma_C))$. For $t\in[0,T]$ a.e. we introduce the bilinear form $a(t;\cdot\,,\cdot): V\times V\to\mathbb{R}$ by

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

and the linear, bounded functional $f\in L^2(0,T;V')$ by

$$\langle f(t),\varphi\rangle_{V',V} = \langle \mathrm{f}(t),\varphi\rangle_H + \int_{\Gamma_N} \mathrm{g}(t)\varphi\,\mathrm{d}\boldsymbol{s} \quad \text{for } t\in[0,T] \text{ a.e. and } \varphi,\psi\in V,$$

where "a.e." stands for "almost everywhere". Then, it follows that the weak formulation of (2.1.3) can be expressed in the form (2.1.2). From $\mathrm{c},\mathrm{a}\in C([0,T];L^\infty(\Omega))$ we infer that the time-dependent bilinear form $a(t;\cdot\,,\cdot)$ satisfies (2.1.1). ♢

EXAMPLE 2.1.2. Let us present a further example for (2.1.2). Suppose that – as in Example 2.1.1 – the set $\Omega\subset\mathbb{R}^{\mathsf{d}}$, $\mathsf{d}\in\{1,2,3\}$, is an open and bounded domain with Lipschitz-continuous boundary $\Gamma=\partial\Omega$. For $T>0$ we set $Q=(0,T)\times\Omega$ and $\Sigma=(0,T)\times\Gamma$. Let $H=L^2(\Omega)$ and $V=H_0^1(\Omega)$. Then, for given initial condition $y_\circ\in H$ we consider the linear heat equation

$$y_t(t,\boldsymbol{x}) - \nabla\cdot\big(\mathrm{c}(t,\boldsymbol{x})\nabla y(t,\boldsymbol{x})\big) + \mathrm{a}(t,\boldsymbol{x})y(t,\boldsymbol{x}) = \mathrm{f}(t,\boldsymbol{x}), \quad (t,\boldsymbol{x})\in Q, \tag{2.1.4a}$$

$$y(t,\boldsymbol{s}) = 0, \quad (t,\boldsymbol{s})\in\Sigma, \tag{2.1.4b}$$

$$y(0,\boldsymbol{x}) = y_\circ(\boldsymbol{x}), \quad \mathbf{x}\in\Omega. \tag{2.1.4c}$$

In (2.1.4a) we suppose that c, a and f satisfies the same assumptions as in Example 2.1.1. Introducing the bilinear form $a(t;\cdot\,,\cdot): V\times V\to\mathbb{R}$ for every $t\in[0,T]$ by

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

it follows that the weak formulation of (2.1.4) can be written in the form (2.1.2).♢

It follows from Theorem C.1 that for every $f\in L^2(0,T;V')$ and $y_\circ\in H$ there exists a unique weak solution $y\in W(0,T)$ satisfying (2.1.2). Moreover, if $f\in L^2(0,T;H)$, $a(t;\cdot\,,\cdot)=a(\cdot\,,\cdot)$ (independent of t) and $y_\circ\in V$ hold, we even have $y\in C([0,T];V)$; see Corollary C.3.

1.2. The Continuous POD Method for Linear Evolution Equations. Let $f \in L^2(0,T;V')$ and $y_\circ \in V$ be given arbitrarily so that the solution $y \in W(0,T)$ to (2.1.2) belongs to $C([0,T];V) \hookrightarrow C([0,T];X)$, where X denotes either the space V or the space H. Then,

$$\mathcal{V} = \operatorname{span}\big\{ y(t) \,|\, t \in [0,T] \big\} \subseteq V \subset X. \tag{2.1.5}$$

If $y_\circ \neq 0$ holds, then $\mathcal{V} \neq \{0\}$ and $d = \dim \mathcal{V} \in [1,\infty]$, but $\mathcal{V}$ may have infinite dimension. Now we proceed similar as in Remark 1.4.4. We define a bounded linear operator $\mathcal{Y} : L^2(0,T) \to X$ 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 $\mathcal{Y}^\star : X \to L^2(0,T)$ satisfying

$$\langle \mathcal{Y}\varphi, \psi \rangle_X = \langle \varphi, \mathcal{Y}^\star \psi \rangle_{L^2(0,T)} \quad \text{for } (\varphi,\psi) \in L^2(0,T) \times X$$

is given by $(\mathcal{Y}^\star\psi)(t) = \langle \psi, y(t) \rangle_X$ for $\psi \in X$ and f.a.a. $t \in [0,T]$. The linear operator $\mathcal{R} = \mathcal{Y}\mathcal{Y}^\star : X \to \mathcal{V} \subset X$ has the form

$$\mathcal{R}\psi = \int_0^T \langle \psi, y(t) \rangle_X \, y(t) \,\mathrm{d}t \quad \text{for } \psi \in X. \tag{2.1.6}$$

Moreover, let $\mathcal{K} = \mathcal{Y}^\star\mathcal{Y} : L^2(0,T) \to L^2(0,T)$ be defined by

$$\big(\mathcal{K}\phi\big)(t) = \int_0^T \langle y(s), y(t) \rangle_X \, \phi(s) \,\mathrm{d}s \quad \text{for } \phi \in L^2(0,T). \tag{2.1.7}$$

LEMMA 2.1.3. *Let X denote either the space V or the space H and $y \in W(0,T)$ hold. Then, the linear operator $\mathcal{R}$ is bounded, compact, nonnegative and symmetric.*

PROOF. Applying the Cauchy-Schwarz inequality we infer that

$$\begin{aligned} \|\mathcal{R}\psi\|_X &\leq \int_0^T \big| \langle \psi, y(t) \rangle_X \big| \, \|y(t)\|_X \,\mathrm{d}t \leq \|\psi\|_X \int_0^T \|y(t)\|_X^2 \,\mathrm{d}t \\ &= \|y\|_{L^2(0,T;X)}^2 \, \|\psi\|_X \quad \text{for } \psi \in X \end{aligned} \tag{2.1.8}$$

holds. By assumption, $y \in W(0,T) \subset L^2(0,T;X)$. Thus, from (2.1.8) we infer that $\mathcal{R}$ is bounded. Again using $y \in W(0,T) \subset L^2(0,T;X)$ the kernel $k(s,t) = \langle y(t), y(s) \rangle_X$ of $\mathcal{K}$ is square integrable over $(0,T) \times (0,T)$; see Exercise 2.3.1. By Remark A.14 we conclude that the integral operator $\mathcal{K}$ is compact. Remark A.16 implies that $\mathcal{R} = \mathcal{K}^\star$ is compact as well. From

$$\langle \mathcal{R}\psi, \psi \rangle_X = \left\langle \int_0^T \langle \psi, y(t) \rangle_X \, y(t) \,\mathrm{d}t, \psi \right\rangle_X = \int_0^T \big| \langle \psi, y(t) \rangle_X \big|^2 \,\mathrm{d}t \geq 0 \text{ for all } \psi \in X$$

we infer that $\mathcal{R}$ is nonnegative. Finally, we have

$$\begin{aligned} \langle \mathcal{R}\psi, \tilde{\psi} \rangle_X &= \left\langle \int_0^T \langle \psi, y(t) \rangle_X \, y(t) \,\mathrm{d}t, \tilde{\psi} \right\rangle_X = \int_0^T \langle \psi, y(t) \rangle_X \, \langle y(t), \tilde{\psi} \rangle_X \mathrm{d}t \\ &= \int_0^T \big\langle \psi, \langle y(t), \tilde{\psi} \rangle_X \, y(t) \big\rangle_X \,\mathrm{d}t = \left\langle \psi, \int_0^T \langle y(t), \tilde{\psi} \rangle_X \, y(t) \,\mathrm{d}t \right\rangle_X \\ &= \langle \psi, \mathcal{R}\tilde{\psi} \rangle_X \quad \text{for all } \psi, \tilde{\psi} \in X. \end{aligned}$$

Hence, the operator $\mathcal{R}$ is selfadjoint. □

From Theorems A.17 and A.18 it follows that there exists a complete orthonormal basis $\{\psi_i\}_{i=1}^{\infty}$ for X and a sequence $\{\lambda_i\}_{i=1}^{\infty}$ of nonnegative real numbers such that

$$\mathcal{R}\psi_i = \lambda_i \psi_i, \quad \lambda_1 \geq \lambda_2 \geq \ldots, \quad \text{and} \quad \lim_{i\to\infty} \lambda_i = 0. \tag{2.1.9}$$

The spectrum of $\mathcal{R}$ is a pure point spectrum except for possibly 0. Each nonzero eigenvalue of $\mathcal{R}$ has finite multiplicity and 0 is the only possible accumulation point of the spectrum of $\mathcal{R}$. Let us note that

$$\int_0^T \|y(t)\|_X^2 \,\mathrm{d}t = \sum_{i=1}^{\infty} \lambda_i \quad \text{and} \quad \|y_\circ\|_X = \sum_{i=1}^{\infty} \left|\langle y_\circ, \psi_i\rangle_X\right|^2.$$

REMARK 2.1.4. 1) Analogously to the theory of singular value decomposition for matrices, we find that the linear, bounded, compact and self-adjoint operator $\mathcal{K}$ has the same eigenvalues $\{\lambda_i\}_{i=1}^{d}$ as the operator $\mathcal{R}$. For all $\lambda_i > 0$ the corresponding eigenfunctions of $\mathcal{K}$ are given by

$$\phi_i(t) = \frac{1}{\sqrt{\lambda_i}} \left(\mathcal{Y}^\star \psi_i\right)(t) = \frac{1}{\sqrt{\lambda_i}} \langle \psi_i, y(t)\rangle_X \text{ f.a.a. } t \in [0,T] \text{ and } 1 \leq i \leq \ell.$$

2) Notice that – independent of the choice for X — $\mathcal{V} \subset V$ implies $\psi_i \in V$ for $1 \leq i \leq \ell$. ◊

In the following theorem we formulate properties of the eigenvalues and eigenvectors of $\mathcal{R}$.

THEOREM 2.1.5. *Let $\{\lambda_i\}_{i=1}^{\infty}$ and $\{\psi_i\}_{i=1}^{\infty}$ denote the eigenvalues and eigenfunctions, respectively, of $\mathcal{R}$. Then, for every $\ell \in \mathbb{N}$ the first ℓ eigenfunctions $\psi_1, \ldots, \psi_\ell \in X$ solve the minimization problem*

$$\begin{cases} \min\limits_{\tilde\psi_1,\ldots,\tilde\psi_\ell \in X} \displaystyle\int_0^T \Big\| y(t) - \sum_{i=1}^{\ell} \langle y(t), \tilde\psi_i\rangle_X \, \tilde\psi_i \Big\|_X^2 \,\mathrm{d}t \\ \quad \textit{s.t. } \langle \tilde\psi_i, \tilde\psi_j\rangle_X = \delta_{ij} \textit{ for } 1 \leq i,j \leq \ell. \end{cases} \tag{2.1.10}$$

Moreover,

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

PROOF. We proceed as in the proof of Theorem 1.1.1. First note that (2.1.10) is equivalent to

$$\begin{cases} \max\limits_{\tilde\psi_1,\ldots,\tilde\psi_\ell \in X} \displaystyle\sum_{i=1}^{\ell} \int_0^T \left|\langle y(t), \tilde\psi_i\rangle_X\right|^2 \mathrm{d}t \\ \quad \text{s.t. } \langle \tilde\psi_i, \tilde\psi_j\rangle_X = \delta_{ij} \text{ for } 1 \leq i,j \leq \ell. \end{cases} \tag{2.1.11}$$

We define the Lagrange function

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

by

$$\mathcal{L}(\psi_1, \ldots, \psi_\ell, \Lambda) = \sum_{i=1}^{\ell} \int_0^T \left|\langle y(t), \tilde\psi_i\rangle_X\right|^2 + \sum_{i,j}^{\ell} \lambda_{ij} \left(\delta_{ij} - \langle \psi_i, \psi_j\rangle_X\right)$$

for $\psi_1, \dots, \psi_\ell \in X$ and $\Lambda = ((\lambda_{ij})) \in \mathbb{R}^{\ell\times\ell}$. To derive first-order optimality conditions we show a constraint qualification for (2.1.11). For that purpose let us introduce the mapping

$$e : \underbrace{X \times \ldots \times X}_{\ell\text{-times}} \to \mathbb{R}^{\ell\times\ell}, \quad e(\tilde\psi_1, \dots, \tilde\psi_\ell) = \delta_{ij} - \langle \tilde\psi_i, \tilde\psi_j \rangle_X$$

Then, we have $e(\tilde\psi_1, \dots, \tilde\psi_\ell) = e(\tilde\psi_1, \dots, \tilde\psi_\ell)^T$. This reflects the fact that we can replace the constraints in (2.1.11) by $\langle \tilde\psi_i, \tilde\psi_j \rangle_X = \delta_{ij}$ for $1 \le i \le \ell$ and $1 \le j \le i$. Moreover, introducing the set of feasible solutions

$$X_{ad}^\ell = \big\{ (\tilde\psi_1, \dots, \tilde\psi_\ell) \,\big|\, \tilde\psi_1, \dots, \tilde\psi_\ell \in X \text{ and } e(\tilde\psi_1, \dots, \tilde\psi_\ell) = 0 \in \mathbb{R}^{\ell\times\ell} \big\}$$

problem (2.1.11) can be expressed as

$$\max \sum_{i=1}^{\ell} \int_0^T \big| \langle y(t), \tilde\psi_i \rangle_X \big|^2 \quad \text{s.t.} \quad (\tilde\psi_1, \dots, \tilde\psi_\ell) \in X_{ad}^\ell. \tag{2.1.12}$$

To derive first-order optimality conditions we show a constraint qualification for the set X_{ad}^ℓ. The Fréchet derivative of e is given by

$$e'(\psi_1, \dots, \psi_\ell)(\psi_1^\delta, \dots, \psi_\ell^\delta) = \big(\big(- \langle \psi_i^\delta, \psi_j \rangle_X - \langle \psi_i, \psi_j^\delta \rangle_X \big) \big)_{1\le i,j\le \ell}$$

for given directions $\psi_1^\delta, \dots, \psi_\ell^\delta \in X$. Suppose that $\{\psi_i\}_{i=1}^\ell$ satisfies $\langle \tilde\psi_i, \tilde\psi_j \rangle_X = \delta_{ij}$ for $1 \le i, j \le \ell$ and that $A = ((a_{ij})) \in \mathbb{R}^{\ell\times\ell}$ is a symmetric matrix. Then a constraint qualification holds at $\{\psi_i\}_{i=1}^\ell$ provided there exists an ℓ-tupel $(\psi_1^\delta, \dots, \psi_\ell^\delta)$ such that

$$e'(\psi_1, \dots, \psi_\ell)(\psi_1^\delta, \dots, \psi_\ell^\delta) = A.$$

Choosing $\psi_i^\delta = -\sum_{k=1}^\ell a_{ik}\psi_k/2$ for $1 \le i \le \ell$ and using $\langle \psi_i, \psi_j \rangle_X = \delta_{ij}$ we have

$$\begin{aligned} -\langle \psi_i^\delta, \psi_j \rangle_X - \langle \psi_i, \psi_j^\delta \rangle_X &= \frac{1}{2} \sum_{k=1}^{\ell} \Big(a_{ik} \langle \psi_k, \psi_j \rangle_X + a_{jk} \langle \psi_i, \psi_k \rangle_X \Big) \\ &= \frac{1}{2} \big(a_{ij} + a_{ji} \big) = a_{ij} \quad \text{for } i, j \in \{1, \dots, \ell\}. \end{aligned}$$

Thus, $\{\psi_i\}_{i=1}^\ell$ satisfies a constraint qualification so that first-order necessary optimality conditions are given by setting the Fréchet derivative of the Lagrangian equal to zero; see Theorem D.4 in the Appendix. Instead of (1.1.10) we get

$$\mathcal{R}\psi_k = \int_0^T \langle y(t), \psi_k \rangle_X \, y(t) \, \mathrm{d}t = \frac{1}{2} \sum_{i=1}^{\ell} (\lambda_{ik} + \lambda_{ki}) \, \psi_i \quad \text{in } X \text{ for all } k \in \{1, \dots, \ell\}.$$

Therefore, we can follow the lines of the proof of Theorem 1.1.1 and conlude that $\lambda_{i,\ell} = -\lambda_{\ell,i}$ for $1 \le i \le \ell - 1$. Setting $\lambda_i = \lambda_{ii}$ for all $i \in \{1, \le i \le \ell\}$ the first-order necessary optimality conditions for (2.1.12) – and hence also for (2.1.10) – are given by

$$\mathcal{R}\psi_i = \lambda_i \psi_i \quad \text{in } X \text{ for all } i \in \{1, \dots, \ell\}. \tag{2.1.13}$$

It follows that $\{\psi_i\}_{i=1}^\ell$ solves (2.1.13). The proof that $\{\psi_i\}_{i=1}^\ell$ is a solution to (2.1.12) and that $\operatorname{argmax}(2.1.12) = \sum_{i=1}^\ell \lambda_i^2$ holds is analogous to the proof for $(\mathbf{P}^1)$; see Exercise 2.3.2. □

1.3. The Truth Approximation for Linear Evolution Problems. To compute the POD basis $\{\psi_i\}_{i=1}^{\ell}$ as described in Section 1.2 we need the snapshots $y(t)$ for $t \in [0,T]$. This is realized numerically by computing approximations for $y(t)$ using a spatial and temporal discretization method. First, we consider the case that the snapshots are given by finite element approximations. In a second step we turn to the temporal discretization.

1.3.1. *Spatial discretization.* Let us introduce a spatial discretization for (2.1.2). For $m \in \mathbb{N}$ the functions $\varphi_1, \dots, \varphi_m$ denote m linearly independent nodal basis functions. Then we define the m-dimensional subspace

$$V^h = \operatorname{span}\{\varphi_1, \dots, \varphi_m\} \subset V$$

endowed with the topology in V. We apply a standard Galerkin scheme for (2.1.2). Thus, we look for a function $y^h \in L^2(0,T;V^h) \cap H^1(0,T;V_h')$ satisfying

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\langle y^h(t), \varphi^h\rangle_H + a(t; y^h(t), \varphi^h) &= \langle f(t), \varphi^h\rangle_{V',V}, \ t \in [0,T], \ \forall \varphi^h \in V^h, \\ \langle y^h(0), \varphi^h\rangle_H &= \langle y_\circ, \varphi^h\rangle_H \qquad \forall \varphi^h \in V^h. \end{aligned} \tag{2.1.14}$$

Since $y^h(t) \in V^h$ holds, we make the Galerkin ansatz of the form

$$y^h(t) = \sum_{i=1}^{m} \mathfrak{y}_i^h(t)\varphi_i \tag{2.1.15}$$

and define the modal coefficient vector

$$\mathfrak{y}^h(t) = \big(\mathfrak{y}_i^h(t)\big)_{1\le i\le m} \quad \text{for } t \in [0,T].$$

From (2.1.14) we derive the linear system of ordinary differential equations

$$M\dot{\mathfrak{y}}^h(t) + A(t)\mathfrak{y}^h(t) = b(t) \text{ f.a.a. } t \in [0,T], \quad M\mathfrak{y}^h(0) = \mathfrak{y}_\circ \tag{2.1.16}$$

with

$$\begin{aligned} M_{ij} &= \langle \varphi_j, \varphi_i\rangle_H, & (A(t))_{ij} &= a(t; \varphi_j, \varphi_i), \\ (\mathfrak{y}_\circ)_i &= \langle y_\circ, \varphi_i\rangle_H, & (b(t))_i &= \langle f(t), \varphi_i\rangle_{V',V}, \end{aligned}$$

for $1 \le i,j \le m$ and $t \in [0,T]$. System (2.1.16) is referred to as the *truth approximation* for (2.1.2). Note that (2.1.16) can then be solved by using an appropriate method for the time discretization. System (2.1.16) can be written in the form (1.4.1) with $A = 0$ and

$$f(t,\mathfrak{y}) = M^{-1}\big(b(t) - A(t)y\big), \quad (t,y) \in [0,T] \times \mathbb{R}^m.$$

From (2.1.1) it follows that (2.1.16) has a unique solution $\mathfrak{y} \in H^1(0,T;\mathbb{R}^m)$; see Exercise 2.3.3. If $f \in C([0,T];V')$ holds and $t \mapsto a(t;\varphi,\phi)$ is continuous for any $\varphi, \psi \in V$, then $\mathfrak{y} \in C([0,T];\mathbb{R}^m)$ and we can proceed as in Section 4.2 of Chapter 1.

REMARK 2.1.6. Suppose that $\mathfrak{u} = (\mathfrak{u}_i)_{1\le i\le m}$ and $\mathfrak{v} = (\mathfrak{v}_i)_{1\le i\le m}$ are two arbitrary vectors in $\mathbb{R}^m$. Then,

$$u^h(x) = \sum_{i=1}^{m} \mathfrak{u}_i\varphi_i(x) \quad \text{and} \quad v^h(x) = \sum_{i=1}^{m} \mathfrak{v}_i\varphi_i(x)$$

are elements in the finite element space V^h. We have

$$\langle u^h, v^h\rangle_H = \langle \mathfrak{u}, \mathfrak{v}\rangle_W \quad \text{and} \quad \|u^h\|_H = \|\mathfrak{u}\|_W$$

with $W = M$, where the symmetric, positive definite mass matrix has been introduced above. Analogously,we obtain

$$\langle u^h, v^h \rangle_V = \langle \mathfrak{u}, \mathfrak{v} \rangle_W \quad \text{and} \quad \|u^h\|_V = \|\mathfrak{u}\|_W$$

with $W = S$, where the symmetric, positive definite stiffness matrix is given by

$$S_{ij} = \langle \varphi_j, \varphi_i \rangle_V \quad \text{for } 1 \le i, j \le m$$

Summarizing, the weighted inner product $\langle \cdot\,, \cdot \rangle_W$ is used to replace the inner products in the m-dimensional finite element space V^h by an inner product in $\mathbb{R}^m$ for the finite element nodal coefficients. ◊

Let $\mathcal{V} = \operatorname{span}\{\mathfrak{y}^h(t) \,|\, t \in [0,T]\} \subset \mathbb{R}^m$ and $d = \dim \mathcal{V} \le m$. For any $\ell \in \{1, \dots, d\}$ we construct a low-dimensional orthonormal basis by solving the optimization problem

$$\min_{\tilde\psi_1,\dots,\tilde\psi_\ell \in \mathbb{R}^m} \int_0^T \Big\| \mathfrak{y}^h(t) - \sum_{i=1}^{\ell} \langle \mathfrak{y}^h(t), \tilde\psi_i \rangle_W \, \tilde\psi_i \Big\|_W^2 \, \mathrm{d}t \tag{2.1.17}$$
$$\text{s.t. } \langle \tilde\psi_i, \tilde\psi_j \rangle_W = \delta_{ij} \text{ for } 1 \le i, j \le \ell.$$

The solution to (2.1.17) is given by the theory presented in Section 4.2 of Chapter 1. Thus, let us define the linear, bounded, nonnegative and selfadjoint operator $\mathcal{R}^h : \mathbb{R}^m \to \mathbb{R}^m$ by

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

Now the solution to (2.1.17) is given by the eigenvectors corresponding to the d largest (positive) eigenvalues $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_d > 0$ solving the symmetric $m \times m$ eigenvalue problem

$$\mathcal{R}^h \psi_i = \lambda_i \psi_i \quad \text{for } i = 1, \dots, d. \tag{2.1.18}$$

Again, we can quantify the POD approximation error as follows

$$\int_0^T \Big\| \mathfrak{y}^h(t) - \sum_{i=1}^{\ell} \langle \mathfrak{y}^h(t), \psi_i \rangle_W \, \psi_i \Big\|_W^2 \, \mathrm{d}t = \sum_{i=\ell+1}^{d} \lambda_i.$$

1.3.2. *Temporal discretization.* In real computations we do not have the whole trajectory $\mathfrak{y}^h(t) \in \mathbb{R}^m$ (or $y^h(t) \in V^h$) for $t \in [0,T]$. For this purpose let $0 = t_1 < t_2 < \ldots < t_n = T$ be a given grid in $[0,T]$ with step sizes $\delta_j = t_j - t_{j-1}$ for $2 \le j \le n$. To solve (2.1.16) we apply an implicit Euler method for the time integration. Of course, other time intergation schemes can be used; see Exercises 2.3.4 and 2.3.5. The sequence $\{\mathfrak{y}_j^h\}_{j=1}^n$ in $\mathbb{R}^m$ is the solution to

$$\big(M + \delta t_j A(t_j)\big)\mathfrak{y}_j^h = M\mathfrak{y}_{j-1}^h + \delta t_j b(t_j) \text{ for } 2 \le j \le n, \quad M\mathfrak{y}^h = \mathfrak{y}_\circ. \tag{2.1.19}$$

Remark 2.1.7. We set

$$y_j^h = \sum_{i=1}^{m} (\mathfrak{y}_j^h)_i \varphi_i \in V^h \quad \text{for } 1 \le j \le n$$

the Galerkin functions y_j^h are approximations for the solution y^h to (2.1.14) at time $t = t_j$. Then, $\{y_j^h\}_{j=1}^n \subset V^h$ satisfies

$$\langle \overline{\partial}_j y_j^h, \varphi^h \rangle_H + a(t_j; y_j^h, \varphi^h) = \langle f(t_j), \varphi^h \rangle_{V',V} \text{ for } 2 \le j \le n, \ \forall \varphi^h \in V^h,$$
$$\langle y_1^h, \varphi^h \rangle_H = \langle y_\circ, \varphi^h \rangle_H \qquad \forall \varphi^h \in V^h.$$

where $\overline{\partial}_j y_j^h = (y_j^h - y_{j-1}^h)/\delta t_j \in V^h$ stands for the backward difference quotient. $\Diamond$

Analogous to Section 1.3.1 we set $\mathcal{V} = \operatorname{span}\{\mathfrak{y}_j^h \,|\, 1 \le j \le n\} \subset \mathbb{R}^m$ and $d = \dim \mathcal{V} \le \min(m,n)$. For any $\ell \in \{1,\dots,d\}$ we construct a low-dimensional orthonormal basis by solving the optimization problem

$$\begin{aligned} \min_{\tilde{\psi}_1,\dots,\tilde{\psi}_\ell \in \mathbb{R}^m} & \sum_{j=1}^n \alpha_j \left\| \mathfrak{y}_j^h - \sum_{i=1}^\ell \langle \mathfrak{y}_j^h, \tilde{\psi}_i \rangle_W \, \tilde{\psi}_i \right\|_W^2 \\ \text{s.t. } & \langle \tilde{\psi}_i, \tilde{\psi}_j \rangle_W = \delta_{ij} \text{ for } 1 \le i,j \le \ell. \end{aligned} \tag{2.1.20}$$

The solution to (2.1.20) is given by the theory presented in Section 4.1 of Chapter 1. As in (2.1.6), let us define the linear, bounded, nonnegative and selfadjoint operator $\mathcal{R}^{h,n} : \mathbb{R}^m \to \mathbb{R}^m$ by

$$\mathcal{R}^{h,n}\psi = \sum_{j=1}^n \alpha_j \, \langle \mathfrak{y}_j^h, \psi \rangle_W \, \mathfrak{y}_j^h \quad \text{for } \psi \in \mathbb{R}^m.$$

Now the solution to (2.1.20) is given by the eigenvectors corresponding to the d largest (positive) eigenvalues $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_d > 0$ solving the symmetric $m \times m$ eigenvalue problem

$$\mathcal{R}^{h,n}\psi_i = \lambda_i \psi_i \quad \text{for } i = 1,\dots,d.$$

Again, we can quantify the POD approximation error as follows

$$\sum_{j=1}^n \alpha_j \left\| \mathfrak{y}_j^h - \sum_{i=1}^\ell \langle \mathfrak{y}_j^h, \psi_i \rangle_W \, \psi_i \right\|_W^2 = \sum_{i=\ell+1}^d \lambda_i.$$

In Exercise 2.3.6 a POD basis is computed for the truth approximation of the heat equation.

REMARK 2.1.8. In [**15**] an asymptotic analysis is carried analogously to the approach presented in Section 4.2 in Chapter 1. $\Diamond$

1.4. POD for Nonlinear Evolution Equations. The application of the POD method can easily be extended for nonlinear evolution problems. Let $\mathcal{N} : [0,T] \times V \to V'$ be a given nonlinearity and $y_\circ \in H$. Instead of (2.1.2) we consider

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \langle y(t), \varphi \rangle_H + a(t; y(t), \varphi) &= \langle \mathcal{N}(t, y(t)), \varphi \rangle_{V',V}, \ t \in [0,T], \ \forall \varphi \in V, \\ \langle y(0), \varphi \rangle_H &= \langle y_\circ, \varphi \rangle_H \qquad\qquad \forall \varphi \in V. \end{aligned} \tag{2.1.21}$$

We suppose that (2.1.21) possesses a unique solution $y \in W(0,T)$. In the following example we present two applications considered in the literature.

REMARK 2.1.9. 1) *A monotonous nonlinearity*: Let the bilinear form a be independent of t, i.e., $a(t;\cdot,\cdot) = a(\cdot,\cdot)$. Moreover $\mathcal{B} : V \to V'$ is a continuous nonlinear operator satisfying

$$a(\varphi,\varphi) + \langle \mathcal{B}(\varphi), \varphi \rangle_{V',V} \ge \kappa \, \|\varphi\|_V^2 - \eta \, \|\varphi\|_H^2 \quad \text{for all } \varphi \in V$$

for constants $\kappa > 0$ and $\eta \ge 0$. For $f \in L^2(0,T;V')$ we set $\mathcal{N}(t,\varphi) = f(t) - \mathcal{B}(\varphi) \in V'$ f.a.a. $t \in [0,T]$ and for $\varphi \in V$. Then, the POD method problem for (2.1.21) is considered in [**14**], for instance.

2) *A general equation in fluid dynamics* [**21**]: Let the bilinear form a be independent of t, i.e., $a(t;\cdot\,,\cdot) = a(\cdot\,,\cdot)$. As in Appendix C we define the linear bounded operator $\mathcal{A} : V \to V'$ by

$$\langle \mathcal{A}\varphi, \phi\rangle_{V',V} = a(\varphi,\phi) \quad \text{for } \varphi,\phi \in V.$$

The domain of $\mathcal{A}$ is given by the set $D(\mathcal{A}) = \{\varphi \in V \,|\, \mathcal{A}\varphi \in H\}$. Further, let us introduce the continuous operator $\mathcal{B} : V \to V'$, which maps $D(\mathcal{A})$ into H and satisfies

$$\begin{aligned} \|\mathcal{B}\varphi\|_H &\le C_{\mathcal{B}}\,\|\varphi\|_V^{1-\delta_1}\|\mathcal{A}\varphi\|_H \quad \text{for all } \varphi \in D(\mathcal{A}),\\ \big|\langle \mathcal{B}\varphi,\varphi\rangle_{V',V}\big| &\le C_{\mathcal{B}}\,\|\varphi\|_V^{1+\delta_2}\|\varphi\|_H^{1-\delta_2} \quad \text{for all } \varphi \in V \end{aligned} \tag{2.1.22}$$

for a constant $C_{\mathcal{B}} > 0$ and for $\delta_1, \delta_2 \in [0,1)$. We also assume that $\mathcal{A}+\mathcal{B}$ satisfies

$$a(\varphi,\varphi) + \langle \mathcal{B}\varphi,\varphi\rangle_{V',V} \ge \kappa\,\|\varphi\|_V^2 - \eta\,\|\varphi\|_H^2 \quad \text{for all } \varphi \in V \tag{2.1.23}$$

with constants $\kappa > 0$ and $\eta \ge 0$. Moreover, let $\mathcal{C} : V \times V \to V'$ be a bilinear continuous operator mapping from $D(\mathcal{A}) \times D(\mathcal{A})$ into H such that there exist constants $C_{\mathcal{C}} > 0$ and $\delta_3, \delta_4, \delta_5 \in [0,1)$ satisfying

$$\begin{aligned} \langle \mathcal{C}(\varphi,\phi),\phi\rangle_{V',V} &= 0,\\ \big|\langle \mathcal{C}(\varphi,\phi),\psi\rangle_{V',V}\big| &\le C_{\mathcal{C}}\,\|\varphi\|_H^{\delta_3}\|\varphi\|_V^{1-\delta_3}\|\phi\|_V^{\delta_3}\|\psi\|_V^{\delta_3}\|\psi\|_H^{\delta_3},\\ \|\mathcal{C}(\varphi,\chi\|_H + \|\mathcal{C}(\chi,\varphi)\|_H &\le C_{\mathcal{C}}\,\|\varphi\|_V\|\chi\|_V^{1-\delta_4}\|\mathcal{A}\chi\|_H^{\delta_4},\\ \|\mathcal{C}(\varphi,\chi)\|_H &\le C_{\mathcal{C}}\,\|\varphi\|_H^{\delta_5}\|\varphi\|_V^{1-\delta_5}\|\chi\|_V^{1-\delta_5}\|\mathcal{A}\chi\|_H^{\delta_5} \end{aligned} \tag{2.1.24}$$

for all $\varphi,\phi,\psi \in V$ and for all $\chi \in D(\mathcal{A})$. Setting

$$\mathcal{N}(t,\varphi) = f(t) - \mathcal{B}\varphi - \mathcal{C}(\varphi,\varphi) \quad \text{f.a.a. } t \in [0,T] \text{ and for } \varphi \in V \tag{2.1.25}$$

problem (2.1.21) is studied in [**15, 16**]. In particular, it is proved in [**21**] that the two-dimensional Navier-Stokes equations can be expressed in the form (2.1.21) taking the nonlinearity (2.1.25). ◊

Let $y \in W(0,T)$ be a solution to (2.1.21). For the snapshot set $\mathcal{V} = \{y(t) \,|\, t \in [0,T]\}$ a POD basis of rank ℓ can be determined as described in Section 1.2. Proceeding as in Section 1.3 we can also compute a POD basis for the aproximate solutions to the nonlinear equations. For later reference we state here the spatial discretization following the arguments in Section 1.3.1. Instead of (2.1.14) the solution $y^h \in L^2(0,T;V^h) \cap H^1(0,T;V_h')$ satisfies

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\,\langle y^h(t),\varphi^h\rangle_H + a(t;y^h(t),\varphi^h) &= \langle \mathcal{N}(t,y(t)),\varphi^h\rangle_{V',V},\\ &\qquad \text{f.a.a. } t \in [0,T],\ \forall \varphi^h \in V^h,\\ \langle y^h(0),\varphi^h\rangle_H &= \langle y_\circ,\varphi^h\rangle_H \ \ \forall \varphi^h \in V^h. \end{aligned} \tag{2.1.26}$$

To derive a system of ordinary differential equations for the coefficients $\mathfrak{y}^h(t) \in \mathbb{R}^m$ of the Galerkin ansatz $y^h(t) = \sum_{i=1}^m \mathfrak{y}_i^h(t)\varphi_i \in V^h$ we introduce the nonlinearity $\mathfrak{f} : [0,T] \times \mathbb{R}^m \to \mathbb{R}^m$ as follows:

$$\mathfrak{f}(t,\mathfrak{y}) = \Big(\langle \mathcal{N}(t,y^h),\varphi_i\rangle_{V',V}\Big)_{1\le i\le m}, \quad (t,\mathfrak{y}) \in [0,T]\times\mathbb{R}^m,$$

where we set $y^h = \sum_{i=1}^m \mathfrak{y}_i \varphi_i \in V^h$ and $\mathfrak{y}_i$ denotes the i-th component of the vector $\mathfrak{y}$. Now, (2.1.26) leads to the following system (compare (2.1.16))

$$M\dot{\mathfrak{y}}^h(t) + A(t)\mathfrak{y}^h(t) = \mathfrak{f}(t, \mathfrak{y}^h(t)) \text{ f.a.a. } t \in [0,T], \quad M\mathfrak{y}^h(0) = \mathfrak{y}_\circ. \tag{2.1.27}$$

2. POD for Parametrized Elliptic Partial Differential Equations

In this section we concentrate on the POD method for parametrized elliptic PDEs. We explain briefly the numerical implementation, which is similar as described in Section 1.3.1. Let us also refer to [**12**].

2.1. Linear Elliptic Equations. Let $\mathcal{D} \subset \mathbb{R}^P$ be a bounded and closed subset. Suppose that for $\mu \in D$ the parameter dependent bilinear form $a(\mu;\cdot\,\cdot) : V \times V \to \mathbb{R}$ satisfies

$$\big|a(\mu;\varphi,\phi)\big| \le \beta \,\|\varphi\|_V \|\phi\|_V \qquad \text{for all } \varphi,\phi \in V \text{ and for } \mu \in \mathcal{D}, \tag{2.2.1a}$$

$$a(\mu;\varphi,\varphi) \ge \kappa \,\|\varphi\|_V^2 \qquad \text{for all } \varphi \in V \text{ and for } \mu \in \mathcal{D} \tag{2.2.1b}$$

for positive constants β, κ. Further, for $\mu \in \mathcal{D}$ let $f(\mu) \in V'$ be a parameter dependent right-hand side. For given parameter $\mu \in \mathcal{D}$ we consider the variational problem: find $y = y(\mu) \in V$ solving

$$a(\mu; y, \varphi) = \langle f(\mu), \varphi \rangle_{V',V} \quad \text{for all } \varphi \in V. \tag{2.2.2}$$

Example 2.2.1. For $\mu_a, \mu_b \in \mathbb{R}$ with $\mu_a < \mu_b$ we define the parameter subset $\mathcal{D} = [\mu_a, \mu_b]$. Then we define the parameter dependent bilinear form $a(\mu;\cdot\,,\cdot) : V \times V \to \mathbb{R}$ as

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

For any $\mu \in \mathcal{D}$ we infer from (B.1) that

$$|a(\mu;\varphi,\phi)| \le \big(1 + C_V^2 \max\{|\mu_a|,|\mu_b|\}\big) \,\|\varphi\|_V \|\phi\|_V \quad \text{for all } \varphi,\phi \in V,$$

i.e., the bilinear form $a(\mu;\cdot\,,\cdot)$ satisfies (2.2.1a) with $\beta = 1 + \max\{|\mu_a|,|\mu_b|\}C_V^2$. Further

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

If $\mu_a \ge 0$ holds, then (2.2.1b) is satisfied with $\kappa = 1$. In the case $\mu_a < 0$ we infer from (B.1) that

$$a(\mu;\varphi,\varphi) \ge \|\varphi\|_V^2 + \mu_a \,\|\varphi\|_H^2 \ge \big(1 + \mu_a C_V^2\big)\|\varphi\|_V^2 \quad \text{for all } \varphi \in V \text{ and } \mu \in \mathcal{D}.$$

Summarizing, (2.2.1b) holds if $\kappa = 1 + \min\{0, \mu_a C_V^2\} > 0$ is fulfilled. ◊

The following theorem ensures that (2.2.2) admits a unique solution.

Theorem 2.2.2. *Suppose that the parameter dependent bilinear form $a(\mu;\cdot\,,\cdot)$ satisfies (2.2.1) and $f(\mu) \in V'$ holds true for any $\mu \in \mathcal{D}$. Then, there exists a unique solution $y = y(\mu) \in V$ to (2.2.2) for every $\mu \in \mathcal{D}$. Moreover, we have*

$$\|y\|_V \le \frac{1}{\kappa} \,\|f(\mu)\|_{V'} \quad \textit{for every } \mu \in \mathcal{D}. \tag{2.2.3}$$

In particular, if the mapping $\mu \mapsto f(\mu) \in V'$ is in $L^2(\mathcal{D})$, $y \in L^2(\mathcal{D};V)$ holds.

PROOF. Since the bilinear form $a(\mu;\cdot,\cdot)$ is bounded and coercive on $V \times V$ for every parameter $\mu \in \mathcal{D}$, the existence of a unique solution to (2.2.2) follows directly from the Lax-Milgram lemma; see [**8**], for instance. Next we prove the a-priori estimate (2.2.3). For that purpose we take $\varphi = y \in V$ in (2.2.2). It follows that

$$\kappa \|y\|_V^2 \leq a(\mu; y, y) = \langle f(\mu), y \rangle_{V',V} \leq \|f(\mu)\|_{V'} \|y\|_V,$$

which gives the claim. □

Together with (2.2.2) we will consider a discretized variational problem, where we apply POD for the discretization of V. For that purpose let $y(\mu) \in V$ the associated solution to (2.2.2) for chosen parameter $\mu \in \mathcal{D}$. We suppose that $f \in L^2(\mathcal{D}; V')$ holds, so that $y \in L^2(\mathcal{D}; V) \hookrightarrow L^2(\mathcal{D}; H)$ by Theorem 2.2.2. Further, X denotes either the space V or the space H. We define the bounded linear operator $\mathcal{Y} : L^2(\mathcal{D}) \to X$ by

$$\mathcal{Y}\phi = \int_{\mathcal{D}} \phi(\mu) y(\mu) \, \mathrm{d}\mu \quad \text{for } \phi \in L^2(\mathcal{D}).$$

Its Hilbert space adjoint $\mathcal{Y}^\star : X \to L^2(\mathcal{D})$ is given by

$$\big(\mathcal{Y}^\star \psi\big)(\mu) = \langle \psi, y(\mu) \rangle_X \quad \text{for } \psi \in X \text{ and } \mu \in \mathcal{D}.$$

Furthermore, we find that the bounded, linear, symmetric and nonnegative operator $\mathcal{R} = \mathcal{Y}\mathcal{Y}^\star : X \to X$ has the form

$$\mathcal{R}\psi = \int_{\mathcal{D}} \langle \psi, y(\mu) \rangle_X y(\mu) \, \mathrm{d}\mu \quad \text{for } \psi \in X. \tag{2.2.4}$$

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

$$\big(\mathcal{K}\phi\big)(\bar{\mu}) = \int_{\mathcal{D}} \langle y(\mu), y(\bar{\mu}) \rangle_X \phi(\mu) \, \mathrm{d}\mu \quad \text{for } \phi \in L^2(\mathcal{D}). \tag{2.2.5}$$

Since the mapping $\mu \mapsto y(\mu) \in V$ is in $L^2(\mathcal{D})$, we conclude that

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

This implies that $\mathcal{K} = \mathcal{Y}^\star \mathcal{Y}$ is compact (see Exercise 2.3.1) and, therefore, $\mathcal{R} = \mathcal{Y}\mathcal{Y}^\star$ is compact as well. From Theorems A.17 and A.18 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 nonnegative real numbers so that

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

Furthermore,

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

REMARK 2.2.3 (Methods of snapshots). Analogous to Remark 1.4.4, we find that the bounded, linear, symmetric and nonnegative operator $\mathcal{K}$ (see (2.2.5) has the same eigenvalues $\{\lambda_i\}_{i \in \mathbb{N}}$ as the operator $\mathcal{R}$ and the eigenfunctions

$$\phi_i(t) = \frac{1}{\sqrt{\lambda_i}} \big(\mathcal{Y}^\star \psi_i\big)(\mu) = \frac{1}{\sqrt{\lambda_i}} \langle \psi_i, y(\mu) \rangle_V$$

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

In the following theorem we formulate properties of the eigenvalues and eigenfunctions of $\mathcal{R}$.

THEOREM 2.2.4. *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* (2.2.4). *Then, for every $\ell \in \mathbb{N}$ the first ℓ eigenfunctions $\psi_1, \dots, \psi_\ell \in X$ solve the minimization problem*

$$\min_{\tilde{\psi}_1,\dots,\tilde{\psi}_\ell \in X} \int_{\mathcal{D}} \left\| y(\mu) - \sum_{i=1}^{\ell} \langle y(\mu), \tilde{\psi}_i \rangle_X \tilde{\psi}_i \right\|_X^2 \mathrm{d}\mu \quad \text{s.t. } \langle \tilde{\psi}_j, \tilde{\psi}_i \rangle_X = \delta_{ij} \text{ for } 1 \le i,j \le \ell. \tag{2.2.6}$$

Moreover,

$$\int_{\mathcal{D}} \left\| y(\mu) - \sum_{i=1}^{\ell} \langle y(\mu), \psi_i \rangle_X \psi_i \right\|_X^2 \mathrm{d}\mu = \sum_{i=\ell+1}^{\infty} \lambda_i \quad \text{for any } \ell \in \mathbb{N}. \tag{2.2.7}$$

PROOF. The proof of the claim relies on the fact that the eigenvalue problem

$$\mathcal{R}\psi_i = \lambda_i \psi_i \quad \text{for } i = 1, \dots, \ell \tag{2.2.8}$$

is the first-order necessary optimality condition for (2.2.6). The proof follows by similar arguments as the proof of Theorem 2.1.5. □

We call a solution to (2.2.6) a *POD basis of rank ℓ*. Analogous to Corollary 1.2.1 we have:

$$\sum_{i=1}^{\ell} \lambda_i = \sum_{i=1}^{\ell} \int_{\mathcal{D}} \left| \langle y(\mu), \psi_i \rangle_X \right|^2 \mathrm{d}\mu \ge \sum_{i=1}^{\ell} \int_{\mathcal{D}} \left| \langle y(\mu), \chi_i \rangle_X \right|^2 \mathrm{d}\mu$$

for every $\ell \in \mathbb{N}$, where $\{\chi_i\}_{i\in\mathbb{N}}$ is an arbitrary orthonormal basis in X.

In applications the weak solution to (2.2.2) is not known for all parameters $\mu \in \mathcal{D}$, but only for a given grid in $\mathcal{D}$. For that purpose let $\{\mu_j\}_{j=1}^n$ be a grid in $\mathcal{D}$ and let $y_i = y(\mu_i)$, $1 \le i \le n$, denote the corresponding solutions to (2.2.2) for the grid points μ_i. Here, we only concentrate on the discretization of the parameter space $\mathcal{D}$. The finite element approximation can be carried analogous to Section 1.3.1. We define the snapshot set $\mathcal{V}^n = \operatorname{span}\{y_1, \dots, y_n\} \subset V$ and determine a POD basis of rank $\ell \le n$ for $\mathcal{V}^n$ by solving

$$\min_{\tilde{\psi}_1,\dots,\tilde{\psi}_\ell \in X} \sum_{j=1}^{n} \alpha_j \left\| y_j - \sum_{i=1}^{\ell} \langle y_j, \tilde{\psi}_i \rangle_X \tilde{\psi}_i \right\|_X^2 \quad \text{s.t. } \langle \tilde{\psi}_j, \tilde{\psi}_i \rangle_X = \delta_{ij} \text{ for } 1 \le i,j \le \ell \tag{2.2.9}$$

where the α_j's are nonnegative weights. The solution to (2.2.9) is given by the solution to the eigenvalue problem

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

with

$$\mathcal{R}^n \psi = \sum_{j=1}^{n} \alpha_j \langle y_j, \psi \rangle_X y_j \quad \text{for } \psi \in X.$$

In contrast to $\mathcal{R}$ introduced in (2.2.4) 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 $d^n \le n$, whereas, in general, the image space of

the operator $\mathcal{R}$ is infinite-dimensional. Since $\mathcal{R}^n$ is a linear, bounded, compact, nonnegative, selfadjoint 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

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

REMARK 2.2.5 (Snapshot POD [**20**]). Let us define the diagonal matrix $D = \operatorname{diag}(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$. We supply $\mathbb{R}^n$ with the weighted inner product

$$\langle u, v \rangle_D = \sum_{i=1}^{n} \alpha_i u_i v_i = u^\top D v \quad \text{for } u = (u_1, \ldots, u_n)^\top, v = (v_1, \ldots, v_n)^\top \in \mathbb{R}^n.$$

If the α_i's are quadrature weights corresponding to the parameter grid $\{\mu_i\}_{i=1}^n$ then the inner product $\langle \cdot, \cdot \rangle_D$ is a discrete version of the inner product in $L^2(\mathcal{D})$. We define the symmetric nonnegative matrix $\mathcal{K}^n \in \mathbb{R}^{n \times n}$ with the elements $\langle y_i, y_j \rangle_X$, $1 \leq i, j \leq n$, and consider the eigenvalue problem

$$\mathcal{K}^n \phi_i^n = \lambda_i^n \phi_i^n, \ 1 \leq i \leq \ell \quad \text{and} \quad \langle \phi_i^n, \phi_j^n \rangle_D = \delta_{ij}, \ 1 \leq i, j \leq \ell \leq d^n. \tag{2.2.10}$$

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

$$\psi_i = \frac{1}{\sqrt{\lambda_i^n}} \sum_{j=1}^{n} \alpha_j (\phi_i^n)_j y_j \quad \text{for } i = 1, \ldots, \ell, \tag{2.2.11}$$

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

2.2. Extension to Nonlinear Elliptic Problems. Let us turn to a certain nonlinear problem; see [**12**], for instance. Suppose that for any $\mu \in \mathcal{D}$ the mapping $\mathcal{N}(\mu; \cdot) : V \to V'$ is a nonlinear, locally Lipschitz-continuous mapping satisfying

$$\langle \mathcal{N}(\mu; \phi) - \mathcal{N}(\mu; \varphi), \phi - \varphi \rangle_{V',V} \geq 0 \quad \text{for all } \phi, \varphi \in V \text{ and for all } \mu \in \mathcal{D}, \tag{2.2.12}$$

i.e., $\mathcal{N}(\mu; \cdot)$ is monotone for any $\mu \in \mathcal{D}$. Instead of (2.2.2) we consider

$$a(\mu; y, \varphi) + \langle \mathcal{N}(\mu; y), \varphi \rangle_{V',V} = \langle f(\mu), \varphi \rangle_{V',V} \quad \text{for all } \varphi \in V. \tag{2.2.13}$$

EXAMPLE 2.2.6. Let us give an example for a semilinear problem satisfying (2.2.12). Suppose that $\Omega \subset \mathbb{R}^{\mathsf{d}}$, $\mathsf{d} \in \{1, 2, 3\}$, is a bounded and open set with Lipschitz-continuous boundary $\Gamma = \partial\Omega$. We consider

$$-\mu_1 \Delta y + \mu_2 y + \mu_3 y^3 = g \text{ in } \Omega \quad \text{and} \quad \mu_1 \frac{\partial y}{\partial n} + \mu_4 y = g_R \text{ on } \Gamma = \partial\Omega, \tag{2.2.14}$$

where $g \in L^2(\Omega)$, $g_R \in L^2(\Gamma)$ and

$$\mathcal{D} = \left\{ \mu = (\mu_1, \ldots, \mu_4) \in \mathbb{R}^4 \,\middle|\, \mu_a \leq \mu_i \leq \mu_b \text{ for } i = 1, \ldots, 4 \right\}$$

with $0 < \mu_a \leq \mu_b$. A *weak solution* to (2.2.14) satisfies $y \in V = H^1(\Omega)$ and

$$\int_\Omega \mu_1 \nabla y \cdot \nabla \varphi + \left(\mu_2 y + \mu_3 y^3\right) \varphi \, \mathrm{d}x + \int_\Gamma \mu_4 y \varphi \, \mathrm{d}s = \int_\Omega g \varphi \, \mathrm{d}x + \int_\Gamma g_R \varphi \, \mathrm{d}s \tag{2.2.15}$$

for all $\varphi \in V$. Next we express (2.2.15) in the form (2.2.13). For that purpose we utilize the parametrized bilinear form $a(\mu;\cdot,\cdot) : V \times V \to \mathbb{R}$ given by

$$a(\mu;\phi,\varphi) = \int_\Omega \mu_1 \nabla\phi \cdot \nabla\varphi + \mu_2 \phi\varphi \,\mathrm{d}x + \int_\Gamma \mu_4 \phi\varphi \,\mathrm{d}s \quad \text{for } \phi,\varphi \in V \text{ and } \mu \in \mathcal{D}.$$

Notice that this bilinear form satisfies (2.2.1). Moreover, let the parameter independent right-hand side be given as

$$\langle f, \varphi\rangle_{V',V} = \int_\Omega g\varphi \,\mathrm{d}x + \int_\Gamma g_R\varphi \,\mathrm{d}s \quad \text{for } \varphi \in V.$$

Finally, we define the nonlinearity

$$\langle \mathcal{N}(\mu;\phi), \varphi\rangle_{V',V} = \int_\Omega \mu_3 y^3\varphi \,\mathrm{d}x \quad \text{for } \phi,\varphi \in V \text{ and } \mu \in \mathcal{D}.$$

Then a weak solution to (2.2.14) satisfies the variational formulation (2.2.13). Recall that $\varphi \in V$ implies $\varphi \in L^6(\Omega)$. Consequently, $\varphi^3(\Omega) \in H = L^2(\Omega) \subset V'$. Let $\phi,\varphi \in V$ and $\chi = \phi - \varphi \in V$. From $\mu_3 \geq \mu_a > 0$ we infer that

$$\begin{aligned}\langle \mathcal{N}(\mu;\phi) - \mathcal{N}(\mu;\varphi), \chi\rangle_{V',V} &= \int_\Omega \mu_3\big(\phi^3 - \varphi^3\big)\chi \,\mathrm{d}x \\ &= \mu_3 \int_\Omega \Big(\int_0^1 3(\varphi + \tau\chi)^2 \chi \mathrm{d}\tau\Big)\chi \,\mathrm{d}x \\ &= \mu_3 \int_0^1 \int_\Omega (\varphi + \tau\chi)^2 \chi^2 d\tau \,\mathrm{d}x \geq 0\end{aligned}$$

holds true. Thus, (2.2.12) is satisfied. ◊

If a solution $y(\mu)$ to (2.2.13) is given then a POD basis can be computed as described above for the linear problem (2.2.2).

Remark 2.2.7. We can also combine the theory of Sections 1 and 2 by considering parametrized evolution problems. In this case the time variable t as well as the parameter are the sampling parameters for the POD method, i.e., we set $\mathcal{D}_T = [0,T] \times \mathcal{D}$ and apply the POD appoach to the set $\mathcal{D}_T$. ◊

3. Exercises

Exercise 2.3.1. Let X be a Hilbert space and $y \in L^2(0,T;X)$. Prove that the kernel $k(s,t) = \langle y(s), y(t)\rangle_X$ f.a.a. $s,t \in [0,T]$ belongs to $L^2((0,T)\times(0,T))$.

Exercise 2.3.2. Show that for $\ell = 1$ the solution ψ_1 to (2.1.13) solves (2.1.12). How can this result be extended to an arbitrary $\ell \leq \dim \mathcal{V}$?

Exercise 2.3.3. Prove that (2.1.16) has a unique solution $\vec{y} \in H^1(0,T;\mathbb{R}^m)$.

Exercise 2.3.4. For a diffusion coeffisient $c > 0$ we consider the linear heat equation

$$\begin{aligned} y_t(t,\boldsymbol{x}) &= c\Delta y(t,\boldsymbol{x}) && \text{for } (t,\boldsymbol{x}) \in Q = (0,T)\times\Omega, \\ y(t,\boldsymbol{x}) &= 0 && \text{for } (t,\boldsymbol{x}) \in \Sigma = (0,T\times\partial\Omega, \\ y(0,\boldsymbol{x}) &= y_\circ(\boldsymbol{x}) && \text{for } \boldsymbol{x} \in \Omega = (0,1)\times(0,T) \subset \mathbb{R}^2. \end{aligned} \tag{2.3.16}$$

We write $\boldsymbol{x} = (x_1, x_2)$ for a point in the spatial domain Ω. Derive a semi-discrete system of the form (2.1.16) by using a discretization with classical finite differences with equidistant mesh size $h = 1/(N+1)$ and with $m = N^2$. To solve (2.1.16) formulate

1) the explicit and implicit Euler method,
2) the trapezoidal (or Crank-Nolson) method

using the time grid $0 = t_1 < t_2 < \ldots < t_n = T$ with step sizes $\delta_j = t_j - t_{j-1}$ for $2 \leq j \leq n$.an equidistant time grid. Discuss whether these solutions methods are well-defined.

Exercise 2.3.5. Implement a code to solve (2.3.16) by the finite difference method following Exercise 2.3.4. For the time integration utilize the following methods:

1) the implicit Euler method (`IE`),
2) the Crank-Nicolson scheme (`CN`) and
3) the Rannacher smoothing method (`RS`), i.e., four half implicit Euler steps $\Delta/2$ followed by regular Crank-Nicolson steps.

For simplicity use an equidistant time grid. Structure your code as follows:
`main` ... main script file, where all parameters are set and the discrete solution is plotted.
`[A,h,X1,X2] = preparation(m)` ... Given the parameter m for the inner spatial grid points, this function returns the discretization of the Laplace operator, the spatial mesh size h, the discretization grids `X1` and `X2` for the x_1- and x_2 axes (includung the boudary points).
`[Y,t] = solve_heat_fdm(c,A,h,n,y0,method)` ... Solves the linear heat equation, where `c` is the diffusion coeffisient, `n` the number of time steps, `y0` the vector of the initial condition evaluated at the inner spatial grid points and `method` classifies the selected solver (`'IE'`, `'CN'`, `'RS'`). The returned values are a matrix $\texttt{Y} \in \mathbb{R}^{m \times n}$, which columns contain the discrete solution to (2.3.16) at the time instances t_j, $1 \leq j \leq n$, and the vector `t` of the corresponding time instances.
`YFDM = add_boundary(Y)` ... Adds the (zero) boundary values to the solution matrix `Y`.
To test your code choose $N = 100$, $n = 100$ and the following setting for c and $y_\circ$:

- $c = 0.01$ and $y_\circ(\boldsymbol{x}) = \sin(2\pi x_1)\sin(2\pi x_2)$;
- $c = 0.5$ and $y_\circ(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in \Omega_1 = (0, 0.25) \times (0.25, 0.75)$, $y_\circ(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \Omega \setminus \Omega_1$;
- $c = 0.01$ and $y_\circ(\boldsymbol{x}) = \texttt{1} * \texttt{rand(size(x1))} < \texttt{0.01}$.

How does the performance of the three methods differ? What do you observe?

Exercise 2.3.6. In this exercise we want to solve (2.1.20) utilizing snapshots computed from a truth approximation for (2.3.16). The α_j's are chosen as trapezoidal weights. Thus, we make use of the code implemented in Exercise 2.3.5. For the inner products we use the Euclidean inner product (`'E'`) and the discretized $L^2(\Omega)$ inner product (`'L2'`); see Exercise 1.5.7. Structure your code as follows:
`main` ... Main script file, where all parameters are set and the desired results are plotted.
`W = weight_matrix(m,wtype)` ... Computes the weighting matrix for the inner product, *i.e.*, $\langle u, v\rangle_W = u^\top W v$ for $u, v \in \mathbb{R}^m$. The input parameters are the total number of inner points $\texttt{m} = N^2$ and the weighting matrix type `wtype` (`'E'`, `'L2'`). The wighting matrix `W` for the inner grid points is the return value.
`[lambda,Psi,traceK] = pod_basis(Y,pod,W,ell)` ... Computes the POD basis by solving (2.1.20). The input variables are the matrix `Y` containing the snapshots (without the boundary points), the method `pod` for computing the POD basis (`'eig'`, `'svd'`), the weighting matrix`W` for the inner product and the number `ell`

of desired POD basis vectors. The output arguments are the eigenvalues `lambda` $\in \mathbb{R}^{\ell}$, the POD basis `Psi` $\in \mathbb{R}^{m\times\ell}$ and the trace `traceK` of the correlation matrix $\bar{Y}^{\top}\bar{Y} = D^{1/2}Y^{\top}\mathtt{W}YD^{1/2}$ with the diagonal matrix $D = \operatorname{diag}\{\alpha_1,\ldots,\alpha_n\} \in \mathbb{R}^{n\times n}$. Utilize the MATLAB routines `eigs` or `svds` in case of `pod = 'eig'` or `pod = 'svd'`, respectively

To compute the snapshots use the code implemented in Exercise 2.3.5 with the Rannacher smoothing scheme (`method = RS`). Further, we set $N = 100$ and $n = 100$. For the diffusion coefficient and the initial condition $y_\circ$ we choose

- $c = 0.01$ and $y_\circ(\boldsymbol{x}) = \sin(2\pi x_1)\sin(2\pi x_2)$;
- $c = 0.5$ and $y_\circ(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in \Omega_1 = (0, 0.25)\times(0.25, 0.75)$, $y_\circ(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \Omega \setminus \Omega_1$.

For the two settings plot the decay of the eigenvalues scaled by `traceK` in a semi-log scale. What do you observe? How do the two methods compare with respect to their performance? Note that for the choice `'eigs'` negative and complex eigenvalues can occur due to numerical issues. Hence only plot the absolute vale of the real part. Further plot the first four POD basis functions.

CHAPTER 3

Reduced-Order Models for Finite-Dimensional Dynamical Systems

In Chapter 1 we have introduced the POD basis of rank ℓ in $\mathbb{R}^m$. In particular in Section 4 of Chapter 1 we discussed its application to the case when the snapshots are given by the solution to an initial-value problem at certain time instances. In Section 1 we utilize the POD basis to compute a so-called *low-dimensional approximation* or a *reduced-order model* (ROM) for (1.4.1). If a solution to the reduced-order model is computed, the question arises whether we can estimate the error between the solution to (1.4.1) and the reduced-order solution. This is the issue of Section 2.

1. Reduced-Order Modelling

Suppose that we have determined a POD basis $\{\psi_j\}_{j=1}^{\ell}$ of rank $\ell \in \{1, \dots, m\}$ in $\mathbb{R}^m$ as described in Section 4 of Chapter 1. Then we make the ansatz

$$y^\ell(t) = \sum_{j=1}^{\ell} \underbrace{\langle y^\ell(t), \psi_j \rangle_W}_{=:\mathfrak{y}_j^\ell(t)} \psi_j \quad \text{for all } t \in [0,T], \tag{3.1.1}$$

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

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

holds, $y^\ell(t)$ is an approximation for $y(t)$ provided $\ell < d$. Inserting (3.1.1) into (1.4.1) yields

$$\sum_{j=1}^{\ell} \dot{\mathfrak{y}}_j^\ell(t)\psi_j = \sum_{j=1}^{\ell} \mathfrak{y}_j^\ell(t)\, A\psi_j + f(t, y^\ell(t)), \quad t \in (0,T], \tag{3.1.2a}$$

$$\sum_{j=1}^{\ell} \mathfrak{y}_j^\ell(0)\psi_j = y_\circ \tag{3.1.2b}$$

Note that (3.1.2) is an initial-value problem in $\mathbb{R}^m$ for $\ell \le m$ coefficient functions $\mathfrak{y}_j^\ell(t)$, $1 \le j \le \ell$ and $t \in [0,T]$, so that the coefficients are overdetermined. Therefore, we assume that (3.1.2) holds after projection on the ℓ dimensional subspace $V^\ell = \operatorname{span}\{\psi_j\}_{j=1}^{\ell}$. From (3.1.2a) and $\langle \psi_j, \psi_i \rangle_W = \delta_{ij}$ we infer that

$$\dot{\mathfrak{y}}_i^\ell(t) = \sum_{j=1}^{\ell} \mathfrak{y}_j^\ell(t)\, \langle A\psi_j, \psi_i \rangle_W + \langle f(t, y^\ell(t)), \psi_i \rangle_W \tag{3.1.3}$$

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

$$A^\ell = ((a_{ij})) \in \mathbb{R}^{\ell\times\ell} \quad \text{with} \quad a_{ij}^\ell = \langle A\psi_j, \psi_i\rangle_W,$$

the vector-valued mapping

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

and the nonlinearity $f^\ell = (f_1^\ell, \dots, f_\ell^\ell)^T : [0,T] \times \mathbb{R}^\ell \to \mathbb{R}^\ell$ with the components

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

Then, (3.1.3) can be expressed as

$$\dot{\mathfrak{y}}^\ell(t) = A^\ell\mathfrak{y}^\ell(t) + f^\ell(t, \mathfrak{y}^\ell(t)) \quad \text{for } t \in (0,T] \tag{3.1.4a}$$

From (3.1.2b) we derive

$$\mathfrak{y}^\ell(0) = \mathfrak{y}_\circ^\ell, \tag{3.1.4b}$$

where

$$\mathfrak{y}_\circ^\ell = \begin{pmatrix} \langle y_\circ, \psi_1\rangle_W \\ \vdots \\ \langle y_\circ, \psi_\ell\rangle_W \end{pmatrix} \in \mathbb{R}^\ell$$

holds. System (3.1.4) is called the *POD-Galerkin projection* for (1.4.1). In case of $\ell \ll m$ the ℓ-dimensional system (3.1.4) is a low-dimensional approximation for (1.4.1). Therefore, (3.1.4) is a reduced-order model for (1.4.1).

2. Error Analysis for the Reduced-Order Model

In this section we focus on error analysis for POD Galerkin approximations. 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 (1.4.1) and $\{\psi_i\}_{i=1}^\ell$ the POD basis of rank ℓ solving

$$\begin{aligned} \min_{\tilde\psi_1,\dots,\tilde\psi_\ell \in \mathbb{R}^m} & \int_0^T \left\| y(t) - \sum_{i=1}^{\ell} \langle y(t), \tilde\psi_i\rangle_W \, \tilde\psi_i \right\|_W^2 \mathrm{d}t \\ \text{s.t. } & \langle \tilde\psi_j, \tilde\psi_i\rangle_W = \delta_{ij}, \ 1 \le i,j \le \ell. \end{aligned} \tag{3.2.1}$$

The reduced-order model for (1.4.1) is given by (3.1.4). 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

$$V^\ell = \operatorname{span}\{\psi_1, \dots, \psi_\ell\} \subset \mathbb{R}^m$$

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

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

Then,

$$\begin{aligned}\mathcal{P}^\ell\big(\alpha\psi+\tilde\alpha\tilde\psi\big)&=\sum_{i=1}^{\ell}\langle\alpha\psi+\tilde\alpha\tilde\psi,\psi_i\rangle_W\,\psi_i=\sum_{i=1}^{\ell}\Big(\alpha\,\langle\psi,\psi_i\rangle_W+\tilde\alpha\,\langle\tilde\psi,\psi_i\rangle_W\Big)\,\psi_i\\&=\alpha\mathcal{P}^\ell\psi+\tilde\alpha\mathcal{P}^\ell\tilde\psi\end{aligned}$$

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

$$\begin{aligned}\|\mathcal{P}^\ell\|^2_{L(\mathbb{R}^m)}&=\sup_{\|\psi\|_W=1}\|\mathcal{P}^\ell\psi\|^2_W=\sup_{\|\psi\|_W=1}\sum_{i=1}^{\ell}\big|\langle\psi,\psi_i\rangle_W\big|^2\\&\le\sup_{\|\psi\|_W=1}\sum_{i=1}^{m}\big|\langle\psi,\psi_i\rangle_W\big|^2=\sup_{\|\psi\|_W=1}\|\psi\|^2_W=1,\end{aligned}\tag{3.2.2}$$

i.e., $\mathcal{P}^\ell$ is bounded and therefore continuous. From $\langle\psi_i,\psi_j\rangle_W=\delta_{ij}$, $1\le i,j\le\ell$ we infer that

$$\big(\mathcal{P}^\ell\big)^2\psi=\mathcal{P}^\ell\big(\mathcal{P}^\ell\psi\big)=\sum_{i=1}^{\ell}\Big\langle\sum_{j=1}^{\ell}\langle\psi,\psi_j\rangle_W\,\psi_j,\psi_i\Big\rangle_W\,\psi_i=\mathcal{P}^\ell\psi\quad\text{for }\psi\in\mathbb{R}^m,$$

i.e., $\mathcal{P}^\ell$ is a projection; see Definition A.8. It is easy to prove that $\mathcal{P}^\ell$ is also selfadjoint. Thus, $\mathcal{P}^\ell$ is an orthogonal projection; see Remark A.9 in the appendix. Notice that, (3.2.2) and $\|\mathcal{P}^\ell\psi\|_W=\|\psi\|_W$ for any $\psi\in V^\ell$ imply $\|\mathcal{P}^\ell\|_{L(\mathbb{R}^m)}=1$, which is well-known for any orthogonal projection.

Throughout we shall use the decomposition

$$y(t)-y^\ell(t)=y(t)-\mathcal{P}^\ell y(t)+\mathcal{P}^\ell y(t)-y^\ell(t)=\varrho^\ell(t)+\vartheta^\ell(t),\tag{3.2.3}$$

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

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

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

$$\int_0^T\|\varrho^\ell(t)\|^2_W\,\mathrm{d}t=\sum_{i=\ell+1}^{m}\lambda_i.\tag{3.2.4}$$

Next we estimate the term $\vartheta^\ell(t)$. Utilizing (1.4.1a) and (3.1.4) we obtain for every $\psi^\ell\in V^\ell$ and $t\in(0,T]$

$$\begin{aligned}\langle\dot\vartheta^\ell(t),\psi^\ell\rangle_W&=\langle\mathcal{P}^\ell\dot y(t)-\dot y(t),\psi^\ell\rangle_W+\langle\dot y(t)-\dot y^\ell(t),\psi^\ell\rangle_W\\&=\langle\mathcal{P}^\ell\dot y(t)-\dot y(t),\psi^\ell\rangle_W\\&\quad+\langle A(y(t)-y^\ell(t))+f(t,y(t))-f(t,y^\ell(t)),\psi^\ell\rangle_W\end{aligned}\tag{3.2.5}$$

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

$$\|A\|=\max_{\|\psi\|_W=1}\|A\psi\|_W$$

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

$$\frac12\frac{\mathrm{d}}{\mathrm{d}t}\|\vartheta^\ell(t)\|^2_W=\langle\dot\vartheta^\ell(t),\vartheta^\ell(t)\rangle_W\quad\text{for every }t\in(0,T].$$

holds. Then, we infer from (3.2.5)

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\vartheta^\ell(t)\right\|_W^2 &\leq \|A\|\big(\|\varrho^\ell(t)\|_W+\|\vartheta^\ell(t)\|_W\big)\|\vartheta^\ell(t)\|_W\\ &\quad+\|f(t,y(t))-f(t,y^\ell(t))\|_W\|\vartheta^\ell(t)\|_W\\ &\quad+\|\mathcal{P}^\ell\dot{y}(t)-\dot{y}(t)\|_W\|\vartheta^\ell(t)\|_W.\end{aligned}\tag{3.2.6}$$

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

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

Moreover, we have

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

for all $t\in(0,T)$. Consequently, (3.2.6) and (3.2.3) imply

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

Consequently,

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

Now we make use of the following lemma; see Exercise 2.1.

LEMMA 3.2.1 (Gronwall's lemma). *For $T>0$ let $u:[0,T]\to\mathbb{R}$ be a nonnegative, differentiable function satisfying*

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

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

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

In particular, if

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

holds, then $u = 0$ *in* $[0,T]$.

Using Lemma 3.2.1 and (3.2.4) we arrive at

$$\begin{aligned}\|\vartheta^\ell(t)\|_W^2 &\le c_1\bigg(\|\vartheta^\ell(0)\|_W^2 + \big(\|A\| + L_f\big)\int_0^t \|\varrho^\ell(s)\|_W^2\,\mathrm{d}s\bigg)\\ &\quad + c_1\sum_{i=\ell+1}^m \int_0^t \big|\langle \dot y(s),\psi_i\rangle_W\big|^2\,\mathrm{d}s \qquad (3.2.7)\\ &\le c_2\bigg(\|\vartheta^\ell(0)\|_W^2 + \sum_{i=\ell+1}^m\Big(\lambda_i + \int_0^T \big|\langle \dot y(t),\psi_i\rangle_W\big|^2\,\mathrm{d}t\Big)\bigg)\end{aligned}$$

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

THEOREM 3.2.2. *Let* $y \in C([0,T];\mathbb{R}^m)\cap C^1(0,T;\mathbb{R}^m)$ *be the unique solution to* (1.4.1), $\ell \in \{1,\dots,m\}$ *be fixed and* $\{\psi_i\}_{i=1}^\ell$ *a POD basis of rank* ℓ *solving* (3.2.1). *Let* y^ℓ *be the unique solution to the reduced-order model* (3.1.4). *Then*

$$\int_0^T \|y(t) - y^\ell(t)\|_W^2\,\mathrm{d}t \le C\sum_{i=\ell+1}^m\Big(\lambda_i + \int_0^T \big|\langle \dot y(t),\psi_i\rangle_W\big|^2\,\mathrm{d}t\Big)$$

for a constant $C > 0$.

PROOF. From (3.2.4), (3.2.7) and $\vartheta^\ell(0) = \mathcal{P}^\ell y_\circ - y^\ell(0) = 0$ we find

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

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

REMARK 3.2.3. The term

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

can not be estimated by the sum over the eigenvalues $\lambda_{\ell+1},\dots,\lambda_m$. If we replace (3.2.1) by

$$\begin{aligned}\min_{\tilde\psi_1,\dots,\tilde\psi_\ell\in\mathbb{R}^m} &\int_0^T \Big\|y(t) - \sum_{i=1}^\ell \langle y(t),\psi_i\rangle_W\,\psi_i\Big\|_W^2\\ &\qquad + \Big\|\dot y(t) - \sum_{i=1}^\ell \langle \dot y(t),\psi_i\rangle_W\,\psi_i\Big\|_W^2\,\mathrm{d}t \qquad (3.2.8)\\ \text{s.t. } &\langle \psi_i,\psi_j\rangle_W = \delta_{ij} \quad \text{for } 1 \le i,j \le \ell,\end{aligned}$$

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 (2.1.6) has to be replaced. It turns out that the POD basis $\{\psi_i\}_{i=1}^{\ell}$ is given by the eigenvalue problem

$$\tilde{\mathcal{R}}\tilde{\psi}_i = \tilde{\lambda}_i \tilde{\psi}_i \text{ for } 1 \le i \le m \quad \text{and} \quad \tilde{\lambda}_1 \ge \tilde{\lambda}_2 \ge \ldots \ge \tilde{\lambda}_m \ge 0 \tag{3.2.9}$$

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

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

for $\psi \in \mathbb{R}^m$; see Exercises 3.4.2 and 3.4.3. ◊

REMARK 3.2.4. Suppose that we build the matrix $Y \in \mathbb{R}^{m\times(2n)}$ using the column vectors $y_j \approx y(t_j)$, $1 \le j \le n$, and $y_j \approx \dot{y}(t_{j-m})$, $m+1 \le j \le 2m$. Then, the discrete variant $\tilde{\mathcal{R}}^n$ of the operator $\tilde{\mathcal{R}}$ introduced in Remark 3.2.3 is given by

$$\begin{aligned}
\tilde{\mathcal{R}}^n\psi &= \sum_{j=1}^{n} \alpha_j \langle y_j, \psi \rangle_W y_j + \alpha_j \langle y_{m+j}, \psi \rangle_W y_{m+j} \\
&= \sum_{j=1}^{n} \alpha_j \Big(\Big(\sum_{k=1}^{m}\sum_{\nu=1}^{m} Y_{kj} W_{k\nu} \psi_\nu \Big) Y_{\cdot,j} + \Big(\sum_{k=1}^{m}\sum_{\nu=1}^{m} Y_{k,m+j} W_{k\nu} \psi_\nu \Big) Y_{\cdot,m+j} \Big) \\
&= \sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{\nu=1}^{m} \Big(\Big(Y_{\cdot,j} D_{jj} Y_{jk}^\top + Y_{\cdot,m+j} D_{jj} Y_{m+j,k}^\top \Big) W_{k\nu} \psi_\nu \Big) \\
&= Y \underbrace{\begin{pmatrix} D & 0 \\ 0 & D \end{pmatrix}}_{=:\tilde{D}\in\mathbb{R}^{2n\times 2n}} Y^\top W \psi = Y \tilde{D} Y^\top W \psi
\end{aligned}$$

with the diagonal matrix $D = \operatorname{diag}(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^{n\times n}$ and nonnegative weights introduced in $(\hat{\mathbf{P}}_W^{n,\ell})$. Thus, we have $\tilde{\mathcal{R}}^n = Y\tilde{D}Y^\top W \in \mathbb{R}^{m\times m}$, which is of the same form as in (1.4.6). The discrete version to (3.2.9) is

$$Y\tilde{D}Y^\top W\psi_i = \lambda_i \psi_i \text{ for } 1 \le i \le m \quad \text{and} \quad \lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_m \ge 0 \tag{3.2.10}$$

Setting $\psi_i = W^{-1/2}\bar{\psi}_i$ in (3.2.10) and multiplying by $W^{1/2}$ from the left yield

$$W^{1/2} Y \tilde{D} Y^\top W^{1/2} \bar{\psi}_i = \lambda_i \bar{\psi}_i. \tag{3.2.11}$$

Let $\bar{Y} = W^{1/2} Y \tilde{D}^{1/2} \in \mathbb{R}^{m\times 2n}$. Using $W^\top = W$ as well as $\tilde{D}^\top = \tilde{D}$ we infer from (3.2.11) that the solution $\{\psi_i\}_{i=1}^{\ell}$ is given by the symmetric $m \times m$ eigenvalue problem

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

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

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

Thus, the POD basis of rank ℓ can also be computed by the methods of snapshots as follows: First solve the symmetric $2n \times 2n$ eigenvalue problem

$$\bar{Y}^\top \bar{Y} \bar{\phi}_i = \lambda_i \bar{\phi}_i, \ 1 \le i \le \ell \quad \text{and} \quad \langle \bar{v}_i, \bar{v}_j \rangle_{\mathbb{R}^{2n}} = \delta_{ij}, \ 1 \le i, j \le \ell.$$

Then we set (by SVD)

$$\psi_i = W^{-1/2}\bar{\psi}_i = \frac{1}{\sqrt{\lambda_i}} W^{-1/2}\bar{Y}\bar{\phi}_i = \frac{1}{\sqrt{\lambda_i}} Y\tilde{D}^{1/2}\bar{\phi}_i$$

for $1 \le i \le \ell$. ◊

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, nonequidistant meshes can be treated analogously [**15**]. We compute a POD basis $\{\psi_i^n\}_{i=1}^{\ell}$ 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 1. Thus,

$$y^\ell(t) = \sum_{i=1}^{\ell} \mathfrak{y}_i^\ell(t)\psi_i^n, \quad t \in [0,T],$$

solves the POD Galerkin projection of (1.4.1)

$$\langle \dot{y}^\ell(t), \psi_i^n\rangle_W = \langle Ay^\ell(t) + f(t, y^\ell(t)), \psi_i^n\rangle_W, \quad i = 1\ldots,\ell \text{ and } t \in (0,T], \tag{3.2.12a}$$

$$\langle y^\ell(0), \psi_i^n\rangle_W = \langle y_\circ, \psi_i^n\rangle_W, \qquad i = 1\ldots,\ell. \tag{3.2.12b}$$

To solve (3.2.12) we apply the implicit Euler method. By y_j^ℓ we denote an approximation for y^ℓ at the time t_j, $1 \le j \le n$. Then, the discrete system for the sequence $\{y_j^\ell\}_{j=1}^n$ in $V_n^\ell = \operatorname{span}\{\psi_1^n, \ldots, \psi_\ell^n\}$ looks like

$$\left\langle \frac{y_j^\ell - y_{j-1}^\ell}{\Delta t}, \psi_i^n \right\rangle_W = \langle Ay_j^\ell + f(t, y_j^\ell), \psi_i^n\rangle_W, \quad i = 1\ldots,\ell,\ 2 \le j \le n, \tag{3.2.13a}$$

$$\langle y_1^\ell, \psi_i^n\rangle_W = \langle y_\circ, \psi_i^n\rangle_W, \qquad i = 1\ldots,\ell. \tag{3.2.13b}$$

We are interested in estimating

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

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

$$\mathcal{P}_n^\ell = \sum_{i=1}^{\ell} \langle \psi, \psi_i^n\rangle_W \, \psi_i^n \quad \text{for } \psi \in \mathbb{R}^m. \tag{3.2.14}$$

It follows that $\mathcal{P}_n^\ell$ is linear and bounded (and therefore continuous). In particular, $\|\mathcal{P}_n^\ell\|_{L(\mathbb{R}^m)} = 1$; see Exercise 3.4.4.

We shall make use of the decomposition

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

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

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

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

$$\sum_{j=1}^{n} \alpha_j \left\|\varrho_j^\ell\right\|_W^2 = \sum_{i=\ell+1}^{m} \lambda_i^n. \tag{3.2.15}$$

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 (1.4.1a) and (3.2.13a)

$$\begin{aligned}
\langle \overline{\partial}\vartheta_j^\ell, \psi_i^n\rangle &= \left\langle \mathcal{P}_n^\ell\left(\frac{y(t_j)-y(t_{j-1})}{\Delta t}\right) - \frac{y_j^\ell - y_{j-1}^\ell}{\Delta t}, \psi_i^n \right\rangle_W \\
&= \langle \dot{y}(t_j) - (Ay_j^\ell + f(t_j, y_j^\ell))), \psi_i^n\rangle_W \\
&\quad + \left\langle \mathcal{P}_n^\ell\left(\frac{y(t_j)-y(t_{j-1})}{\Delta t}\right) - \dot{y}(t_j), \psi_i^n \right\rangle_W \\
&= \langle A(y(t_j) - y_j^\ell) + f(t_j, y(t_j)) - f(t_j, y_j^\ell), \psi_i^n\rangle_W \\
&\quad + \left\langle \mathcal{P}_n^\ell\left(\frac{y(t_j)-y(t_{j-1})}{\Delta t}\right) - \frac{y(t_j)-y(t_{j-1})}{\Delta t}, \psi_i^n \right\rangle_W \\
&\quad + \left\langle \frac{y(t_j)-y(t_{j-1})}{\Delta t} - \dot{y}(t_j), \psi_i^n \right\rangle_W \\
&= \langle A(y(t_j) - y_j^\ell) + f(t_j, y(t_j)) - f(t_j, y_j^\ell) + z_j^\ell + w_j^\ell, \psi_i^n\rangle_W
\end{aligned} \tag{3.2.16}$$

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

$$\langle \overline{\partial}\vartheta_j^\ell, \vartheta_j^\ell\rangle = \langle A(y(t_j) - y_j^\ell) + f(t_j, y(t_j)) - f(t_j, y_j^\ell) + z_j^\ell + w_j^\ell, \vartheta_j^\ell\rangle_W \tag{3.2.17}$$

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

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

for all $\psi, \tilde{\psi} \in \mathbb{R}^m$. Choosing $\psi = \vartheta_j^\ell$ and $\tilde{\psi} = \vartheta_{j-1}^\ell$ we infer from (3.2.17)

$$2\langle \overline{\partial}\vartheta_j^\ell, \vartheta_j^\ell\rangle = \frac{1}{\Delta t}\left(\|\vartheta_j^\ell\|_W^2 - \|\vartheta_{j-1}^\ell\|_W^2 + \|\vartheta_j^\ell - \vartheta_{j-1}^\ell\|_W^2 \right). \tag{3.2.18}$$

Inserting (3.2.18) into (3.2.17) and using the Cauchy-Schwarz inequality we obtain

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

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

$$\|f(t_j, y(t_j)) - f(t_j, y_j^\ell)\|_W \le L_f \|y(t_j) - y_j^\ell\|_W \quad \text{for } j = 2, \dots, n.$$

Hence, by Young's inequality we find

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

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

$$0 < \Delta t \leq \frac{1}{2c_2} \tag{3.2.19}$$

holds. With (3.2.19) 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,

$$\frac{1}{1 - c_2\Delta t} = \frac{1 - c_2\Delta t + c_2\Delta t}{1 - c_2\Delta t} = 1 + \frac{c_2\Delta t}{1 - c_2\Delta t}. \leq 1 + 2c_2\Delta t \tag{3.2.20}$$

Using (3.2.20) 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, $\vartheta_0^\ell = \mathcal{P}_n^\ell y_\circ - Y_1 = 0$ and $0 \leq j\Delta t \leq T$ for $j = 0, \ldots, n-1$. Summarizing

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

where the constant $c_3 = e^{2c_2 T} \max\{7/3, 2c_1, 8\}$ is independent of ℓ and $\{t_j\}_{j=1}^n$. From $\sum_{k=1}^n \alpha_k = T$ and (3.2.15) we infer

$$\begin{aligned}
\sum_{j=1}^{n} \alpha_j \|\vartheta_j^\ell\|_W^2 &\leq c_3 T \bigg(\sum_{j=1}^{n} \alpha_j \left(\|\mathcal{P}_n^\ell \dot{y}(t_j) - \dot{y}(t_j)\|_W^2 + \|\varrho_j^\ell\|_W^2 \right) \\
&\qquad + (\Delta t)^2 \|\ddot{y}\|_{L^2(0,T;\mathbb{R}^m)}^2 \bigg) \\
&\leq c_4 \bigg(\sum_{i=\ell+1}^{m} \left(\lambda_i^n + \sum_{j=1}^{n} \alpha_j \left| \langle \dot{y}(t_j), \psi_i^n \rangle_W \right|^2 \right) + (\Delta t)^2 \bigg)
\end{aligned} \tag{3.2.21}$$

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

THEOREM 3.2.5. *Let $y \in C([0,T];\mathbb{R}^m) \cap C^1(0,T;\mathbb{R}^m)$ be the unique solution to (1.4.1) satisfying $\ddot{y} \in L^2(0,T;\mathbb{R}^m)$ and $\ell \in \{1, \ldots, m\}$ be fixed. Suppose that $\{\psi_i^n\}_{i=1}^\ell$ is a POD basis of rank ℓ solving $(\hat{\mathbf{P}}_W^{n,\ell})$. Assume that (3.2.13) possesses a unique solution $\{y_j^\ell\}_{j=1}^n$. Then there exists a constant $C > 0$ such that*

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

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

PROOF. The claim follows directly from (3.2.15), (3.2.21), and

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

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

REMARK 3.2.6. Compared to the estimate in Theorem 3.2.2 we observe the term

$$\sum_{j=1}^{n} \alpha_j \left| \langle \dot{y}(t_j), \psi_i^n \rangle_W \right|^2 \tag{3.2.22}$$

instead of the term

$$\int_0^T \left| \langle \dot{y}(t), \psi_i \rangle_W \right|^2 \mathrm{d}t. \tag{3.2.23}$$

Note that (3.2.22) is the trapezoidal approximation of (3.2.23). Further, the error $O((\Delta t)^2)$ appears in the estimate of Theorem 3.2.5 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 1.4.5 that there exists a number $\bar{n} \in \mathbb{N}$ satisfying

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

for $n \geq \bar{n}$ provided $\sum_{i=\ell+1}^m \lambda_i \neq 0$ and $\int_0^T \left| \langle \dot{y}(t), \psi_i \rangle_W \right|^2 \mathrm{d}t \neq 0$ hold. Thus, we infer from Theorems 3.2.2 and 3.2.5 the following result.

THEOREM 3.2.7. *Let all hypothesis of Theorems* 1.4.5, 3.2.2 *and* 3.2.5 *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*

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

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

3. Empirical Interpolation Method for Nonlinear Problem

The ROM introduced in (3.1.4) is a nonlinear system. Hence the problem with the POD Galerkin approach is the complexity of the evaluation of the nonlinearity. To illustrate this we have a look at the nonlinearity f^ℓ in (3.1.4a). Setting $\Psi = [\psi_1 \,|\, \dots \,|\, \psi_\ell] \in \mathbb{R}^{m\times\ell}$ we can write

$$f^\ell(t, \mathfrak{y}^\ell(t)) = \Psi^\top W f(t, \Psi\mathfrak{y}^\ell(t)).$$

This can be interpreted in the way that the variable $\mathfrak{y}^\ell(t) \in \mathbb{R}^\ell$ is first expanded to a vector $U\mathrm{y}^\ell(t)$ of dimension m, then the nonlinearity $f^\ell(t, \Psi\mathfrak{y}^\ell(t))$ is evaluated and at last the result is reduced back to the low dimension ℓ of the reduced-order model. This is computationally expensive. Further this means that our reduced-order model is not independent of the full dimension m. Note that when applying a Newton method to the system (3.1.4) the Jacobian of the nonlinearity is also needed. For instance, we have

$$\frac{\partial f^\ell}{\partial \mathfrak{y}}(t, \Psi\mathfrak{y}^\ell(t)) = \Psi^\top g^\ell(t, \Psi\mathfrak{y}^\ell(t))\Psi \quad \text{for } t \in [0, T],$$

where

$$g^\ell(t, \Psi\mathfrak{y}^\ell)(t)) = \Big(\big\langle f_y(t, \Psi\mathfrak{y}^\ell(t))\psi_j, \psi_i\big\rangle_{V',V}\Big)_{1\le i,j\le m}.$$

Again the same problem can be observed. Note that here the computation expenses are larger since the Jacobians are of dimension $m \times m$. Hence not only a vector is transformed but a matrix of full dimension. To avoid this computational expensive evaluation the empirical interpolation method (EIM) was introduced [**1**]. This method is often used in combination with the reduced basis approach [**9**]. The second approach we will investigate here is the discrete empirical interpolation method (DEIM) as introduced in [**3, 5, 4**]. While the EIM implementation is based on a greedy algorithm the DEIM implementation is based on a POD approach combined with a greedy algorithm. We will now discuss both methods. We define

$$b(t) = f(t, \Psi\mathfrak{y}^\ell(t)) \in \mathbb{R}^m \quad \text{for } t \in [0, T].$$

Now, $b(t)$ is approximated by a Galerkin ansatz utilizing $\wp$ linearly independent functions $\phi_1, \dots, \phi_\wp \in \mathbb{R}^m$, i.e.

$$b(t) \approx \sum_{k=1}^{\wp} \phi_k c_k(t) = \Phi c(t) \tag{3.3.1}$$

with $c(t) = [c_1(t), \dots, c_\wp(t)]^\top \in \mathbb{R}^\wp$ and $\Phi = [\phi_1 \,\big|\, \dots \,\big|\, \phi_\wp] \in \mathbb{R}^{m\times\wp}$. Hence we can write the approximation of $f^\ell(t, \cdot)$ as

$$f^\ell(t, \mathfrak{y}^\ell(t)) = \Psi^\top W f(t, \Psi\mathfrak{y}^\ell(t)) = \Psi^\top W b(t) \approx \Psi^\top W \Phi c(t).$$

The question arising is how to compute the matrix Φ and the vector $c(t)$. Let $\vec{\imath} \in \mathbb{R}^\wp$ be an index vector and $B \in \mathbb{R}^{m\times\wp}$ a given matrix. Then by $B_{\vec{\imath}}$ we denote the submatrix consisting of the rows of B corresponding to the indices in $\vec{\imath}$. Obviously, if we choose $\wp$ indices then the overdetermined system $b(t) = \Phi c(t)$ can be solved by choosing $\wp$ rows of $b(t)$ and Φ. Here it is assumed that the submatrix $\Phi_{\vec{\imath}} \in \mathbb{R}^{\wp\times\wp}$ is invertible.

Assuming we have computed Φ and $\vec{\imath}$ by an algorithm. Then we proceed as follows. For simplicity we introduce here the matrix $P = (e_{\vec{\imath}_1} \,|\, \dots \,|\, e_{\vec{\imath}_\wp}) \in \mathbb{R}^{m\times\wp}$, where $e_{\vec{\imath}_i} = (0, \dots, 0, 1, 0, \dots, 0)^\top \in \mathbb{R}^m$ is a vector with all zeros and at the $\vec{\imath}_i$-th

row a one. Note that $\Phi_{\vec{\imath}} = P^\top \Phi$ holds. To evaluate the approximate nonlinearity we need $c(t)$. Since we know Φ and the index vector $\vec{\imath}$ we can compute

$$c(t) = \left(P^\top \Phi\right)^{-1} P^\top b(t) = \left(P^\top \Phi\right)^{-1} P^\top f(t, \Psi \mathfrak{y}^\ell(t)) \quad \text{for } t \in [0, T].$$

Suppose that the matrix P can be moved into the nonlinearity. Then, we obtain

$$P^\top f(t, \Psi \mathfrak{y}^\ell(t)) = \left(f(t, \Psi \mathfrak{y}^\ell(t))\right)_{\vec{\imath}} = f(t, P^\top \Psi \mathfrak{y}^\ell(t)).$$

An extension for general nonlinearities is shown in [**5**]. Let us now have a look at the computational expenses. The matrices $P^\top \Psi \in \mathbb{R}^{\wp \times \ell}$, $(P^\top \Phi)^{-1} \in \mathbb{R}^{\wp \times \wp}$ and $\Psi^\top W \Phi \in \mathbb{R}^{\ell \times \wp}$ can be precomputed. All the precomputed quantities are independent of the full dimension m. Additionally, during the iterations the nonlinearity only has to be evaluated at the interpolation points, i.e. only at $\wp$ points. This allows the reduced-order model to be completely independent of the full dimension. Note that the used method is an interpolation and therefore is exact at the interpolation points. For the Jacobian the approach is similar.

Let us now turn to the EIM and DEIM algorithms. When (1.4.1) is solved the nonlinearity $f(t, y(t))$ is evaluated for each time step. If these evaluations are stored the procedure to determine Φ and the index vector $\vec{\imath}$ does not involve any further evaluations of the nonlinearity. We denote by $F \in \mathbb{R}^{m \times n}$ the matrix with columns $f(t_i, y(t_i)) \in \mathbb{R}^m$ for $i = 1, \ldots, n$. Next let us have a look at the two algorithms of interest and let us present some numerical results. In the algorithms $\|\cdot\|_\infty$ stands for the maximum norm in $\mathbb{R}^m$ and the operation 'arg max' returns the index, where the maximum entry occurs. In Algorithm 6 we state the EIM using a greedy algorithm. Here the basis ϕ^i, $i = 1, \ldots, \wp$, is chosen from the

Algorithm 6 (The empirical interpolation method (EIM))

Require: $\wp$ and matrix $F = [f(t_1, y(t_1)) \,|\, \ldots \,|\, f(t_n, y(t_n))] \in \mathbb{R}^{m \times n}$;
1: $k \leftarrow \arg\max_{j=1,\ldots,n} \|f(t_j, y(t_j))\|_\infty$;
2: $\xi \leftarrow f(t_k, y(t_k))$;
3: idx $\leftarrow \arg\max_{j=1,\ldots,m} |\xi_j|$;
4: $\phi_1 \leftarrow \xi / \xi_{\{\text{idx}\}}$;
5: $\Phi = [\phi^1]$ and $\vec{\imath} =$ idx;
6: **for** $i = 2$ to ℓ^{EI} **do**
7: Solve $\Phi_{\{\wp^{EI}\}} c_j = f(t_j, y(t_j))_{\{\wp^{EI}\}}$ for $j = 1, \ldots, n$;
8: $k \leftarrow \arg\max_{j=1,\ldots,n} \|f(t_j, y(t_j)) - \Phi c_j\|_\infty$;
9: $\xi \leftarrow f(t_k, y(t_k))$;
10: idx$\leftarrow \arg\max_{j=1,\ldots,m} |(\xi - \Phi c_k)_{\{j\}}|$;
11: $\phi^i \leftarrow (\xi - \Phi c_k) / (\xi - \Phi c_k)_{\{\text{idx}\}}$;
12: $\Phi \leftarrow [\Phi, \phi^i]$ and $\vec{\imath} \leftarrow [\vec{\imath}, \text{idx}]$;
13: **end for**
14: **return** Φ and $\vec{\imath}$

provided snapshots of $f(t, y(t))$ by scaling and shifting. The obtained basis is not orthonormal. The advantage of this method is that the submatrix $\Phi_{\vec{\imath}}$ is an upper triangular matrix. Hence solving for $c(t)$ is computationally cheap. The drawback of this method is that the computation of the basis is more expensive than the DEIM algorithm presented in Algorithm 7. The DEIM algorithm on the other hand generates the basis using the POD approach. Here the previously introduced

Algorithm 7 (The discrete empirical interpolation method (DEIM))

Require: $\wp$ and matrix $F = [f(t_1, y(t_1)) \,|\, \dots \,|\, f(t_n, y(t_n))] \in \mathbb{R}^{m\times n}$;
1: Compute POD basis $\Phi = [\phi_1, \dots, \phi_\wp]$ for F;
2: idx$\leftarrow \arg\max_{j=1,\dots,m} |(\phi_1)_{\{j\}}|$;
3: $U = [\phi_1]$ and $\vec{\imath} =$ idx;
4: **for** $i = 2$ to $\wp$ **do**
5: $\quad u \leftarrow \phi_i$;
6: $\quad$ Solve $U_{\vec{\imath}}c = u_{\vec{\imath}}$;
7: $\quad r \leftarrow u - Uc$;
8: $\quad$ idx$\leftarrow \arg\max_{j=1,\dots,m} |(r)_{\{j\}}|$;
9: $\quad U \leftarrow [U, u]$ and $\vec{\imath} \leftarrow [\vec{\imath}, \text{idx}]$;
10: **end for**
11: **return** Φ and $\vec{\imath}$

POD approach is applied to the snapshots of the nonlinearity $b(t) = f(t, y(t))$ to compute Φ. The matrix $\Phi_{\vec{\imath}}$ obtained by the DEIM method has no special structure. Hence evaluating the nonlinearity using DEIM is more expensive compared to EIM. The computational cost can be reduced by precomputing a LU decomposition of $\Phi_{\vec{\imath}}$. Then the evaluation of the nonlinearity using DEIM involves two solves compared to one solve for the EIM. Further when comparing the two algorithms it can be seen that the computation for the EIM basis is more expensive compared to the DEIM basis. This can be seen when comparing line 7 in Algorithm 6 and line 6 in Algorithm 7. In each iteration of Algorithm 6 one has to solve n linear systems compared to one linear system in Algorithm 7. The selection for the interpolation points in both algorithms is similar and is based on a greedy algorithm. The idea is to successively select spatial points to limit the growth of an error bound. The indices are constructed inductively from the input data. For more details we refer the reader to [**1, 3**].

4. Exercises

Exercise 3.4.1. Prove Gronwall's lemma; see Lemma 3.2.1.

Exercise 3.4.2. Prove that the first-order necessary optimality condition for (3.2.8) is given by $\tilde{\mathcal{R}}\tilde{u}_i = \tilde{\lambda}_i\tilde{u}_i$, $1 \le i \le \ell$.

Exercise 3.4.3. Show that $\tilde{\mathcal{R}}$ is linear, bounded, self-adjoint and nonnegative provided $y \in H^1(0, T; \mathbb{R}^m)$, i.e.,

$$\int_0^T \|y(t)\|_W^2 + \|\dot{y}(t)\|_W^2 \,\mathrm{d}t < \infty$$

holds.

Exercise 3.4.4. Show that the operator $\mathcal{P}_n^\ell$ defined in (3.2.14) is linear, bounded and satisfies $\|\mathcal{P}_n^\ell\|_{L(\mathbb{R}^m)} = 1$.

CHAPTER 4

Balanced Truncation Method

1. The linear-quadratic control problem

In this section we introduce the optimal state-feedback and the linear-quadratic regulator (LQR) problem. Utilizing dynamic programming necessary optimality conditions are derived. It turns out that for the LQR problem the state-feedback solution can be determined by solving a differential matrix Riccati equation. The presented theory is taken from the book [**7**].

1.1. The linear-quadratic regulator (LQR) problem. The goal is to find a state-feedback control law of the form

$$u(t) = -Kx(t) \quad \text{for } t \in [0,T]$$

with $u : [0,T] \to \mathbb{R}^{m_u}$, $x : [0,T] \to \mathbb{R}^{m_x}$, $K \in \mathbb{R}^{m_u \times m_x}$ so that u minimizes the quadratic cost functional

$$J(x,u) = \int_0^T x(t)^T Q x(t) + u(t)^T R u(t)\,\mathrm{d}t + x(T)^T M x(T), \tag{4.1.1a}$$

where the state x and the control u are related by the linear initial value problem

$$\dot{x}(t) = Ax(t) + Bu(t) \text{ for } t \in (0,T] \quad \text{and} \quad x(0) = x_0. \tag{4.1.1b}$$

In (4.1.1a) the matrices Q, $M \in \mathbb{R}^{m_x \times m_x}$ are symmetric, positive semi-definite, $R \in \mathbb{R}^{m_u \times m_u}$ is symmetric, positive definite and in (4.1.1b) we have $A \in \mathbb{R}^{m_x \times m_x}$, $B \in \mathbb{R}^{m_x \times m_u}$ and $x_0 \in \mathbb{R}^{m_x}$. The final time T is fixed, but the final state $x(T)$ is free. Thus, we aim to track the state to the state $\bar{x} = 0$ as good as possible. The terms $x(t)^T Q x(t)$ and $x(T)^T M x(T)$ are measures for the control accuracy and the term $u(t)^T R u(t)$ measures the control effort. Problem (4.1.1) is called the *linear-quadratic regulator problem* (*LQR problem*).

1.2. The Hamilton-Jacobi-Bellman equation. In this section we derive first-order necessary optimality conditions for the LQR problem. Since generalizing the problem to a non-linear problem does not cause more difficulties in the deviation, we consider the problem to find a state-control feedback control law

$$u(t) = \Phi(x(t),t), \quad t \in [0,T],$$

such that the cost-functional

$$J_t(x,u) = \int_t^T L(x(s),u(s),s)\,\mathrm{d}s + g(x(T)) \tag{4.1.2a}$$

is minimized subject to the non-linear system dynamics

$$\dot{x}(s) = F(x(s),u(s),s) \text{ for } s \in (0,T] \quad \text{and} \quad x(t) = x_t. \tag{4.1.2b}$$

We suppose that the functions $L : \mathbb{R}^{m_x} \times \mathbb{R}^{m_u} \times [0,T] \to [0,\infty)$ and $g : \mathbb{R}^{m_x} \to [0,\infty)$ satisfy

$$L(0,0,s) = 0 \text{ for } s \in [0,T] \quad \text{and} \quad g(0) = 0$$

Moreover, let $F : \mathbb{R}^{m_x} \times \mathbb{R}^{m_u} \times [0,T] \to \mathbb{R}^{m_x}$ be continuous and locally Lipschitz-continuous with respect to the variable x. Moreover, $x_t \in \mathbb{R}^{m_x}$ holds. To derive optimality conditions we use the so-called *Bellman principle* (or *dynamic programming principle*). The essential assumption is that the system can be characterized by its state $x(t)$ at the time $t \in [0,T]$ which completely summarizes the effect of all $u(s)$ for $0 \le s \le t$. The dynamic programming principle was first proposed by Bellman [**2**].

THEOREM 4.1.1 (Bellman principle). *Let $t \in [0,T]$. If $u^*(s)$ is optimal for $s \in [t,T]$ and x^* is the associated optimal state, starting at the state $x_t \in \mathbb{R}^{m_x}$, then $u^*(s)$ is also optimal over the subinterval $[t+\Delta t, T]$ for any $\Delta t \in [0, T-t]$ starting at $x_{t+\Delta t} = x^*(t+\Delta t)$.*

PROOF. We show Theorem 4.1.1 by contradiction. Suppose that there exists a control u^{**} so that

$$\begin{aligned} \int_{t+\Delta t}^{T} L(x^{**}(s), u^{**}(s), s)\,\mathrm{d}s + g(x^{**}(T)) \\ < \int_{t+\Delta t}^{T} L(x^*(s), u^*(s), s)\,\mathrm{d}s + g(x^*(T)), \end{aligned} \tag{4.1.3}$$

where

$$\dot{x}^*(s) = F(x^*(s), u^*(s), s) \quad \text{and} \quad \dot{x}^{**}(s) = F(x^{**}(s), u^{**}(s), s)$$

hold for $s \in [t+\Delta t, T]$. We define the control

$$u(s) = \begin{cases} u^*(s) & \text{if } s \in [t, t+\Delta t], \\ u^{**}(s) & \text{if } s \in (t+\Delta t, T]. \end{cases} \tag{4.1.4}$$

By $x(s)$ we denote the state satisfying $\dot{x}(s) = F(x(s), u(s), s)$ for $s \in [t,T]$ and $x(t) = x_t$. Then we derive from (4.1.3) and (4.1.4) that

(4.1.5)

$$\begin{aligned} &\int_t^T L(x(s), u(s), s)\,\mathrm{d}s + g(x(T)) \\ &= \int_t^{t+\Delta t} L(x^*(s), u^*(s), s)\,\mathrm{d}s + \int_{t+\Delta t}^{T} L(x^{**}(s), u^{**}(s), s)\,\mathrm{d}s + g(x^{**}(T)) \\ &< \int_t^{t+\Delta t} L(x^*(s), u^*(s), s)\,\mathrm{d}s + \int_{t+\Delta t}^{T} L(x^*(s), u^*(s), s)\,\mathrm{d}s + g(x^*(T)) \\ &= \int_t^T L(x^*(s), u^*(s), s)\,\mathrm{d}s + g(x^*(T)). \end{aligned}$$

Recall that $u^*(s)$ is optimal for $s \in [t,T]$ by assumption. From (4.1.5) it follows that the control u given by (4.1.4) yields a smaller value of the cost functional. This is a contradiction. □

Next we derive the Hamilton-Jacobi-Bellman equation for (4.1.2). Let $V^* : \mathbb{R}^{m_x} \times [0,T] \to \mathbb{R}$ denote the minimal value function given by

(4.1.6)
$$V^*(x_t,t) = \min_{u:[t,T]\to\mathbb{R}^{m_u}} \Big\{ J_t(x,u) \,\Big|\, \dot{x}(s) = F(x(s),u(s),s),\ s \in (t,T] \text{ and } x(t) = x_t \Big\}$$

for $(x_t,t) \in \mathbb{R}^{m_x} \times [0,T]$, where

$$J_t(x,u) = \int_t^T L(x(s),u(s),s)\,\mathrm{d}s + g(x(T)).$$

From the linearity of the integral and (4.1.6) we conclude

(4.1.7)
$$\begin{aligned} V^*(x_t,t) = \min_{u:[t,t+\Delta t]\to\mathbb{R}^{m_u}} \Big\{ \int_t^{t+\Delta t} L(x(s),u(s),s)\,\mathrm{d}s + V^*(x(t+\Delta t),t+\Delta t) \,\Big|\, \\ \dot{x}(s) = F(x(s),u(s),s),\ s \in (t,t+\Delta t] \text{ and } x(t) = x_t \Big\} \end{aligned}$$

for $(x_t,t) \in \mathbb{R}^{m_x} \times [0,T-\Delta t]$, where we have used the Bellman principle. Thus, by using the Bellman principle the problem of finding an optimal control over the interval $[t,T]$ has been reduced to the problem of finding an optimal control over the interval $[t,t+\Delta t]$.

Now we replace the integral in (4.1.7) by $L(x(t),u(t),t)\Delta t$, perform a Taylor approximation for $V^*(x(t+\Delta t),t+\Delta t)$ about the point $(x_t,t) = (x(t),t)$ and approximate $x(t+\Delta t) - x(t)$ by $F(x(t),u(t),t)\Delta t$. Then we find

$$\begin{aligned} V^*(x_t,t) &= \min_{u_t\in\mathbb{R}^{m_u}} \Big\{ L(x_t,u_t,t)\Delta t + V^*(x_t,t) + \frac{\partial V^*}{\partial t}(x_t,t)\,\Delta t \\ &\qquad + \nabla V^*(x_t,t)^T F(x_t,u_t,t)\Delta t + o(\Delta t) \Big\} \\ &= V^*(x_t,t) + \frac{\partial V^*}{\partial t}(x_t,t)\Delta t \\ &\quad + \Delta t \min_{u_t\in\mathbb{R}^{m_u}} \Big\{ L(x_t,u_t,t) + \nabla V^*(x_t,t)^T F(x_t,u_t,t) + \frac{o(\Delta t)}{\Delta t} \Big\} \end{aligned}$$

for any $\Delta t > 0$. Thus,

$$-\frac{\partial V^*}{\partial t}(x_t,t) = \min_{u_t\in\mathbb{R}^{m_u}} \Big\{ L(x_t,u_t,t) + \nabla V^*(x_t,t)^T F(x_t,u_t,t) + \frac{o(\Delta t)}{\Delta t} \Big\}.$$

Taking the limit $\Delta t \to 0$ and using $V^*(x_t,T) = g(x_t)$ we obtain

$$-\frac{\partial V^*}{\partial t}(x_t,t) = \min_{u_t\in\mathbb{R}^{m_u}} \big\{ L(x_t,u_t,t) + \nabla V^*(x_t,t)^T F(x_t,u_t,t) \big\} \tag{4.1.8a}$$

for all $(x_t,t) \in \mathbb{R}^{m_x} \times [0,T)$ and

$$V^*(x_t,T) = g(x_t) \tag{4.1.8b}$$

for all $x_t \in \mathbb{R}^{m_x}$.

To solve (4.1.8) we proceed in two steps. First we compute a solution u_t to

$$u^*(t) = \underset{u_t \in \mathbb{R}^{m_u}}{\operatorname{argmin}} \left\{ L(x_t, u_t, t) + \nabla V^*(x_t, t)^T F(x_t, u_t, t) \right\}$$

and set

$$\Psi(\nabla V^*(x_t, t), x_t, t) = u^*(t), \tag{4.1.9}$$

which gives us a control law. Then we insert (4.1.9) into (4.1.8a) and solve

$$\begin{aligned} -\frac{\partial V^*}{\partial t}(x_t, t) = {} & L(x_t, \Psi(\nabla V^*(x_t, t), x_t, t), t) \\ & + \nabla V^*(x_t, t)^T F(x_t, \Psi(\nabla V^*(x_t, t), x_t, t), t) \end{aligned}$$

for all $(x_t, t) \in \mathbb{R}^{m_x} \times [0, T)$. Finally, we can compute the gradient $\nabla V^*(x_t, t)$ and deduce the state-feedback law

$$u^*(t) = \Phi(x_t, t) = \Psi(\nabla V^*(x_t, t), x_t, t) \quad \text{for all } (x_t, t) \in \mathbb{R}^{m_x} \times [0, T).$$

REMARK 4.1.2.
1) In general, it is not possible to solve (4.1.8) analytically. However, for the LQR problem we can derive an explicit solution for the state-feedback law.
2) Note that the Hamilton-Jacobi-Bellman equation are only necessary optimality conditions. ◊

1.3. The state-feedback law for the LQR problem. For the LQR problem we have

$$L(x, u, t) = x^T Q x + u^T R u, \quad g(x) = x^T M x, \quad F(x, u, t) = Ax + Bu$$

for $(x, u, t) \in \mathbb{R}^{m_x} \times \mathbb{R}^{m_u} \times [0, T]$. For brevity, we focus on the situation, where the matrices A, B, Q, R are time-invariant. However, most of the presented theory also holds for the time-varying case.

First we minimize

$$x^T Q x + u^T R u + \nabla V^*(x, t)^T \big(Ax + Bu\big)$$

with respect to u. First-order necessary optimality conditions are given by

$$u^T R \tilde{u} + \tilde{u}^T R u + \nabla V^*(x, t)^T B \tilde{u} = 0 \quad \text{for all } \tilde{u} \in \mathbb{R}^{m_u}.$$

By assumption, R is symmetric and positive definite. Then we find

$$\big(2Ru + B^T \nabla V^*(x, t)\big)^T \tilde{u} = 0 \quad \text{for all } \tilde{u} \in \mathbb{R}^{m_u}$$

and

$$u^* = -\frac{1}{2} R^{-1} B^T \nabla V^*(x, t). \tag{4.1.10}$$

For the minimal value function V^* we make the quadratic ansatz

$$V^*(x, t) = x^T P(t) x, \quad P(t) \in \mathbb{R}^{m_x \times m_x} \text{ symmetric.} \tag{4.1.11}$$

Then, we have $\nabla V^*(x, t) = 2P(t)x$ so that

$$u^* = -R^{-1} B^T P(t) x.$$

Note that

$$\begin{aligned}\frac{\partial V^*}{\partial t}(x_t,t) &= x_t^T\dot{P}(t)x_t,\\ L(x_t,-R^{-1}B^TP(t)x_t,t) &= x_t^TQx_t + x_t^TP(t)BR^{-1}B^TP(t)x_t\\ &= x_t^T\big(Q+P(t)BR^{-1}B^TP(t)\big)x_t,\\ F(x_t,-R^{-1}B^TP(t)x_t,t) &= Ax_t - BR^{-1}B^TP(t)x_t = \big(A-BR^{-1}B^TP(t)\big)x_t,\\ \nabla V^*(x_t,t) &= 2P(t)x_t.\end{aligned}$$

Consequently,

$$\begin{aligned}-x_t^T\dot{P}(t)x_t &= -\frac{\partial V^*}{\partial t}(x_t,t)\\ &= x_t^T\big(Q+P(t)BR^{-1}B^TP(t)\big)x_t + \big(2P(t)x_t\big)^T\big(A-BR^{-1}B^TP(t)\big)x_t\end{aligned}$$

for all $x_t\in\mathbb{R}^{m_x}$, which yields

$$\begin{aligned}&-x_t^T\dot{P}(t)x_t\\ &\quad= x_t^T\big(Q+P(t)BR^{-1}B^TP(t)+2P(t)A-2P(t)BR^{-1}B^TP(t)\big)x_t\\ &\quad= x_t^T\big(2P(t)A+Q-P(t)BR^{-1}B^TP(t)\big)x_t\end{aligned}$$

for all $x_t\in\mathbb{R}^{m_x}$. From $P(t)=P(t)^T$ we deduce that

$$2x_t^TP(t)Ax_t = x_t^TP(t)Ax_t + x_t^TA^TP(t)x_t = x_t^T\big(A^TP(t)+P(t)A\big)x_t.$$

Using $V^*(x_t,T)=x_t^TP(T)x_t$ and (4.1.8b) we get

(4.1.12a)
$$-x_t^T\dot{P}(t)x_t = x_t^T\big(A^TP(t)+P(t)A+Q-P(t)BR^{-1}B^TP(t)\big)x_t,\quad t\in[0,T)$$

(4.1.12b)
$$x_t^TP(T)x_t = x_t^TMx_t.$$

Since (4.1.12) holds for all $x_t\in\mathbb{R}^{m_x}$ we obtain the following *matrix Riccati equation*

(4.1.13a) $$-\dot{P}(t) = A^TP(t)+P(t)A+Q-P(t)BR^{-1}B^TP(t),\quad t\in[0,T)$$

(4.1.13b) $$P(T)=M.$$

Finally, the optimal state-feedback is given by

$$u^*(t) = -K(t)x(t)\quad\text{and}\quad K(t)=R^{-1}B^TP(t).$$

Example 4.1.3. Let us consider the problem

$$\min\int_0^T |x(t)|^2+|u(t)|^2\,\mathrm{d}t\quad\text{s.t.}\quad\dot{x}(t)=u(t)\text{ for }t\in(0,T].$$

Choosing $m_x=m_u=1$, $A=M=0$ and $B=Q=R=1$ the matrix Riccati equation has the form

$$-\dot{P}(t) = 1-P(t)^2\text{ for }t\in[0,T)\quad\text{and}\quad P(T)=0.$$

This scalar ordinary differential equation can be solved by separation of variables. Its solution is

$$P(t)=\frac{1-e^{-2(T-t)}}{1+e^{-2(T-t)}}$$

with the optimal control $u^*(t)=-P(t)x(t)$. ◊

2. Balanced truncation

Let us consider the linear time-invariant system

$$\dot{x}(t) = Ax(t) + Bu(t) \text{ for } t \in (0,\infty) \quad \text{and} \quad x(0) = x_0, \tag{4.2.14a}$$

$$y(t) = Cx(t) \qquad \text{for } t \in [0,\infty) \tag{4.2.14b}$$

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 (4.2.14) through the Laplace transform.

Definition 4.2.4. *Let $f(t)$ be a time-varying vector. Then its* Laplace transform *is defined by*

$$\mathcal{L}[f](s) = \int_0^\infty e^{-st} f(t)\,\mathrm{d}t \quad \textit{for } s \in \mathbb{R}. \tag{4.2.15}$$

The Laplace transform is defined for those values of s, for which (4.2.15) *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 (4.2.14b). 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 (4.2.14a) 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,

$$\mathcal{L}[y](s) = C\mathcal{L}[x](s) = C(sI - A)^{-1}x(0) + C(sI - A)^{-1}B\mathcal{L}[u](s). \tag{4.2.16}$$

For $x(0) = 0$ the expression (4.2.16) reduces to

$$\mathcal{L}[y](s) = G(s)\mathcal{L}[u](s) \tag{4.2.17}$$

where

$$G(s) = C(sI - A)^{-1}B \tag{4.2.18}$$

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*.

DEFINITION 4.2.5. *The dynamical system* (4.2.14a) *or the pair* (A, B) *are called* controllable *if for any* $x_0 \in \mathbb{R}^{m_x}$ *and final state* $x_T \in \mathbb{R}^{m_x}$ *there exists a (piecewise continuous) input* u *such that the solution to* (4.2.14a) *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 [**24**].

THEOREM 4.2.6. *The following claims are equivalent:*

1) (A, B) *are controllable.*
2) *The controllability gramian*

$$W_c(t) = \int_0^t e^{sA} B B^T e^{sA^T} \, \mathrm{d}s$$

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 rank.

DEFINITION 4.2.7. 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* (4.2.14a) *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 [**24**], is a consequence of Theorem 4.2.6.

THEOREM 4.2.8. *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.,* $\Re e\lambda < 0$.

Let us now consider the dual notions of observability.

DEFINITION 4.2.9. *The dynamical system* (4.2.14) *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 [**24**].

THEOREM 4.2.10. *The following claims are equivalent:*

1) (A, C) *is observable.*
2) *The observability gramian*

$$W_o(t) = \int_0^t e^{sA^T} C^T C e^{sA} \, \mathrm{d}s$$

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 rank.

We set

$$W_c = \int_0^\infty e^{sA} B B^T e^{sA^T} \,\mathrm{d}s \quad \text{and} \quad W_o = \int_0^\infty e^{sA^T} C^T C e^{sA} \,\mathrm{d}s.$$

It can be proved that W_c and W_o can be determined numerically by solving the *Lyapunov equations*

$$AW_c + W_c A^T + BB^T = 0 \in \mathbb{R}^{n_x \times n_x}, \tag{4.2.19a}$$

$$A^T W_o + W_o A + C^T C = 0 \in \mathbb{R}^{n_x \times n_x}. \tag{4.2.19b}$$

The controllability gramian is a measure to what degree each state is excited by an input. Suppose that x_1, $x_2 \in \mathbb{R}^{n_x}$ are two states with $\|x_1\|_{\mathbb{R}^{n_x}} = \|x_2\|_{\mathbb{R}^{n_x}}$. If $x_1^T W_c x_1 > x_2^T W_c x_2$ holds, then we say that the state x_1 is more controllable than x_2. This means, it takes a smaller input to drive the system from x_0 to x_1 than to x_2. It can be proved that the gramian W_c is positive definite if and only if all states are reachable with some input u. On the other hand, the observability gramian W_o is a measure to what degree each state excites future outputs y. Let x_0 be an initial state. If $u = 0$ holds, we have

$$\begin{aligned} \|y\|^2_{L^2(0,\infty;\mathbb{R}^{m_y})} &= \int_0^\infty y(s)^T y(s) \,\mathrm{d}s = \int_0^\infty x(s)^T C^T C x(s) \,\mathrm{d}s \\ &= \int_0^\infty x_0^T e^{sA^T} C^T C e^{sA} x_0 \,\mathrm{d}s = x_0^T W_o x_0. \end{aligned}$$

We say that the state x_1 is *more observable* than another state x_2 if the corresponding output $y_1 = Cx_1$ yields a larger value of the L^2-norm than for $y_2 = Cx_2$

The gramians depend on the coordinates. Suppose that

$$x = \mathcal{T} z \tag{4.2.20}$$

where $\mathcal{T} \in \mathbb{R}^{n_x \times n_x}$ is a regular matrix. Then we obtain instead of (4.2.14) the system

$$\dot{z}(t) = \tilde{A} z(t) + \tilde{B} u(t) \text{ for } t \in (0,\infty) \quad \text{and} \quad z(0) = z_0, \tag{4.2.21a}$$

$$y(t) = \tilde{C} z(t) \qquad \text{for } t \in [0,\infty) \tag{4.2.21b}$$

with

$$\tilde{A} = \mathcal{T}^{-1} A \mathcal{T}, \quad \tilde{B} = \mathcal{T}^{-1} B, \quad \tilde{C} = C\mathcal{T}, \quad z_0 = \mathcal{T}^{-1} x_0.$$

Let W_c solve (4.2.19a). The controllability gramian $\tilde{W}_c$ for (4.2.21) satisfies

$$\tilde{A}\tilde{W}_c + \tilde{W}_c \tilde{A}^T + \tilde{B}\tilde{B}^T = 0$$

i.e.,

$$\mathcal{T}^{-1} A \mathcal{T} \tilde{W}_c + \tilde{W}_c \mathcal{T}^T A^T \mathcal{T}^{-T} + \mathcal{T}^{-1} B B^T \mathcal{T}^{-T} = 0. \tag{4.2.22}$$

Multiplying (4.2.22) by $\mathcal{T}$ from the left and by $\mathcal{T}^T$ from the right yields

$$A\mathcal{T}\tilde{W}_c\mathcal{T}^T + \mathcal{T}\tilde{W}_c\mathcal{T}^T A^T + BB^T = 0. \tag{4.2.23}$$

From (4.2.19a) and (4.2.23) we infer that $W_c = \mathcal{T}\tilde{W}_c\mathcal{T}^T$ holds. Thus, the coordinate transformation (4.2.20) implies that the controllability gramian W_c is transformed as

$$W_c \mapsto \tilde{W}_c = \mathcal{T}^{-1} W_c \mathcal{T}^{-T}.$$

Now we suppose that W_o solves (4.2.19b). The observability gramian $\tilde{W}_o$ for (4.2.21) satisfies

$$\tilde{A}^T\tilde{W}_o + \tilde{W}_o\tilde{A} + \tilde{C}^T\tilde{C} = 0$$

i.e.,

$$\mathcal{T}^T A^T \mathcal{T}^{-T}\tilde{W}_o + \tilde{W}_o\mathcal{T}^{-1}A\mathcal{T} + \mathcal{T}^T C^T C\mathcal{T} = 0. \tag{4.2.24}$$

Multiplying (4.2.22) by $\mathcal{T}^{-T}$ from the left and by $\mathcal{T}^{-1}$ from the right yields

$$A^T\mathcal{T}^{-T}\tilde{W}_o\mathcal{T}^{-1} + \mathcal{T}^{-T}\tilde{W}_o\mathcal{T}^{-1}A + C^T C = 0. \tag{4.2.25}$$

From (4.2.19b) and (4.2.25) we infer that $W_o = \mathcal{T}^{-T}\tilde{W}_o\mathcal{T}^{-1}$ holds. Thus, the coordinate transformation (4.2.20) implies that the observability gramian W_o is transformed as

$$W_o \mapsto \tilde{W}_o = \mathcal{T}^T W_o \mathcal{T}.$$

The goal is to find a transformation $\mathcal{T}$ such that

$$\mathcal{T}^{-1}W_c\mathcal{T}^{-T} = \mathcal{T}^T W_o\mathcal{T} = \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_{m_x}). \tag{4.2.26}$$

The elements $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{m_x}$ are called *Hankel singular values* of the system. They are independent of the coordinate system. It can be shown that a regular matrix $\mathcal{T}$ which satisfies (4.2.26) exists if the system is controllable and observable, i.e., the matrices W_c and W_o are positive definite. The coordinate transformation $\mathcal{T}$ is said to be a *balancing transformation*. Computing appropriately scaled eigenvalues of the product W_cW_o, the matrix $\mathcal{T}$ can be determined. In the balanced coordinates, the states which are least influenced by the input u also have least influence on the output y. In *balanced truncation* the least controllable and observable states having little effect on the input-output performance are truncated.

Instead of (4.2.21) we only consider the system for the first $\ell \in \{1, \ldots, m_x\}$ components of z:

$$\dot{z}_\ell(t) = \tilde{A}_\ell z_\ell(t) + \tilde{B}_\ell u(t) \text{ for } t \in (0,\infty) \quad \text{and} \quad z_\ell(0) = z_{0\ell}, \tag{4.2.27a}$$

$$y_\ell(t) = \tilde{C}_\ell z_\ell(t) \qquad \text{for } t \in [0,\infty), \tag{4.2.27b}$$

where

$$\tilde{A} = \left(\begin{array}{c|c} \tilde{A}_\ell & * \\ \hline * & * \end{array}\right), \quad \tilde{B} = \begin{pmatrix} \tilde{B}_\ell \\ \hline * \end{pmatrix}, \quad \tilde{C} = \left(\begin{array}{c|c} \tilde{C}_\ell & * \end{array}\right), \quad z_{0\ell} = \begin{pmatrix} \tilde{z}_{0\ell} \\ \hline * \end{pmatrix},$$

and $\tilde{A}_\ell \in \mathbb{R}^{\ell\times\ell}$, $\tilde{B}_\ell \in \mathbb{R}^{\ell\times m_u}$, $\tilde{C}_\ell \in \mathbb{R}^{m_y\times\ell}$ and $z_{0\ell} \in \mathbb{R}^\ell$.

One big advantage of balanced truncation is that a-priori error bounds are known. These bounds are formulated for the transfer function. Suppose that $G(s) = C(sI - A)^{-1}B \in \mathbb{R}^{m_y\times m_u}$ is the transfer function of the system (4.2.14) and $G_\ell(s) = C_\ell(sI - A_\ell)^{-1}B_\ell \in \mathbb{R}^{m_y\times m_u}$ is the transfer function of the reduced system (4.2.27). Then we have

$$\|G - G_\ell\| = \max\left\{ \|(G - G_\ell)u\|_{L^2(0,\infty;\mathbb{R}^{m_y})} \,:\, \|u\|_{L^2(0,\infty;\mathbb{R}^{m_u})} = 1 \right\} > \sigma_{\ell+1}$$

and

$$\|G - G_\ell\| < 2 \sum_{i=\ell+1}^{m_x} \sigma_i.$$

3. Exercises

Let us consider the one-dimensional heat equation

$$\theta_t(t,x) = \theta_{xx}(t,x) + u(t)\chi(x) \qquad \text{for all } (t,x) \in Q = (0,T) \times \Omega, \tag{4.3.28a}$$

$$\theta_x(t,0) = \theta_x(t,1) = 0 \qquad \text{for all } t \in (0,T), \tag{4.3.28b}$$

$$\theta(0,x) = \theta_0(x) \qquad \text{for all } x \in \Omega = (0,1) \subset \mathbb{R}, \tag{4.3.28c}$$

where $\theta = \theta(t,x)$ is the temperature, $u = u(t)$ the control input, $\chi = \chi(x)$ a given control shape function and $\theta_0 = \theta_0(x)$ a given initial condition.

Exercise 5.3.1. Apply a classical finite difference approximation for the spatial variable x (compare Example 1.4.1) and derive the finite-dimensional initial value problem for the finite difference approximations.

Exercise 5.3.2. Utilizing the trapezoidal rule deduce a discretization for the quadratic cost functional

$$J(\theta,u) = \frac{1}{2}\int_\Omega |\theta(T,x) - \theta_T(x)|^2 \,\mathrm{d}x + \frac{\kappa}{2}\int_0^T |u(t)|^2 \,\mathrm{d}t,$$

where $\theta_T = \theta_T(x)$ is a given desired terminal state and $\kappa > 0$ denotes a fixed regularization parameter.

Exercise 5.3.3. Formulate the matrix Riccati equation for the discretized quadratic cost functional — see Exercise 5.3.2 — and the discretized heat equation — see Exercise 5.3.1.

Exercise 5.3.4. What is the matrix Riccati equation in the case if we apply a POD Galerkin approximation instead of a finite difference discretization? How can we solve the matrix Riccati equation numerically?

CHAPTER 5

The Appendix

A. Linear and Compact Operators

Let $\mathcal{X}$ and $\mathcal{Y}$ denote two real normed linear spaces with norms $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$.

DEFINITION A.1. *A* bounded linear operator $\mathcal{A}:\mathcal{X}\to\mathcal{Y}$ *satisfies the following two conditions*

1) $\mathcal{A}(\alpha_1 x_1+\alpha_2 x_2)=\alpha_1\mathcal{A}x_1+\alpha_2\mathcal{A}x_2$ *for all* $\alpha_1,\alpha_2\in\mathbb{R}$ *and* $x_1,x_2\in\mathcal{X}$*;*
2) *there exists a constant* $C_{\mathcal{A}}>0$ *such that* $\|\mathcal{A}x\|_{\mathcal{Y}}\le C_{\mathcal{A}}\|x\|_{\mathcal{X}}$ *for all* $x\in\mathcal{X}$.

The space of all bounded and linear operators from $\mathcal{X}$ *to* $\mathcal{Y}$ *is denoted by* $L(\mathcal{X},\mathcal{Y})$. *We shortly write* $L(\mathcal{X})$ *for* $L(\mathcal{X},\mathcal{X})$.

The following proposition is proved in [**19**, p. 70].

PROPOSITION A.2. *The space* $L(\mathcal{X},\mathcal{Y})$ *equipped with the norm*

$$\|\mathcal{A}\|_{L(\mathcal{X},\mathcal{Y})}=\sup_{\|x\|_{\mathcal{X}}=1}\|\mathcal{A}x\|_{\mathcal{Y}}\quad \textit{for } \mathcal{A}\in L(\mathcal{X},\mathcal{Y})$$

is a normed linear space. Furthermore, if $\mathcal{Y}$ *is even a Banach space then* $L(\mathcal{X},\mathcal{Y})$ *is a Banach space.*

REMARK A.3. The smallest constant $C_{\mathcal{A}}$ in Definition A.1-b) is given by the norm $\|\mathcal{A}\|_{L(\mathcal{X},\mathcal{Y})}$. ◊

DEFINITION A.4. *Let* $\mathcal{X}$ *and* $\mathcal{Y}$ *be two Banach spaces and* $A\in L(\mathcal{X},\mathcal{Y})$. *The* Banach space adjoint $\mathcal{A}':\mathcal{Y}'\to\mathcal{X}'$ *is defined by*

$$\langle\mathcal{A}'f,x\rangle_{\mathcal{X}',\mathcal{X}}=\langle f,\mathcal{A}x\rangle_{\mathcal{Y}',\mathcal{Y}}\quad \textit{for all } (f,x)\in\mathcal{Y}'\times\mathcal{X},$$

where $\langle\cdot\,,\cdot\rangle_{\mathcal{X}',\mathcal{X}}$ *stands for the dual pairing of* $\mathcal{X}'$ *and* $\mathcal{X}$.

It is proved in [**19**, p. 186] that $\|\mathcal{A}\|_{L(\mathcal{X},\mathcal{Y})}=\|\mathcal{A}'\|_{L(\mathcal{Y}',\mathcal{X}')}$ for $\mathcal{A}\in L(\mathcal{X},\mathcal{Y})$. Let $\mathcal{A}\in L(\mathcal{H})$ holds and $\mathcal{X}$ be a real Hilbert space, then we can introduce the Riesz isomorhism $\mathcal{J}_{\mathcal{X}}:\mathcal{X}\to\mathcal{X}'$ as follows: for given $x_1\in\mathcal{X}$ the element $\mathcal{J}_{\mathcal{X}}x_1$ satisfies

$$\langle\mathcal{J}_{\mathcal{X}}x_1,x_2\rangle_{\mathcal{X}',\mathcal{X}}=\langle x_1,x_2\rangle_{\mathcal{X}}\quad \text{for all } x_2\in\mathcal{X}.$$

By the Riesz theorem $\mathcal{J}_{\mathcal{X}}$ is well-defined. Moreover, $\|\mathcal{J}_{\mathcal{X}}\|_{L(\mathcal{X},\mathcal{X}')}=1$. For more details we refer the reader to [**19**, p. 43].

DEFINITION A.5. *Let* $\mathcal{X}$, $\mathcal{Y}$ *be two real Hilbert spaces and* $\mathcal{A}\in L(\mathcal{X},\mathcal{Y})$. *Then the* Hilbert space adjoint $\mathcal{A}^\star:\mathcal{Y}\to\mathcal{X}$ *is defined by* $\mathcal{A}^\star=\mathcal{J}_{\mathcal{X}}^{-1}\mathcal{A}'\mathcal{J}_{\mathcal{Y}}$.

REMARK A.6. Let $\mathcal{X}$, $\mathcal{Y}$ be two real Hilbert space and $\mathcal{A}\in L(\mathcal{X},\mathcal{Y})$. From $\mathcal{A}'\in L(\mathcal{Y}',\mathcal{X}')$ and $\|\mathcal{J}_{\mathcal{X}}\|_{L(\mathcal{X},\mathcal{X}')}=\|\mathcal{J}_{\mathcal{Y}}\|_{L(\mathcal{Y},\mathcal{Y}')}=1$ we infer that $\mathcal{A}^\star\in L(\mathcal{Y},\mathcal{Y})$. In

particular, $\|\mathcal{A}^\star\|_{L(\mathcal{Y},\mathcal{X})} = \|\mathcal{A}\|_{L(\mathcal{X},\mathcal{Y})}$. Further, we have

$$\begin{aligned}\langle \mathcal{A}^\star y, x\rangle_{\mathcal{X}} &= \langle \mathcal{J}_{\mathcal{X}}^{-1}\mathcal{A}'\mathcal{J}_{\mathcal{Y}} y, x\rangle_{\mathcal{X}} = \langle \mathcal{A}'\mathcal{J}_{\mathcal{Y}} y, x\rangle_{\mathcal{X}',\mathcal{X}} = \langle \mathcal{J}_{\mathcal{Y}} y, \mathcal{A}x\rangle_{\mathcal{Y}',\mathcal{Y}} \\ &= \langle y, \mathcal{A}x\rangle_{\mathcal{Y}}\end{aligned}$$

for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. ◊

The following theorem is proved in [**19**, pp. 186-187].

THEOREM A.7. *Let $\mathcal{X}$ be a real Hilbert space and $\mathcal{A}$, $\mathcal{B} \in L(\mathcal{X})$. Then, $(\mathcal{A}\mathcal{B})^\star = \mathcal{B}^\star\mathcal{A}$ and $(\mathcal{A}^\star)^\star = \mathcal{A}$.*

DEFINITION A.8. *Suppose that $\mathcal{X}$ is a real Hilbert space and $\mathcal{A} \in L(\mathcal{X})$. Then, $\mathcal{A}$ is called* selfadjoint *if $\mathcal{A} = \mathcal{A}^\star$ holds true. If $\mathcal{A}^2 = \mathcal{A}$ is valid, $\mathcal{A}$ is called a* projection. *If a projection $\mathcal{A}$ is selfadjoint, then $\mathcal{A}$ is an* orthogonal projection.

REMARK A.9. Suppose that $\mathcal{X}$ is a real Hilbert space and $\mathcal{A} \in L(\mathcal{X})$ is an orthogonal projection. Then, we have $\mathcal{A}^\star\mathcal{A} = \mathcal{A}^2 = \mathcal{A}$. Hence, it follows that for an arbitrary $x \in \mathcal{X}$

$$\langle \mathcal{A}x, x - \mathcal{A}x\rangle_{\mathcal{X}} = \langle \mathcal{A}x, x\rangle_{\mathcal{X}} - \langle \mathcal{A}x, \mathcal{A}x\rangle_{\mathcal{X}} = \langle \mathcal{A}x, x\rangle_{\mathcal{X}} - \langle \mathcal{A}^\star\mathcal{A}x, x\rangle_{\mathcal{X}} = 0.$$

Thus, the elements $\mathcal{A}x$ and $x - \mathcal{A}x$ are orthogonal in $\mathcal{X}$. ◊

DEFINITION A.10. *Let $\mathcal{X}$ be a Banach space and $\mathcal{A}$ belong to $L(\mathcal{X})$. A complex number λ is in the* resolvent set $\rho(\mathcal{A})$ *of $\mathcal{A}$ if $\lambda\mathcal{I} - \mathcal{A}$ is a bijection with bounded inverse $(\lambda\mathcal{I} - \mathcal{A})^{-1}$. The operator $\mathcal{R}_\lambda = (\lambda\mathcal{I} - \mathcal{A})^{-1}$ is called the* resolvent *of $\mathcal{A}$ at $\lambda \in \rho(\mathcal{A})$. If $\lambda \notin \rho(\mathcal{A})$, then λ is an element of the* spectrum $\sigma(\mathcal{A})$ *of $\mathcal{A}$. Let $x \in X \setminus \{0\}$ and $\lambda \in \mathbb{C}$ satisfying $\mathcal{A}x = \lambda x$. Then, λ is called an* eigenvalue *of $\mathcal{A}$ and x is an associated* eigenvector *of $\mathcal{A}$. The set of all eigenvalues is said to be the* point spectrum *of $\mathcal{A}$.*

The following theorem is taken from [**19**, p. 192].

THEOREM A.11. *Let $\mathcal{X}$ be a Banach space and $\mathcal{A} \in L(\mathcal{X})$. Then, $\sigma(\mathcal{A}) = \sigma(\mathcal{A}')$ holds. Moreover, for any $\lambda \in \rho(\mathcal{A})$ we have $\mathcal{R}_\lambda(\mathcal{A}') = \mathcal{R}_\lambda(\mathcal{A})'$. If $\mathcal{X}$ is a real Hilbert space, then*

$$\sigma(\mathcal{A}^\star) = \left\{\lambda \in \mathbb{C} \,\middle|\, \bar{\lambda} \in \sigma(\mathcal{A})\right\}$$

and $\mathcal{R}_\lambda(\mathcal{A}^\star) = \mathcal{R}_\lambda(\mathcal{A})^\star$, where $\bar{\lambda}$ denotes the complex conjugate of λ.

DEFINITION A.12. *Let $\mathcal{X}$ be a Hilbert space. Then, $\mathcal{A} \in L(\mathcal{X})$ is called a* positve operator *if $\langle Ax, x\rangle_{\mathcal{X}} \geq 0$ holds for all $x \in \mathcal{X}$.*

Suppose that $\mathcal{X}$ and $\mathcal{Y}$ are two real Banach spaces. Recall that a set $\mathcal{D} \subset \mathcal{Y}$ is called *precompact* of the closure $\overline{\mathcal{D}}$ of $\mathcal{D}$ is compact in $\mathcal{Y}$.

DEFINITION A.13. *An operator $A \in L(\mathcal{X}, \mathcal{Y})$ is called* compact *if for every sequence $\{x_n\}_{n\in\mathbb{N}} \subset \mathcal{X}$ the sequence $\{Ax_n\}_{n\in\mathbb{N}} \subset \mathcal{Y}$ has a convergent subsequence.*

REMARK A.14. Let $\mathcal{D} \subset \mathbb{R}^{\mathsf{d}}$ be an open, bounded and convex subset. Then, $\mathcal{X} = L^2(\mathcal{D})$ is a Hilbert space. Suppose that $k \in L^2(\mathcal{D} \times \mathcal{D})$ is a given kernel function. It is proved in [**23**, pp. 67-68] that the linear integral operator

$$\mathcal{T} : L^2(\mathcal{D}) \to L^2(\mathcal{D}), v \mapsto \mathcal{T}v = \int_{\mathcal{D}} k(\mu, \nu)v(\nu)\,\mathrm{d}\nu$$

is a compact operator. ◊

The proof of the following results can be found in [**19**, pp. 199-203], for instance.

THEOREM A.15. *Let $\mathcal{X}, \mathcal{Y}$ be two real Banach spaces and $\mathcal{A} \in L(\mathcal{X}, \mathcal{Y})$.*

1) *If $\{x_n\}_{n\in\mathbb{N}} \subset \mathcal{X}$ converges weakly to an element $x \in \mathcal{X}$ and $\mathcal{A}$ is compact, then the sequence $\{\mathcal{A}x_n\}_{n\in\mathbb{N}}$ converges strongly to $\mathcal{A}x$.*
2) *If $\{\mathcal{A}_n\}_{n\in\mathbb{N}} \subset L(\mathcal{X}, \mathcal{Y})$ is a sequence of compact operators with $\|\mathcal{A}_n - \mathcal{A}\|_{L(\mathcal{X},\mathcal{Y})} \to 0$ for $n \to \infty$. Then $\mathcal{A}$ is compact as well.*
3) *If $\mathcal{A}$ is a compact operator, then its Banach space adjoint $\mathcal{A}'$ is also compact.*
4) *If $\mathcal{Z}$ is a Banach space and $\mathcal{B} \in L(\mathcal{Y}, \mathcal{Z})$ holds, then $\mathcal{BA} : \mathcal{X} \to \mathcal{Z}$ is compact if $\mathcal{A}$ or $\mathcal{B}$ is a compact operator.*

REMARK A.16. Suppose that $\mathcal{X}$ and $\mathcal{Y}$ are two Hilbert spaces. If $\mathcal{A} \in L(\mathcal{X}, \mathcal{Y})$ is compact, its Banach space adjoint $\mathcal{A}' \in L(\mathcal{Y}', \mathcal{X}')$ is compact by Theorem A.15-3). Due to Definition A.5 the associated Hilbert space adjoint is given by $\mathcal{A}^\star = \mathcal{J}_{\mathcal{X}}^{-1}\mathcal{A}'\mathcal{J}_{\mathcal{Y}}$. Since $\mathcal{J}_{\mathcal{X}}^{-1}$ and $\mathcal{J}_{\mathcal{Y}}$ are isomorphisms, the operator $\mathcal{A}^\star$ is compact by Theorem A.15-4). ◊

THEOREM A.17 (Riesz-Schauder). *Let $\mathcal{X}$ be a Hilbert space and $\mathcal{A} \in L(\mathcal{X})$ be a compact operator. Then the spectrum $\sigma(\mathcal{A}A)$ is a discrete set having no limit points except perhaps 0. Furthermore, the space of eigenvectors corresponding to each nonzero $\lambda \in \sigma(\mathcal{A})$ is finite dimensional.*

THEOREM A.18 (Hilbert-Schmidt). *Let $\mathcal{X}$ be a Hilbert space and $\mathcal{A} \in L(\mathcal{X})$ compact and selfadjoint. Then, there is a complete orthonormal basis $\{\psi_i\}_{i\in\mathbb{N}} \subset \mathcal{X}$ with*

$$\mathcal{A}\psi_i = \lambda_i\psi_i \quad \textit{and} \quad \lambda_i \to 0 \textit{ as } i \to \infty.$$

B. Function Spaces

Let $\emptyset \neq \Omega \subset \mathbb{R}^{\mathsf{d}}$ be an open and bounded set. By

$$\int_\Omega \varphi(x)\,\mathrm{d}x$$

we denote the Lebesgue integral of $\varphi : \Omega \to \mathbb{R}$. For $1 \leq p < \infty$ we define

$$\|\varphi\|_{L^p(\Omega)} = \Big(\int_\Omega \big|\varphi(x)\big|^p\,\mathrm{d}x\Big)^{1/p}$$

and for $p = \infty$ set

$$\|\varphi\|_{L^\infty(\Omega)} = \operatorname{esssup}\big\{\big|\varphi(x)\big| \,:\, x \in \Omega\big\}.$$

For $p \in [1, \infty]$ the associated Lebesgue space $L^p(\Omega)$ is defined as

$$L^p(\Omega) = \big\{\varphi : \Omega \to \mathbb{R} \,\big|\, \varphi \text{ is Lebesgue measurable and } \|\varphi\|_{L^p(\Omega)} < \infty\big\}.$$

We identify two functions $\varphi, \phi \in L^p(\Omega)$ proivided $\|\varphi - \phi\|_{L^p(\Omega)} = 0$ holds true. It is well-known that $L^p(\Omega)$ is a Banach-space for any $p \in [1, \infty]$; see [**8**], for instance. Further, $L^2(\Omega)$ is a Hilbert space. By $C_0^\infty(\Omega)$ we denote the set of all $C^\infty(\Omega)$ functions with compact support in Ω. Further, the set of locally integrable functions $L^1_{\mathrm{loc}}(\Omega)$ is given by

$$L^1_{\mathrm{loc}}(\Omega) = \big\{\varphi : \Omega \to \mathbb{R} \,\big|\, \varphi \in L^1(\mathcal{K}) \text{ for any compact } \mathcal{K} \subset \Omega\big\}.$$

Before we turn to the notion of weak derivatives we introduce some notation. The d-tupel $\alpha = (\alpha_1, \dots, \alpha_{\mathsf{d}})$ is called a multi-index of nonegative integers α_i. We set

$$|\alpha| = \sum_{i=1}^{\mathsf{d}} \alpha_i.$$

For a function $\varphi \in C^\infty(\Omega)$ we set D^α for the partial derivative

$$\frac{\partial^{|\alpha|}\varphi}{\partial x_1^{\alpha_1} \dots \partial x_{\mathsf{d}}^{\alpha_{\mathsf{d}}}}.$$

Moreover, we define $x^\alpha = x_1^{\alpha_1} \dots x_{\mathsf{d}}^{\alpha_{\mathsf{d}}}$ for $x = (x_1, \dots, x_{\mathsf{d}}) \in \Omega$.

DEFINITION B.1. *Let $\alpha = (\alpha_1, \dots, \alpha_{\mathsf{d}})$ be a multi-index of nonegative integers α_i. A function $\varphi \in L^1_{\mathrm{loc}}(\Omega)$ has a* weak derivative *$D^\alpha_w \varphi$ provided there is a function $\phi \in L^1_{\mathrm{loc}}(\Omega)$ satisfying*

$$\int_\Omega \phi\psi \,\mathrm{d}x = (-1)^{|\alpha|} \int_\Omega \varphi D^\alpha \psi \,\mathrm{d}x \quad \textit{for all } \psi \in C_0^\infty(\Omega).$$

If such a function ϕ exists, we set $D^\alpha_w \varphi = \phi$.

Now we define function spaces for weakly differentiable functions.

DEFINITION B.2. *Let k be a nonnegative integer and $\varphi \in L^1_{\mathrm{loc}}(\Omega)$. Suppose that the weak derivative $D^\alpha_w \varphi$ exists for all multi-indices α satisfying $|\alpha| \le k$. Then, the* Sobolev norm *of φ is defined by*

$$\|\varphi\|_{W^{k,p}(\Omega)} = \Big(\sum_{|\alpha| \le k} \|D^\alpha_w \varphi\|^p_{L^p(\Omega)} \Big)^{1/p} \qquad \textit{for } p \in [1, \infty),$$

$$\|\varphi\|_{W^{k,\infty}(\Omega)} = \max_{|\alpha| \le k} \|D^\alpha_w \varphi\|_{L^\infty(\Omega)} \qquad \textit{for } p = \infty.$$

The Sobolev space *$W^{k,p}(\Omega)$ is given as*

$$W^{k,p}(\Omega) = \big\{ \varphi \in L^1_{\mathrm{loc}}(\Omega) \,\big|\, \|\varphi\|_{W^{k,p}(\Omega)} < \infty \big\}$$

for $1 \le p \le \infty$.

Next we introduce the so-called Bochner spaces. Let $\mathcal{X}$ be a Banach space and $T > 0$.

DEFINITION B.3 (Bochner spaces). 1) *We denote by $L^p(0,T;\mathcal{X})$, $1 \le p < \infty$, the space of (classes of) functions $t \mapsto \varphi(t) \in \mathcal{X}$ satisfying*

1a) *$t \mapsto \varphi(t)$ is measurable for $t \in [0,T]$;*

1b) $\|\varphi\|_{L^p(0,T;\mathcal{X})} = \Big(\int_\Omega \|\varphi(t)\|^p_{\mathcal{X}} \,\mathrm{d}t \Big)^{1/p} < \infty.$

2) *By $L^\infty(0,T;\mathcal{X})$ we denote the space of (classes of) functions $\varphi : [0,T] \to \mathcal{X}$ satisfying* 1a)

It is well-known that $L^p(0,T;\mathcal{X})$, $p \in [1,\infty]$, is a Banach space provided $\mathcal{X}$ is a Banach space.

Let V and H be two real, separable Hilbert spaces with inner product spaces $\langle \cdot, \cdot \rangle_V$ and $\langle \cdot, \cdot \rangle_H$, respectively. Moreover, we assume that V is dense in H with compact embedding. Hence, there exists a constant $C_V > 0$ satisfying

$$\|\varphi\|_H \le C_V \|\varphi\|_V \quad \text{for all } \varphi \in V. \tag{B.1}$$

By identifying H with its dual space (by using the Riesz theorem) we have

$$V \hookrightarrow H \simeq H' \hookrightarrow V',$$

where each space is dense in the following one. Then, the space

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

equipped with the norm

$$\|\varphi\|_{W(0,T)} = \left(\|\varphi\|^2_{L^2(0,T;V)} + \|\varphi_t\|^2_{L^2(0,T;V')} \right)^{1/2}, \quad \varphi \in W(0,T),$$

is a Hilbert space. Moreover, $W(0,T) \hookrightarrow C([0,T];H)$; see [**6**, p. 473]. Hence, $\varphi(0)$ and $\varphi(T)$ are meaningful for an element $\varphi \in W(0,T)$. The integration by parts formula reads

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

for φ, $\phi \in W(0,T)$. Moreover, we have the formula

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

see [**6**, p. 477], for example.

C. Evolution Problems

Let V and H be two real, separable Hilbert spaces with inner product spaces $\langle \cdot\,, \cdot \rangle_V$ and $\langle \cdot\,, \cdot \rangle_H$, respectively. Moreover, we assume that V is dense in H with compact embedding. Then, there exists a constant $C_V > 0$ satisfying (B.1). Suppose that f.a.a. $t \in [0,T]$ the bilinear form $a(t;\cdot\,,\cdot) : V \times V \to \mathbb{R}$ satisfies the following conditions:

1) $t \mapsto a(t;\cdot\,,\cdot)$ is measurable,
2) there exists a constant $\beta > 0$ (independent of t) so that

$$\left| a(t;\varphi,\phi) \right| \le \beta \|\varphi\|_V \|\phi\|_V \text{ for all } \varphi,\, \phi \in V \text{ and f.a.a. } t \in [0,T], \tag{C.1}$$

3) there are constants $\kappa > 0$ and $\eta \ge 0$, which are independent of t, with

$$a(t;\varphi,\varphi) \ge \kappa \|\varphi\|^2_V - \eta \|\varphi\|^2_H \text{ for all } \varphi \in V \text{ and f.a.a. } t \in [0,T]. \tag{C.2}$$

The bilinear form $a(t;\cdot\,,\cdot)$ defines a linear operator $\mathcal{A}(t) : V \to V'$ f.a.a. $t \in [0,T]$ by

$$\langle \mathcal{A}(t)\varphi, \phi \rangle_{V',V} = a(t;\varphi,\phi) \quad \text{for all } \varphi, \phi \in V \text{ and f.a.a. } t \in [0,T].$$

It follows from (C.1) that

$$\|\mathcal{A}(t)\|_{L(V,V')} \le \beta \quad \text{f.a.a. } t \in [0,T].$$

The *domain* of $\mathcal{A}(t)$ is defined as

$$D(\mathcal{A}(t)) = \left\{ \varphi \in V \,\middle|\, \mathcal{A}(t)\varphi \in H \right\}.$$

The following result is proved in [**6**, pp. 512-520].

THEOREM C.1. *Let the spaces V, H and the bilinearform $a(t;\cdot\,,\cdot)$ as introduced above. Then, for every $y_\circ \in H$ and $f \in L^2(0,T;V')$ there exists a unique solution $y \in W(0,T)$ satisfying*

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle y(t),\varphi\rangle_{V',V} + a(t;y(t),\varphi) = \langle f(t),\varphi\rangle_{V',V} \quad \text{for all } \varphi \in V \text{ and f.a.a. } t \in [0,T],$$
$$\langle y(0),\phi\rangle_H = \langle y_\circ,\phi\rangle_H \qquad \text{for all } \phi \in H.$$

REMARK C.2. Setting $y(t) = \exp(\eta t)z(t)$ with η from (C.2) we infer that $v(t) = \exp(-\eta t)y(t)$ solves

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle v(t),\varphi\rangle_{V',V} + \tilde{a}(t;v(t),\varphi) = \langle \tilde{f}(t),\varphi\rangle_{V',V} \quad \text{for all } \varphi \in V \text{ and f.a.a. } t \in [0,T],$$
$$\langle v(0),\phi\rangle_H = \langle y_\circ,\phi\rangle_H \qquad \text{for all } \phi \in H$$

with

$$\tilde{a}(t;\varphi,\phi) = a(t;\varphi,\phi) + \langle v(t),\varphi\rangle_H \quad \text{for all } \varphi \in V \text{ and f.a.a. } t \in [0,T]$$

and $\tilde{f}(t) = \exp(-\eta t)f(t) \in V'$ f.a.a. $t \in [0,T]$. Using (B.1), (C.1) and (C.2) we obtain

$$\big|\tilde{a}(t;\varphi,\phi)\big| \le \beta\,\|\varphi\|_V\|\phi\|_V + \eta\,\|\varphi\|_H\|\phi\|_H \le \big(\beta + \eta C_V^2\big)\|\varphi\|_V\|\phi\|_V$$

for all φ, $\phi \in V$ and f.a.a. $t \in [0,T]$. Thus, $\tilde{a}(t;\cdot\,,\cdot)$ is a bounded bilinear form. Moreover,

$$\tilde{a}(t;\varphi,\varphi) \ge \kappa\,\|\varphi\|_V^2 \quad \text{for all } \varphi \in V \text{ and f.a.a. } t \in [0,T],$$

i.e., the bilinear form $\tilde{a}(t;\cdot\,,\cdot)$ is coercive. ◊

COROLLARY C.3. *Let all assumptions of Theorem* C.1 *be satisfied. In addition, we have $a(t;\cdot\,,\cdot) = a(\cdot\,,\cdot)$, i.e., the bilinear form is independent of t. If $y_\circ \in V$ and $f \in L^2(0,T;H)$ hold, then $u \in L^\infty(0,T;V) \cap L^2(0,T;D(\mathcal{A}))$, where the operator $\mathcal{A} \in L(V,V')$ is given by*

$$\langle \mathcal{A}\varphi,\phi\rangle_{V',V} = a(\varphi,\phi) \quad \textit{for all } \varphi,\phi \in V.$$

For a proof we refer the reader to [**6**, pp. 532-533].

D. Nonlinear Optimization

We consider the problem

$$(\mathtt{P}) \qquad \min J(x) \quad \text{s.t.} \quad e(x) = 0,$$

where $J : \mathbb{R}^n \to \mathbb{R}$ denotes the *cost functional* or *objective* and $e : \mathbb{R}^n \to \mathbb{R}^m$, $m \le n$ are the *equality constraints*. A point $x \in \mathbb{R}^n$ is called *admissible* provided $e(x) = 0$ holds true. The set of admissible solutions is defined as

$$\mathcal{F}(\mathtt{P}) = \big\{x \in \mathbb{R}^n \,\big|\, e(x) = 0\big\}.$$

DEFINITION D.1. *Let $\bar{x} \in \mathbb{R}^n$ be given.*

1) *The point $\bar{x}$ is called a* local solution *to* $(\mathtt{P})$ *if $\bar{x} \in \mathcal{F}(\mathtt{P})$ holds and $J(\bar{x}) \le J(x)$ for all $x \in \mathcal{U}(\bar{x}) \cap \mathcal{F}(\mathtt{P})$, where $\mathcal{U}(\bar{x}) \subset \mathbb{R}^n$ is an open, nonempty neighborhood of $\bar{x}$.*
2) *The point $\bar{x}$ is called a* strict local solution *to* $(\mathtt{P})$ *if $\bar{x} \in \mathcal{F}(\mathtt{P})$ holds and $J(\bar{x}) < J(x)$ for all $x \in \mathcal{U}(\bar{x}) \cap \mathcal{F}(\mathtt{P})$, where $\mathcal{U}(\bar{x}) \subset \mathbb{R}^n$ is an open, nonempty neighborhood of $\bar{x}$.*

3) *The point $\bar{x}$ is called a* global solution *to* (P) *if $\bar{x} \in \mathcal{F}(\mathtt{P})$ holds and $J(\bar{x}) \leq J(x)$ for all $x \in \mathcal{F}(\mathtt{P})$.*
4) *The point $\bar{x}$ is called a* strict global solution *to* (P) *if $\bar{x} \in \mathcal{F}(\mathtt{P})$ holds and $J(\bar{x}) < J(x)$ for all $x \in \mathcal{F}(\mathtt{P})$.*

To characterize solutions to (P) we need the notion of the tangent plane. A *curve* in a hyperplane $\mathcal{H} \subset \mathbb{R}^n$ is a family of points $x(t) \in \mathcal{H}$, where $x : [a, b] \to \mathcal{H}$ is continuous and $a < b$ holds. The curve x is *differentiable* in t provided $\dot{x}(t) = \frac{\mathrm{d}}{\mathrm{d}t}x(t)$ exists. If $\ddot{x}(t) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}x(t)$ is defined, the curve x is said to be *twice differentiable*. We say that the curve x goes through the point $\bar{x} \in \mathcal{H}$ if there exists a $\bar{t} \in [a, b]$ so that $x(\bar{t}) = \bar{x}$ is satisfied. The set of the tangential vectors $\dot{x}(\bar{t})$ of all differentiable curves going through $\bar{x}$ is called the *tangent plane* at $\bar{x}$.

Definition D.2. *A point $\bar{x} \in \mathcal{F}(\mathtt{P})$ is called* regular *with respect to the constraint $e(x) = 0$ if the m gradients $\{\nabla e_i(\bar{x})\}_{i=1}^m \in \mathbb{R}^n$ are linearly independent in $\mathbb{R}^n$.*

For a proof of the following characterization of the tangent plane we refer the reader to [**18**].

Theorem D.3. *Suppose that $\bar{x} \in \mathcal{F}(\mathtt{P})$ is a regular point. Then the tangent plane at $\bar{x}$ is equal to the set*

$$\ker \nabla e(\bar{x}) = \left\{ v \in \mathbb{R}^n \,\middle|\, \nabla e(\bar{x})v = 0 \right\} \subset \mathbb{R}^n,$$

where

$$\nabla e(\bar{x}) = \begin{pmatrix} \nabla e_1(\bar{x})^\top \\ \vdots \\ \nabla e_m(\bar{x})^\top \end{pmatrix} \in \mathbb{R}^{m \times n}$$

is the Jacobian of e at $\bar{x}$.

Now we can formulate the following first-order necessary optimality conditions for (P). A proof can be found in [**18**], for instance.

Theorem D.4 (First-order necessary optimality conditions). *Suppose that J and e are continuously differentiable. Moreover, let $\bar{x}$ be a local solution to* (P) *and a regular point for $e(x) = 0$. Then, there exists a unique Lagrange multiplier $\bar{\lambda} = (\bar{\lambda}_1, \ldots, \bar{\lambda}_m) \in \mathbb{R}^m$ solving*

$$\nabla J(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i \nabla e_i(\bar{x}) = \nabla J(\bar{x}) + \nabla e(\bar{x})^\top \bar{\lambda} = 0. \tag{D.1}$$

Let us introduce the *Lagrange function* $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ by

$$\mathcal{L}(x, \lambda) = J(x) + \langle \lambda, e(x) \rangle_{\mathbb{R}^m} = J(x) + \lambda^\top e(x).$$

Then, we can express (D.1) as

$$\nabla_x \mathcal{L}(\bar{x}, \bar{\lambda}) = \nabla J(\bar{x}) + \nabla e(\bar{x})^\top \bar{\lambda} = 0 \in \mathbb{R}^n. \tag{D.2a}$$

Moreover, the equality constraint is satisfied at $\bar{x}$, so that we have

$$\nabla_\lambda \mathcal{L}(\bar{x}, \bar{\lambda}) = e(\bar{x}) = 0 \in \mathbb{R}^m. \tag{D.2b}$$

System (D.2) consists of $n + m$ equations for the unknown vectors $\bar{x} \in \mathbb{R}^n$ and $\bar{\lambda} \in \mathbb{R}^m$.

If J and e are more regular, we can formulate necessary and sufficient second-order optimality conditions.

THEOREM D.5 (Second-order necessary optimality conditions). *Suppose that J and e are twice continuously differentiable. Moreover, let $\bar{x}$ be a local solution to* (P) *and a regular point for $e(x) = 0$. Then, the $n \times n$ matrix*

$$\nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{\lambda}) = \nabla^2 J(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i \nabla^2 e_i(\bar{x})$$

is positive semidefinite on the set $\ker \nabla e(\bar{x}) \subset \mathbb{R}^n$, *i.e.*

$$v^\top \nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{\lambda}) v \geq 0 \quad \textit{for all } v \in \ker \nabla e(\bar{x}).$$

Here, $\bar{\lambda} = (\bar{\lambda}_1, \ldots, \bar{\lambda}_m)^\top \in \mathbb{R}^m$ denotes the unique Lagrange multiplier introduced in Theorem D.4.

For a proof of Theorem D.5 we refer the reader to [**18**]. To ensure that a point $\bar{x} \in \mathcal{F}(\mathtt{P})$ is a solution to (P) we have to guarantee sufficient optimality conditions. A proof of the following second-order condition can be found in [**18**], for instance.

THEOREM D.6 (Second-order sufficient optimality conditions). *Suppose that J and e are twice continuously differentiable. Moreover, let the pair $(\bar{x}, \bar{\lambda}) \in \mathbb{R}^n \times \mathbb{R}^m$ satisfy the necessary optimality conditions* (D.2). *Further, $\bar{x}$ is a regular point for $e(x) = 0$. Then the matrix $\nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{\lambda})$ is positive definite on the set* $\ker \nabla e(\bar{x}) \subset \mathbb{R}^n$, *i.e.*

$$v^\top \nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{\lambda}) v > 0 \quad \textit{for all } v \in \ker \nabla e(\bar{x}) \setminus \{0\}.$$

Bibliography

[1] M. Barrault, Y. Maday, N. C. Nguyen and A. T. Patera. An 'empirical interpolation' method: application to efficient reduced-basis descretization of partial differential equations. *Comptes Rendus Mathematique*, 339(9):667-672, 2004.

[2] R.E. Bellman. The theory of dynamic programming. *Proc. Nat. Acad. Sci.*, USA, 38:716-719, 1952.

[3] S. Chaturantabut and D.C. Sorensen. Application of POD and DEIM on a Dimension Reduction of Nonlinear Miscible Viscous Fingering in Porous Media. *Technical Report*, TR09-25, RICE University, 2009.

[4] S. Chaturantabut and D.C. Sorensen. A state space estimate for POD-DEIM Nonlinear Model Reduction. *Technical Report*, TR10-32, RICE University, 2010.

[5] S. Chaturantabut and D.C. Sorensen. Nonlinear model reduction via discrete empirical interpolation. *SIAM J. Sci. Comput.*, 32:2737-2764, 2010.

[6] R. Dautray and J.-L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology. Volume 5: Evolution Problems I.* Springer-Verlag, Berlin, 2000.

[7] P. Dorato, C. Abdallah, and V. Cerone. *Linear-Quadratic Control. An Introduction.* Prentice Hall, Englewood Cliffs, New Jersey 07632, 1995.

[8] L.C. Evans. *Partial Differential Equations.* American Math. Society, Providence, Rhode Island, 2002.

[9] M. Grepl. *Certified reduced basis method for nonaffine linear time-varying and nonlinear parabolic partial differential equations. WSPC: M3AS*, 22(3), 2012.

[10] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints.* Mathematical Modelling: Theory and Applications, vol. 23, Springer Verlag, 2009.

[11] P. Holmes, J.L. Lumley, G. Berkooz, and C.W. Romley. *Turbulence, Coherent Structures, Dynamical Systems and Symmetry.* Cambridge Monographs on Mechanics, Cambridge University Press, 2nd edition, 2012.

[12] 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.

[13] T. Kato. *Perturbation Theory for Linear Operators.* Springer-Verlag, Berlin, 1980.

[14] K. Kunisch and S. Volkwein. Galerkin proper orthogonal decomposition methods for parabolic problems. *Numerische Mathematik*, 90:117-148, 2001.

[15] K. Kunisch and S. Volkwein. Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics. *SIAM Journal on Numerical Analysis*, 40:492-515, 2002.

[16] 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.

[17] B. Noble. *Applied Linear Algebra.* Englewood Cliffs, NJ : Prentice-Hall, 1969.

[18] J. Nocedal and S.J. Wright. *Numerical Optimization.* Springer Series in Operation Research, second edition, 2006.

[19] M. Reed and B. Simon. *Methods of Modern Mathematical Physics I: Functional Analysis.* Academic Press, New York, 1980.

[20] L. Sirovich. Turbulence and the dynamics of coherent structures, parts I-III. *Quarterly of Applied Mathematicss*, XLV:561-590, 1987.

[21] R. Temam. *Infinite-Dimensional Dynamical Systems in Mechanics and Physics.* Applied Mathematical Sciences 68, Springer, New York, 1988.

[22] S. Volkwein. Optimal control of a phase-field model using proper orthogonal decomposition. *Zeitschrift für Angewandte Mathematik und Mechanik*, 81:83-97, 2001.

[23] D. Werner. *Funktionalanalysis*. Springer, Berlin, 2011.
[24] K. Zhou, J.C. Doyle, and K. Glover. *Robust and Optimal Control*. Prentice Hall, Upper Saddle River, New Jersey, 07458, 1996.