MODEL REDUCTION USING PROPER ORTHOGONAL DECOMPOSITION

S. VOLKWEIN

Abstract. In this lecture notes an introduction to model reduction utilizing proper orthogonal decomposition (POD) is given. The close connection between POD and singular value decomposition (SVD) of rectangular matrices is emphasized. As an application POD is used to derive a reduced-order model for non-linear initial value problems. The strategy is extended to linear-quadratic optimal control problems governed by ordinary differential equations. The relationship to classical model reduction techniques like balanced truncation is studied.

1. The POD method in $\mathbb{R}^m$

In this section we introduce the POD method in the Euclidean space $\mathbb{R}^m$ and study the close connection to the SVD of rectangular matrices; see [6]. We also refer to the monograph [3].

1.1. POD and 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,

$$\bar{y} = \frac{1}{n} \sum_{j=1}^{n} y_j \tag{1.1}$$

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

SVD [10] guarantees the existence of real numbers $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d > 0$ and orthogonal matrices $U \in \mathbb{R}^{m \times m}$ with columns $\{u_i\}_{i=1}^m$ and $V \in \mathbb{R}^{n \times n}$ with columns $\{v_i\}_{i=1}^n$ such that

$$U^T Y V = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} =: \Sigma \in \mathbb{R}^{m \times n}, \tag{1.2}$$

where $D = \operatorname{diag}(\sigma_1, \ldots, \sigma_d) \in \mathbb{R}^{d \times d}$ and the zeros in (1.2) denote matrices of appropriate dimensions. Moreover the vectors $\{u_i\}_{i=1}^d$ and $\{v_i\}_{i=1}^d$ satisfy

$$Y v_i = \sigma_i u_i \quad \text{and} \quad Y^T u_i = \sigma_i v_i \quad \text{for } i = 1, \ldots, d. \tag{1.3}$$

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

From (1.2) we deduce that

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

Date: December 7, 2011.

Key words and phrases. Proper orthogonal decomposition, singular value decomposition, model reduction, error estimates, linear-quadratic regulator problems, balanced truncation.

It follows that Y can also be expressed as

$$Y = U^d D (V^d)^T, \tag{1.4}$$

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

$$U^d_{ij} = U_{ij} \quad \text{for } 1 \le i \le m,\ 1 \le j \le d,$$
$$V^d_{ij} = V_{ij} \quad \text{for } 1 \le i \le n,\ 1 \le j \le d.$$

Setting $B^d = D(V^d)^T \in \mathbb{R}^{d\times n}$ we can write (1.4) in the form

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

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

$$\begin{aligned} y_j = \sum_{i=1}^d B^d_{ij} U^d_{\cdot,i} = \sum_{i=1}^d \big(D(V^d)^T\big)_{ij} u_i = \sum_{i=1}^d \big(\underbrace{(U^d)^T U^d}_{=I^d\in\mathbb{R}^{d\times d}} D(V^d)^T\big)_{ij} u_i \\ \overset{(1.4)}{=} \sum_{i=1}^d \big((U^d)^T Y\big)_{ij} u_i = \sum_{i=1}^d \Big(\underbrace{\sum_{k=1}^m U^d_{ki} Y_{kj}}_{=u_i^T y_j}\Big) u_i = \sum_{i=1}^d \langle u_i, y_j\rangle_{\mathbb{R}^m} u_i, \end{aligned}$$

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

$$y_j = \sum_{i=1}^d \langle y_j, u_i\rangle_{\mathbb{R}^m} u_i \quad \text{for } j = 1, \dots, n \tag{1.5}$$

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

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

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

Note that $(\mathbf{P}^1)$ is a constrained optimization problem that can be solved by considering first-order necessary optimality conditions. We introduce the function $e : \mathbb{R}^m \to \mathbb{R}$ by $e(u) = 1 - \|u\|^2_{\mathbb{R}^m}$ for $u \in \mathbb{R}^m$. Then, the equality constraint in $(\mathbf{P}^1)$ can be expressed as $e(u) = 0$. Notice that $\nabla e(u) = 2u^T$ is linear independent if $u \neq 0$ holds. In particular, a solution to $(\mathbf{P}^1)$ satisfies $u \neq 0$. Thus, any solution to $(\mathbf{P}^1)$ is a *regular point*. Let $\mathcal{L} : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ be the Lagrange functional associated with $(\mathbf{P}^1)$, i.e.,

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

Suppose that $u \in \mathbb{R}^m$ is a solution to $(\mathbf{P}^1)$. Since u is regular, there exists a Lagrange multiplier satisfying the first-order necessary optimality condition

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

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

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

Thus,

$$\nabla_u \mathcal{L}(u,\lambda) = 2\big(YY^Tu - \lambda u\big) \overset{!}{=} 0 \quad \text{in } \mathbb{R}^m. \tag{1.6}$$

Equation (1.6) yields the eigenvalue problem

$$YY^Tu = \lambda u \quad \text{in } \mathbb{R}^m. \tag{1.7a}$$

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

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

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

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

$$\|u\|_{\mathbb{R}^m} = 1. \tag{1.7b}$$

Due to SVD the vector u_1 solves (1.7) and

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

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

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

Thus,

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

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

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

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

SVD implies that u_2 is a solution to $(\mathbf{P}^2)$ and $\operatorname{argmax}(\mathbf{P}^2) = \sigma_2^2 = \lambda_2$. In fact, u_2 solves the first-order necessary optimality conditions (1.7) and for

$$\tilde{u} = \sum_{i=2}^{m} \langle \tilde{u}, u_i\rangle_{\mathbb{R}^m} u_i \in \operatorname{span}\{u_1\}^{\perp}$$

we have

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

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

Theorem 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 $Y = U\Sigma V^T$ be the singular value decomposition of Y, where $U = [u_1, \dots, u_m] \in \mathbb{R}^{m\times m}$, $V = [v_1, \dots, v_n] \in \mathbb{R}^{n\times n}$ are orthogonal matrices and the matrix $\Sigma \in \mathbb{R}^{m\times n}$ has the form as (1.2). Then, for any $\ell \in \{1, \dots, d\}$ the solution to*

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

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

$$\operatorname{argmax}(\mathbf{P}^{\ell}) = \sum_{i=1}^{\ell} \sigma_i^2 = \sum_{i=1}^{\ell} \lambda_i. \tag{1.8}$$

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

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

by

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

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

$$\frac{\partial \mathcal{L}}{\partial \psi_k}(\psi_1, \ldots, \psi_\ell, \Lambda)\delta\psi_k = 0 \quad \text{for all } \delta\psi_k \in \mathbb{R}^m \text{ and } k \in \{1, \ldots, \ell\}. \tag{1.9}$$

From

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

and (1.9) we infer that

$$\sum_{j=1}^{n} \langle y_j, \psi_k \rangle_{\mathbb{R}^m} y_j = \frac{1}{2} \sum_{i=1}^{\ell} (\lambda_{ik} + \lambda_{ki}) \psi_i \quad \text{in } \mathbb{R}^m \text{ and for all } k \in \{1, \ldots, \ell\}. \tag{1.10}$$

Note that

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

Thus, condition (1.10) can be expressed as

$$YY^T\psi_k = \frac{1}{2} \sum_{i=1}^{\ell} (\lambda_{ik} + \lambda_{ki}) \psi_i \quad \text{in } \mathbb{R}^m \text{ and for all } k \in \{1, \ldots, \ell\}. \tag{1.11}$$

Now we proceed by induction. For $\ell = 1$ we have $k = 1$. It follows from (1.11) that

$$YY^T\psi_1 = \lambda_1\psi_1 \quad \text{in } \mathbb{R}^m \tag{1.12}$$

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

$$YY^T\psi_k = \lambda_k\psi_k \quad \text{in } \mathbb{R}^m \text{ and for all } k \in \{1, \ldots, \ell\}. \tag{1.13}$$

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

$$YY^T\psi_k = \lambda_k\psi_k \quad \text{in } \mathbb{R}^m \text{ and for all } k \in \{1,\dots,\ell+1\}. \tag{1.14}$$

By assumption we have (1.13). Thus, we only have to prove that

$$YY^T\psi_{\ell+1} = \lambda_{\ell+1}\psi_{\ell+1} \quad \text{in } \mathbb{R}^m. \tag{1.15}$$

Due to (1.11) we have

$$YY^T\psi_{\ell+1} = \frac{1}{2}\sum_{i=1}^{\ell+1}(\lambda_{i,\ell+1} + \lambda_{\ell+1,i})\,\psi_i \quad \text{in } \mathbb{R}^m. \tag{1.16}$$

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 (1.13) and the symmetry of YY^T we have for any $j \in \{1,\dots,\ell\}$

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

This gives

$$\lambda_{\ell+1,i} = -\lambda_{i,\ell+1} \quad \text{for any } i \in \{1,\dots,\ell\}. \tag{1.17}$$

Inserting (1.17) into (1.16) we obtain

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

Setting $\lambda_{\ell+1} = \lambda_{\ell+1,\ell+1}$ we obtain (1.15).
Summarizing, the necessary optimaity conditions for $(\mathbf{P}^\ell)$ are given by the symmetric $m \times m$ eigenvalue problem

$$YY^Tu_i = \lambda_i u_i \quad \text{for } i = 1,\dots,\ell. \tag{1.18}$$

It follows from SVD that $\{u_i\}_{i=1}^{\ell}$ solves (1.18). The proof that $\{u_i\}_{i=1}^{\ell}$ 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 1.2). □

Motivated by the previous theorem we give the next definition.

Definition 1.2. *For $\ell \in \{1,\dots,d\}$ the vectors $\{u_i\}_{i=1}^{\ell}$ are called* POD basis of rank ℓ.

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

Corollary 1.3 (Optimality of the POD basis)**.** *Let all hypotheses of Theorem* 1.1 *be satisfied. Suppose that $\hat{U}^d \in \mathbb{R}^{m\times d}$ denotes a matrix with pairwise orthonormal vectors $\hat{u}_i$ and that the expansion of the columns of Y in the basis $\{\hat{u}_i\}_{i=1}^{d}$ be given by*

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

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

$$\|Y - U^\ell B^\ell\|_F \leq \|Y - \hat{U}^\ell C^\ell\|_F. \tag{1.19}$$

In (1.19), $\|\cdot\|_F$ *denotes the Frobenius norm given by*

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

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

Remark 1.4. Notice that

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

Analogously,

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

Thus, (1.19) implies that

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

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

$$\min_{\tilde{u}_1,\ldots,\tilde{u}_\ell \in \mathbb{R}^m} \sum_{j=1}^{n} \left\|y_j - \sum_{i=1}^{\ell} \langle y_j, \tilde{u}_i\rangle_{\mathbb{R}^m} \tilde{u}_i\right\|_{\mathbb{R}^m}^2 \text{ s.t. } \langle \tilde{u}_i, \tilde{u}_j\rangle_{\mathbb{R}^m} = \delta_{ij},\ 1 \leq i, j \leq \ell. \tag{1.20}$$

◊

Proof of Corollary 1.3. Note that (see Exercise 1.3))

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

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

$$\begin{aligned}\|Y - U^\ell B^\ell\|_F^2 &= \|U^k(B^d - B_0^\ell)\|_F^2 = \|B^d - B_0^\ell\|_F^2 = \sum_{i=\ell+1}^{d}\sum_{j=1}^{n} \left|B_{ij}^d\right|^2 \\ &= \sum_{i=\ell+1}^{d}\sum_{j=1}^{n} \left|\langle y_j, u_i\rangle_{\mathbb{R}^m}\right|^2 \\ &= \sum_{i=\ell+1}^{d}\sum_{j=1}^{n} \left\langle \langle y_j, u_i\rangle_{\mathbb{R}^m} y_j, u_i\right\rangle_{\mathbb{R}^m} = \sum_{i=\ell+1}^{d} \langle YY^T u_i, u_i\rangle_{\mathbb{R}^m} \\ &= \sum_{i=\ell+1}^{d} \sigma_i^2,\end{aligned} \tag{1.21}$$

By Theorem 1.1 the vectors $u_1, \ldots, u_\ell$ solve $(\mathbf{P}^\ell)$. From (1.21),

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

and

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

we infer that

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

which gives (1.19). □

Remark 1.5. It follows from Corollary 1.3 that the POD basis of rank ℓ is optimal in the sense of representing in the mean the columns $\{y_j\}_{j=1}^n$ of Y as a linear combination by an orthonormal basis of rank ℓ:

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

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

The next corollary states that the POD coefficients are uncorrelated.

Corollary 1.6 (Uncorrelated POD coefficients)**.** *Let all hypotheses of Theorem* 1.1 *hold. Then.*

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

Proof. The claim follows from (1.18) and $\langle u_i, u_k\rangle_{\mathbb{R}^m} = \delta_{ik}$ for $1 \le i, k \le \ell$:

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

□

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

$$Y^T Y v_i = \lambda_i v_i \quad \text{for } i = 1, \dots, \ell \tag{1.22}$$

and set, by (1.3),

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

For historical reasons [13] 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 (1.18).

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

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

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

1.2. The POD method with a weighted inner product. Let us endow the Euclidean space $\mathbb{R}^m$ with the weighted inner product

$$\langle u, \tilde{u}\rangle_W = u^T W \tilde{u} = \langle u, W\tilde{u}\rangle_{\mathbb{R}^m} = \langle Wu, \tilde{u}\rangle_{\mathbb{R}^m} \quad \text{for } u, \tilde{u} \in \mathbb{R}^m, \tag{1.23}$$

where $W \in \mathbb{R}^{m\times m}$ is a symmetric, positive-definite matrix. Furthermore, let $\|u\|_W = \sqrt{\langle u, u\rangle_W}$ for $u \in \mathbb{R}^m$ be the associated induced norm. For the choice $W = I$, the inner product (1.23) coincides the Euclidean inner product.

Example 1.7. Let us motivate the weighted inner product by an example. Suppose that $\Omega = (0, 1) \subset \mathbb{R}$ holds. We consider the space $L^2(\Omega)$ of square integrable functions on Ω:

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

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

$$\langle \varphi, \tilde{\varphi} \rangle_{L_h^2(\Omega)} = h\left(\frac{\varphi_1^h \tilde{\varphi}_1^h}{2} + \sum_{i=2}^{m-1} \left(\varphi_i^h \tilde{\varphi}_i^h \right) + \frac{\varphi_m^h \tilde{\varphi}_m^h}{2} \right), \tag{1.24}$$

where

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

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

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

Thus, the discrete L^2-inner product can be written as a weighted inner product of the form (1.23). ◊

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

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

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

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

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

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

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

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

Thus,

$$\nabla_u \mathcal{L}(u,\lambda) = 2\big(WYY^TWu - \lambda Wu\big) \stackrel{!}{=} 0 \quad \text{in } \mathbb{R}^m. \tag{1.25}$$

Equation (1.25) yields the generalized eigenvalue problem

$$(WY)(WY)^T u = \lambda W u. \tag{1.26}$$

Since W is symmetric and positive definite, W possesses an eigenvalue decomposition of the form $W = QDQ^T$, where $D = \operatorname{diag}(\eta_1, \ldots, \eta_m)$ 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

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

Note that $(W^\alpha)^{-1} = W^{-\alpha}$ and $W^{\alpha+\beta} = W^\alpha W^\beta$ for α, $\beta \in \mathbb{R}$; see Exercise 1.4). Moreover, we have

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

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

Setting $\bar{u} = W^{1/2}u \in \mathbb{R}^m$ and $\bar{Y} = W^{1/2}Y \in \mathbb{R}^{m\times n}$ and multiplying (1.26) by $W^{-1/2}$ from the left we deduce the symmetric, $m \times m$ eigenvalue problem

$$\bar{Y}\bar{Y}^T\bar{u} = \lambda\bar{u} \quad \text{in } \mathbb{R}^m. \tag{1.27a}$$

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

$$\|\bar{u}\|_{\mathbb{R}^m} = 1. \tag{1.27b}$$

Thus, the first-order optimality conditions (1.27) for $(\mathbf{P}^1_W)$ are — as for $(\mathbf{P}^1)$ (compare (1.7)) — an $m \times m$ eigenvalue problem, but the matrix Y as well as the vector u have to be weighted by the matrix $W^{1/2}$.

It can be shown (see Exersice 1.4.1)) that

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

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

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

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

$$u_1 = W^{-1/2}\bar{u}_1 = \frac{1}{\sqrt{\lambda_1}} W^{-1/2}\bar{Y}\bar{v}_1 = \frac{1}{\sqrt{\lambda_1}} Y\bar{v}_1. \tag{1.28}$$

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

Theorem 1.8. *Let $Y \in \mathbb{R}^{m\times n}$ be a given matrix with rank $d \leq \min\{m,n\}$, W a symmetric, positive definite matrix, $\bar{Y} = W^{1/2}Y$ and $\ell \in \{1,\ldots,d\}$. Further, let $\bar{Y} = \bar{U}\Sigma\bar{V}^T$ be the singular value decomposition of $\bar{Y}$, where $\bar{U} = [\bar{u}_1, \ldots, \bar{u}_m] \in \mathbb{R}^{m\times m}$, $\bar{V} = [\bar{v}_1, \ldots, \bar{v}_n] \in \mathbb{R}^{n\times n}$ are orthogonal matrices and the matrix Σ has the form*

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

Then the solution to

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

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

$$\operatorname{argmax}(\mathbf{P}_W^\ell) = \sum_{i=1}^{\ell} \sigma_i^2 = \sum_{i=1}^{\ell} \lambda_i. \tag{1.29}$$

Proof. Using similar arguments as in the proof of Theorem 1.1 one can prove that $\{u_i\}_{i=1}^\ell$ solves $(\mathbf{P}_W^\ell)$; see Exersice 1.4). □

Remark 1.9. Due to SVD and $\bar{Y}^T\bar{Y} = Y^TWY$ the POD basis $\{u_i\}_{i=1}^\ell$ of rank ℓ can be determined by the method of snapshots as follows: Solve the symmetric $n \times n$ eigenvalue problem

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

and set

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

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

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

For $m \gg n$ the method of snapshots turns out to be faster than computing the POD basis via (1.27). Notice that the matrix $W^{1/2}$ is also not required for the method of snapshots. ◊

1.3. Application to time-dependent systems. For $T > 0$ we consider the semi-linear initial value problem

$$\dot{y}(t) = Ay(t) + f(t, y(t)) \quad \text{for } t \in (0,T], \tag{1.30a}$$

$$y(0) = y_0, \tag{1.30b}$$

where $y_0 \in \mathbb{R}^m$ is a chosen initial condition, $A \in \mathbb{R}^{m\times m}$ is a given matrix, $f : [0,T] \times \mathbb{R}^m \to \mathbb{R}^m$ is continuous in both arguments and locally Lipschitz-continuous with respect to the second argument. It is well known that (1.30) has a unique (classical) solution $y \in C^1(0,T;\mathbb{R}^m) \cap C([0,T];\mathbb{R}^m)$ given by the implicit integral representation

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

with $e^{tA} = \sum_{i=0}^\infty t^n A^n/(n!)$. Let $0 \le t_1 < t_2 < \ldots < t_n \le T$ be a given time grid in the interval $[0,T]$. For simplicity of the presentation, the time grid is assumed to be equidistant with step-size $\Delta t = T/(n-1)$, i.e., $t_j = (j-1)\Delta t$. We suppose that we know the solution to (1.30) at the given time instances t_j, $j \in \{1,\dots,n\}$. Our goal is to determine a POD basis of rank $\ell \le n$ that desribes the ensemble

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

as well as possible with respect to the weighted inner product:
$(\hat{\mathbf{P}}_W^{n,\ell})$

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

where the α_j's denote non-negative weights which will be specified later on. Note that for $\alpha_j = 1$ for $j = 1,\dots,n$ and $W = I$ problem $(\hat{\mathbf{P}}_W^{n,\ell})$ coincides with (1.20).

Example 1.10. Let us consider the following one-dimensional heat equation:

$$\theta_t(t,x) = \theta_{xx}(t,x) \qquad \text{for all } (t,x) \in Q = (0,T)\times\Omega, \tag{1.31a}$$

$$\theta_x(t,0) = \theta_x(t,1) = 0 \qquad \text{for all } t \in (0,T), \tag{1.31b}$$

$$\theta(0,x) = \theta_0(x) \qquad \text{for all } x \in \Omega = (0,1) \subseteq \mathbb{R}, \tag{1.31c}$$

where $\theta_0 \in C(\overline{\Omega})$ is a given initial condition. To solve (1.31) numerically we apply a classical finite difference approximation for the spatial variable x. In Example 1.7 we have introduced the spatial grid $\{x_i\}_{i=1}^m$ in the interval $[0,1]$. Let us denote by $y_i : [0,T] \to \mathbb{R}$ the numerical approximation for $\theta(\cdot\,, x_i)$ for $i = 1,\dots,m$. The second partial derivative θ_{xx} in (1.31a) and the boundary conditions (1.31b) 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:

$$\left\{ \begin{array}{rcl} \dot{y}_1(t) &=& \dfrac{-2y_1(t) + 2y_2(t)}{h^2}, \\ \dot{y}_i(t) &=& \dfrac{y_{i-1}(t) - 2y_i(t) + y_{i+1}(t)}{h^2}, \quad i = 2,\dots,m-1, \\ \dot{y}_m(t) &=& \dfrac{-2y_m(t) + 2y_{m-1}(t)}{h^2} \end{array} \right. \tag{1.32a}$$

for $t \in (0,T]$. From (1.31c) we infer the initial condion

$$y_i(0) = \theta_0(x_i), \quad i = 1,\dots,m. \tag{1.32b}$$

Introducing the matrix

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

and the vectors

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

we can express (1.32) in the form

$$\begin{array}{rcl} \dot{y}(t) &=& Ay(t) \quad \text{for } t \in (0,T], \\ y(0) &=& y_0 \end{array} \tag{1.33}$$

Setting $f \equiv 0$ the linear initial-value problem coincides with (1.30). 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 (1.24); see Example 1.7. Next we choose a time grid $\{t_j\}_{j=1}^n$ in the interval $[0,T]$ and define $y_j = y(t_j)$ for $j = 1,\dots,n$. If we are interested in finding a POD basis of rank $\ell \leq n$ 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 1.1 and 1.2, i.e., we use the Lagrangian framework. Thus, we introduce the Lagrange functional

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

by

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

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

$$\nabla_{u_i}\mathcal{L}(u_1,\dots,u_\ell,\Lambda) \stackrel{!}{=} 0 \quad \text{in } \mathbb{R}^m,\ 1 \leq i \leq \ell, \tag{1.34a}$$

and

$$\langle u_i, u_j\rangle_W \stackrel{!}{=} \delta_{ij}, \quad 1 \leq i,j \leq \ell. \tag{1.34b}$$

From (1.34a) we derive

$$YDY^TWu_i = \lambda_i u_i \quad \text{for } i = 1,\dots,\ell, \tag{1.35}$$

where $D = \operatorname{diag}(\alpha_1,\dots,\alpha_n) \in \mathbb{R}^{n\times n}$. Inserting $u_i = W^{-1/2}\bar{u}_i$ in (1.35) and multiplying (1.35) by $W^{1/2}$ from the left yield

$$W^{1/2}YDY^TW^{1/2}\bar{u}_i = \lambda_i\bar{u}_i. \tag{1.36a}$$

From (1.34b) we find

$$\langle \bar{u}_i, \bar{u}_j\rangle_{\mathbb{R}^m} = \bar{u}_i^T\bar{u}_j = u_i^TWu_j = \langle u_i, u_j\rangle_W = \delta_{ij}, \quad 1 \leq i,j \leq \ell. \tag{1.36b}$$

Setting $\bar{Y} = W^{1/2}YD^{1/2} \in \mathbb{R}^{m\times n}$ and using $W^T = W$ as well as $D^T = D$ we infer from (1.36) that the solution $\{u_i\}_{i=1}^\ell$ to $(\hat{\mathbf{P}}_W^{n,\ell})$ is given by the symmetric $m \times m$ eigenvalue problem

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

Note that

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

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

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

Then we set (by SVD)

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

compare (1.28).

Note that

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

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

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

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

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

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

To address these two questions we will introduce a *continuous version* of POD. Let $y : [0, T] \to \mathbb{R}^m$ be the unique solution to (1.30). If we are interested to find a POD basis of rank ℓ that describes the whole trajectory $\{y(t) \,|\, t \in [0, T]\} \subset \mathbb{R}^m$ as good as possible we have to consider the following minimization problem

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

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

$$\min_{\tilde{u} \in \mathbb{R}^m} \int_0^T \Big\| y(t) - \langle y(t), \tilde{u} \rangle_W \, \tilde{u} \Big\|_W^2 \, \mathrm{d}t \quad \text{s.t.} \quad \|\tilde{u}\|_W^2 = 1, \tag{1.37}$$

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

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

Thus,

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

we conclude that (1.37) is equivalent with the following maximization problem

$$\max_{\tilde{u} \in \mathbb{R}^m} \int_0^T \big| \langle y(t), \tilde{u} \rangle_W \big|^2 \, \mathrm{d}t \quad \text{s.t.} \quad \|\tilde{u}\|_W^2 = 1. \tag{1.38}$$

The Lagrange functional $\mathcal{L} : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ associated with (1.38) is given by

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

First-order necessary optimality conditions are given by

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

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

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

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

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

which gives

$$\int_0^T \langle y(t),u\rangle_W W y(t)\,\mathrm{d}t = \lambda W u \quad \text{in } \mathbb{R}^m. \tag{1.39}$$

Multiplying (1.39) by W^{-1} from the left yields

$$\int_0^T \langle y(t),u\rangle_W y(t)\,\mathrm{d}t = \lambda u \quad \text{in } \mathbb{R}^m. \tag{1.40}$$

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

$$\mathcal{R}u = \int_0^T \langle y(t),u\rangle_W\, y(t)\,\mathrm{d}t \quad \text{for } u \in \mathbb{R}^m. \tag{1.41}$$

Lemma 1.11. *The operator $\mathcal{R}$ is linear and bounded (i.e., continuous). Moreover,*

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

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

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

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

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

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

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

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

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

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

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

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

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

Utilizing the operator $\mathcal{R}$ we can write (1.40) as the eigenvalue problem

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

It follows from Lemma 1.11 that $\mathcal{R}$ possesses eigenvectors $\{u_i\}_{i=1}^m$ and associated real eigenvalues $\{\lambda_i\}_{i=1}^m$ such that

$$\mathcal{R}u_i = \lambda_i u_i \text{ for } 1 \le i \le m \quad \text{and} \quad \lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_m \ge 0. \tag{1.42}$$

Note that

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

for $i \in \{1,\ldots,m\}$ so that u_1 solves (1.37).

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

Theorem 1.12. *Let $y \in C([0,T];\mathbb{R}^m)$ be the unique solution to (1.30). Then the POD basis of rank ℓ solving the minimization problem $(\hat{\mathbf{P}}_W^\ell)$ is given by the eigenvectors $\{u_i\}_{i=1}^\ell$ of $\mathcal{R}$ corresponding to the ℓ largest eigenvalues $\lambda_1 \ge \ldots \ge \lambda_\ell$.*

Remark 1.13 (Methods of snapshots)**.** Let us introduce the linear and bounded operator $\mathcal{Y} : L^2(0,T) \to \mathbb{R}^m$ by

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

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

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

is given as

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

Then we have

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

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

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

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

$$\mathcal{K}v_i = \lambda_i v_i \text{ for } 1 \le i \le \ell, \quad \lambda_1 \ge \ldots \ge \lambda_\ell > 0, \quad \int_0^T v_i(t)v_j(t)\,\mathrm{d}t = \delta_{ij} \tag{1.43}$$

and set

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

Note that (1.43) is a symmetric eigenvalue problem in the infinite-dimensional function space $L^2(0,T)$. For the functional analytic theory we refer, e.g., to [11]. $\Diamond$

Let us turn back to the optimality conditions (1.35). For any $u \in \mathbb{R}^m$ and $i \in \{1,\ldots,m\}$ we derive

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

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

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

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

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

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

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

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

$$\mathcal{R}^n u_i^n = \lambda_i^n u_i^n, \qquad \lambda_1^n \ge \ldots \lambda_\ell^n \ge \ldots \lambda_{d(n)}^n > \lambda_{d(n)+1}^n = \ldots = \lambda_m^n = 0, \tag{1.44a}$$

$$\mathcal{R}u_i = \lambda_i u_i, \qquad \lambda_1 \ge \ldots \lambda_\ell \ge \ldots \lambda_d > \lambda_{d+1} = \ldots = \lambda_m = 0. \tag{1.44b}$$

Let us note that

$$\int_0^T \|y(t)\|_W^2 \,\mathrm{d}t = \sum_{i=1}^{d} \lambda_i = \sum_{i=1}^{m} \lambda_i. \tag{1.45}$$

In fact,

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

Taking the inner product with u_i, using (1.44b) and summing over i we arrive at

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

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

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

and hence

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

which is (1.45). Analogously, we obtain

$$\sum_{j=1}^{n} \alpha_j \, \|y(t_j)\|_W^2 = \sum_{i=1}^{d(n)} \lambda_i^n = \sum_{i=1}^{m} \lambda_i^n \quad \text{for every } n \in \mathbb{N}. \tag{1.46}$$

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

$$\sum_{j=1}^{n} \alpha_j \, \|y(t_j)\|_W^2 \to \int_0^T \|y(t)\|_W^2 \,\mathrm{d}t \quad \text{as } \Delta t \to 0 \tag{1.47}$$

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

$$\alpha_1 = \frac{\Delta t}{2},\ \alpha_j = \Delta t \text{ for } 2 \le j \le n-1,\ \alpha_n = \frac{\Delta t}{2}. \tag{1.48}$$

Suppose that we have

$$\lim_{n\to\infty} \|\mathcal{R}^n - \mathcal{R}\|_{L(\mathbb{R}^m)} = \lim_{n\to\infty} \sup_{\|u\|_W = 1} \|\mathcal{R}^n u - \mathcal{R}u\|_W = 0 \tag{1.49}$$

provided $y \in C^1([0,T];\mathbb{R}^m)$ is satisfied. In (1.49) $L(\mathbb{R}^m)$ denotes the Banach space of all linear and bounded operators mapping from $\mathbb{R}^m$ into itself. Combining (1.47) with (1.45) and (1.46) we find

$$\sum_{i=1}^{m} \lambda_i^n \to \sum_{i=1}^{m} \lambda_i \quad \text{as } n \to \infty. \tag{1.50}$$

Now choose and fix

$$\ell \quad \text{such that} \quad \lambda_\ell \neq \lambda_{\ell+1}. \tag{1.51}$$

Then by spectral analysis of compact operators ([5, pp. 212–214]) and (1.49) it follows that

$$\lambda_i^n \to \lambda_i \quad \text{for } 1 \le i \le \ell \text{ as } n \to \infty. \tag{1.52}$$

Combining (1.50) and (1.52) there exists $\bar{n} \in \mathbb{N}$ such that

$$\sum_{i=\ell+1}^{m} \lambda_i^n \leq 2 \sum_{i=\ell+1}^{m} \lambda_i \quad \text{for all } n \geq \bar{n}, \tag{1.53}$$

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

$$\sum_{i=\ell+1}^{d(n)} \left|\langle y_0, u_i^n\rangle_W\right|^2 \leq 2 \sum_{i=\ell+1}^{m} \left|\langle y_0, u_i\rangle_W\right|^2 \quad \text{for all } n \geq \bar{n}, \tag{1.54}$$

provided that $\sum_{i=\ell+1}^{m} |\langle y_0, u_i\rangle_W|^2 \neq 0$ (1.49) hold. Recall that the vector $y_0 \in \mathbb{R}^m$ stands for the initial condition in (1.30b). Then we have

$$\|y_0\|_W^2 = \sum_{i=1}^{m} \left|\langle y_0, u_i\rangle_W\right|^2. \tag{1.55}$$

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

$$\|y_0\|_W^2 = \sum_{i=1}^{d(n)} \left|\langle y_0, u_i^n\rangle_W\right|^2. \tag{1.56}$$

Therefore, for $\ell < d(n)$ by (1.55) and (1.56)

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

As a consequence of (1.49) and (1.51) we have $\lim_{n\to\infty} \|u_i^n - u_i\|_W = 0$ for $i = 1, \ldots, \ell$ and hence (1.54) follows.

Summarizing we have the following theorem.

Theorem 1.14. *Assume that $y \in C^1([0,T];\mathbb{R}^m)$ is the unique solution to* (1.30). *Let $\{(u_i^n, \lambda_i^n)\}_{i=1}^m$ and $\{(u_i, \lambda_i)\}_{i=1}^m$ be the eigenvector-eigenvalue pairs given by* (1.44). *Suppose that $\ell \in \{1, \ldots, m\}$ is fixed such that* (1.51) *and*

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

hold. Then we have

$$\lim_{n\to\infty} \|\mathcal{R}^n - \mathcal{R}\|_{L(\mathbb{R}^m)} = 0. \tag{1.57}$$

This implies

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

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

Proof. We only have to verify (1.57). For that purpose we choose an arbitrary $u \in \mathbb{R}^m$ with $\|u\|_W = 1$ and introduce $f_u : [0,T] \to \mathbb{R}^m$ by

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

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

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

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

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

Hence,

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

Consequently,

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

which is (1.57). □

1.4. Exercises.

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

1.2) Verify the claim in Theorem 1.1 that $\operatorname{argmax}(\mathbf{P}^\ell) = \sum_{i=1}^{\ell} \sigma_i^2$ holds true.

1.3) 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 $U^d \in \mathbb{R}^{m\times d}$ is a matrix with pairwise orthonormal vectors $u_i \in \mathbb{R}^m$, $1 \le i \le d$. Prove that

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

1.4) 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 $W^\alpha = Q\operatorname{diag}(\eta_1^\alpha, \ldots, \eta_m^\alpha)Q^T$ be the eigenvalue decomposition of W. We define

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

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

1.5) Verify the claims of Theorem 1.8.

1.5.1) Prove that $u_i = W^{-1/2}\bar{u}_i$, $1 \le i \le \ell$, solves $(\mathbf{P}_W^\ell)$, where the matrix W and the vectors $\bar{u}_1, \ldots, \bar{u}_m$ are introduced in Theorem 1.8.

1.5.2) Show that (1.29) holds.

1.6) Prove that u_1 given by (1.42) is a global solution to (1.37).

1.7) Verify (1.46).

2. Reduced-order modeling (ROM)

In Section 1 we have introduced the POD basis of rank ℓ in $\mathbb{R}^m$ and discussed its application to initial-value problems. If the POD basis is computed, it can be used to derive a so-called *low-dimensional approximation* or a *reduced-order model* for (1.30). This is the focus of this section.

2.1. ROM for time-dependent systems. Suppose that we have determined a POD basis $\{u_j\}_{j=1}^\ell$ of rank $\ell \in \{1, \ldots, m\}$ in $\mathbb{R}^m$. Then we make the ansatz

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

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

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

holds, $y^\ell(t)$ is an approximation for $y(t)$ provided $\ell < m$. Inserting (2.1) into (1.30) yields

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

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

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

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

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

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

the vector-valued mapping

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

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

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

Then, (2.3) can be expressed as

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

From (2.2b) we derive

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

where

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

holds. System (2.4) is called the *POD-Galerkin projection* for (1.30). In case of $\ell \ll m$ the ℓ-dimensional system (2.4) is a low-dimensional approximation for (1.30). Therefore, (2.4) is a reduced-order model for (1.30).

2.2. Error analysis for the reduced-order model. In this section we focus on error analysis for POD Galerkin approximations. For a more detailed presentation we refer the reader to [7, 8, 9] and [4].

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 (1.30) and $\{u_i\}_{i=1}^\ell$ the POD basis of rank ℓ solving

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

The reduced-order model for (1.30) is given by (2.4). We are interested in estimating the error

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

Let us introduce the finite-dimensional space

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

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

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

Then,

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

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

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

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

Throughout we shall use the decomposition

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

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

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

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

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

Next we estimate the term $\vartheta^\ell(t)$. Utilizing (1.30a) and (2.4) we obtain for every $u^\ell \in V^\ell$ and $t \in (0, T]$

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

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

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

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

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

holds. Then, we infer from (2.9)

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

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

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

Moreover, we have

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

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

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

Consequently,

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

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

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

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

Theorem 2.1. *Let $y \in C([0,T];\mathbb{R}^m) \cap C^1(0,T;\mathbb{R}^m)$ be the unique solution to (1.30), $\ell \in \{1,\ldots,m\}$ be fixed and $\{u_i\}_{i=1}^{\ell}$ a POD basis of rank ℓ solving (2.5). Let y^ℓ be the unique solution to the reduced-order model (2.4). Then*

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

for a constant $C > 0$.

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

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

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

Remark 2.2. The term

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

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

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

subject to

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

we end up with the estimate

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

for a constant $\tilde{C} > 0$. In this case the time derivatives are also included in the snapshot ensemble. Of course, the operator $\mathcal{R}$ defined in (1.41) has to be replaced. It turns out that the POD basis $\{u_i\}_{i=1}^{\ell}$ is given by the eigenvalue problem

$$\tilde{\mathcal{R}} \tilde{u}_i = \tilde{\lambda}_i \tilde{u}_i \text{ for } 1 \le i \le m \quad \text{and} \quad \tilde{\lambda}_1 \ge \tilde{\lambda}_2 \ge \ldots \ge \tilde{\lambda}_m \ge 0 \tag{2.13}$$

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

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

for $u \in \mathbb{R}^m$. ◊

Remark 2.3. 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 2.2 is given by

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

with non-negative weights introduced in $(\hat{\mathbf{P}}_W^{n,\ell})$ and the diagonal matrix $D = \operatorname{diag}(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^{n\times n}$. Thus, we have $\tilde{\mathcal{R}} = Y\tilde{D}Y^T W \in \mathbb{R}^{m\times m}$, which is of the same form as in (1.35). The discrete version to (2.13) is

$$Y\tilde{D}Y^T W \tilde{u}_i = \tilde{\lambda}_i \tilde{u}_i \text{ for } 1 \le i \le m \quad \text{and} \quad \tilde{\lambda}_1 \ge \tilde{\lambda}_2 \ge \ldots \ge \tilde{\lambda}_m \ge 0 \tag{2.14}$$

Setting $\tilde{u}_i = W^{-1/2}\bar{u}_i$ in (2.14) and multiplying by $W^{1/2}$ from the left yield

$$W^{1/2} Y \tilde{D} Y^T W^{1/2} \bar{u}_i = \lambda_i \bar{u}_i. \tag{2.15}$$

Let $\bar{Y} = W^{1/2} Y \tilde{D}^{1/2} \in \mathbb{R}^{m\times 2n}$. Using $W^T = W$ as well as $\tilde{D}^T = \tilde{D}$ we infer from (2.15) that the solution $\{\tilde{u}_i\}_{i=1}^{\ell}$ is given by the symmetric $m \times m$ eigenvalue problem

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

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

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

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

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

Then we set (by SVD)

$$\tilde{u}_i = W^{-1/2}\bar{u}_i = \frac{1}{\sqrt{\lambda_i}} W^{-1/2} \bar{Y} \bar{v}_i = \frac{1}{\sqrt{\lambda_i}} Y \tilde{D}^{1/2} \bar{v}_i$$

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, non-equidistant meshes can be treated analogously [8]. We compute a POD basis $\{u_i^n\}_{i=1}^{\ell}$ of rank ℓ by solving the constrained minimization problem $(\hat{\mathbf{P}}_W^{n,\ell})$. After the POD basis has been determined,

we derive the reduced-order model as described in Section 2.1. Thus,

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

solves the POD Galerkin projection of (1.30)

$$\langle \dot{y}^\ell(t), u_i^n \rangle_W = \langle A y^\ell(t) + f(t, y^\ell(t)), u_i^n \rangle_W \quad \text{for } i = 1 \dots, \ell \text{ and } t \in (0,T], \tag{2.16a}$$

$$\langle y^\ell(0), u_i^n \rangle_W = \langle y_0, u_i^n \rangle_W \quad \text{for } i = 1 \dots, \ell. \tag{2.16b}$$

To solve (2.16) we apply the implicit Euler method. By Y_j we denote an approximation for y^ℓ at the time t_j, $1 \le j \le n$. Then, the discrete system for the sequence $\{Y_j\}_{j=1}^n$ in $V_n^\ell = \operatorname{span}\{u_1^n, \dots, u_\ell^n\}$ looks like

$$\left\langle \frac{Y_j - Y_{j-1}}{\Delta t}, u_i^n \right\rangle_W = \langle A Y_j + f(t, Y_j), u_i^n \rangle_W \quad \text{for } i = 1 \dots, \ell,\ 2 \le j \le n, \tag{2.17a}$$

$$\langle Y_1, u_i^n \rangle_W = \langle y_0, u_i^n \rangle_W \quad \text{for } i = 1 \dots, \ell. \tag{2.17b}$$

We are interested in estimating

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

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

$$\mathcal{P}_n^\ell = \sum_{i=1}^{\ell} \langle u, u_i^n \rangle_W \, u_i^n \quad \text{for } u \in \mathbb{R}^m. \tag{2.18}$$

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

We shall make use of the decomposition

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

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

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

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

$$\sum_{j=1}^{n} \alpha_j \, \|\varrho_j^\ell\|_W^2 = \sum_{i=\ell+1}^{m} \lambda_i^n. \tag{2.19}$$

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 (1.30a) and (2.17a)

$$\begin{aligned}
\langle \overline{\partial}\vartheta_j^\ell, u_i\rangle &= \left\langle \mathcal{P}_n^\ell\left(\frac{y(t_j)-y(t_{j-1})}{\Delta t}\right) - \frac{Y_j - Y_{j-1}}{\Delta t}, u_i^n\right\rangle_W \\
&= \langle \dot y(t_j) - (AY_j + f(t_j, Y_j))), u_i^n\rangle_W \\
&\quad + \left\langle \mathcal{P}_n^\ell\left(\frac{y(t_j)-y(t_{j-1})}{\Delta t}\right) - \dot y(t_j), u_i^n\right\rangle_W \\
&= \langle A(y(t_j)-Y_j) + f(t_j, y(t_j)) - f(t_j, Y_j), u_i^n\rangle_W \\
&\quad + \left\langle \mathcal{P}_n^\ell\left(\frac{y(t_j)-y(t_{j-1})}{\Delta t}\right) - \frac{y(t_j)-y(t_{j-1})}{\Delta t}, u_i^n\right\rangle_W \\
&\quad + \left\langle \frac{y(t_j)-y(t_{j-1})}{\Delta t} - \dot y(t_j), u_i^n\right\rangle_W \\
&= \langle A(y(t_j)-Y_j) + f(t_j, y(t_j)) - f(t_j, Y_j) + z_j^\ell + w_j^\ell, u_i^n\rangle_W
\end{aligned} \tag{2.20}$$

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

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

Multiplying (2.20) by $\langle \vartheta_j^\ell, u_i^n\rangle_W$ and adding all ℓ equations we arrive at

$$\langle \overline{\partial}\vartheta_j^\ell, \vartheta_j^\ell\rangle = \langle A(y(t_j)-Y_j) + f(t_j, y(t_j)) - f(t_j, Y_j) + z_j^\ell + w_j^\ell, \vartheta_j^\ell\rangle_W \tag{2.21}$$

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

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

for all $u, \tilde u \in \mathbb{R}^m$. Choosing $u = \vartheta_j^\ell$ and $\tilde u = \vartheta_{j-1}^\ell$ we infer from (2.21)

$$2\langle \overline{\partial}\vartheta_j^\ell, \vartheta_j^\ell\rangle = \frac{1}{\Delta t}\left(\|\vartheta_j^\ell\|_W^2 - \|\vartheta_{j-1}^\ell\|_W^2 + \|\vartheta_j^\ell - \vartheta_{j-1}^\ell\|_W^2\right). \tag{2.22}$$

Inserting (2.22) into (2.21) and using the Cauchy-Schwarz inequality we obtain

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

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

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

Hence, by Young's inequality we find

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

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

$$0 < \Delta t \le \frac{1}{2c_2} \tag{2.23}$$

holds. With (2.23) holding we have

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

Thus,

$$\frac{1}{1 - c_2\Delta t} = \frac{1 - c_2\Delta t + c_2\Delta t}{1 - c_2\Delta t} = 1 + \frac{c_2\Delta t}{1 - c_2\Delta t}. \le 1 + 2c_2\Delta t \tag{2.24}$$

Using (2.24) we infer that

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

Summation on j yields

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

Note that

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

Thus,

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

We next estimate the term involving w_k^ℓ:

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

The term z_k^ℓ can be estimated as follows:

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

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

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

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

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

where $c_3 = e^{2c_2 T} \max\{7/3, 2c_1, 8\}$ is independent of ℓ and $\{t_j\}_{j=1}^n$. From $\sum_{k=1}^n \alpha_k = T$ and (2.19) we infer

$$\begin{aligned}
\sum_{j=1}^{n} \alpha_j \|\vartheta_j^\ell\|_W^2 \quad &\le c_3 T \Bigg(\sum_{j=1}^{n} \alpha_j \left(\|\mathcal{P}_n^\ell \dot{y}(t_j) - \dot{y}(t_j)\|_W^2 + \|\varrho_j^\ell\|_W^2 \right) \\
&\qquad + (\Delta t)^2 \|\ddot{y}\|_{L^2(0,T;\mathbb{R}^m)}^2 \Bigg) \\
&\le c_4 \Bigg(\sum_{i=\ell+1}^{m} \left(\lambda_i^n + \sum_{j=1}^{n} \alpha_j \left| \langle \dot{y}(t_j), u_i^n \rangle_W \right|^2 \right) + (\Delta t)^2 \Bigg)
\end{aligned} \tag{2.25}$$

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

Theorem 2.4. *Let $y \in C([0,T];\mathbb{R}^m) \cap C^1(0,T;\mathbb{R}^m)$ be the unique solution to (1.30) satisfying $\ddot{y} \in L^2(0,T;\mathbb{R}^m)$ and $\ell \in \{1, \dots, m\}$ be fixed. Suppose that $\{u_i^n\}_{i=1}^\ell$ is a POD basis of rank ℓ solving $(\hat{\mathbf{P}}_W^{n,\ell})$. Assume that (2.17) possesses a unique solution $\{Y_j\}_{j=1}^n$. Then there exists a constant $C > 0$ such that*

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

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

Proof. The claim follows directly from (2.19), (2.25), and

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

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

Remark 2.5. Compared to the estimate in Theorem 2.1 we observe the term

$$\sum_{j=1}^n \alpha_j \left|\langle \dot{y}(t_j), u_i^n\rangle_W\right|^2 \tag{2.26}$$

instead of the term

$$\int_0^T \left|\langle \dot{y}(t), u_i\rangle_W\right|^2 \mathrm{d}t. \tag{2.27}$$

Note that (2.26) is the trapezoidal approximation of (2.27). Furthermore, the error $O((\Delta t)^2)$ appears in the estimate of Theorem 2.4 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 1.14 that there exists a number $\bar{n} \in \mathbb{N}$ satisfying

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

for $n \geq \bar{n}$ provided $\sum_{i=\ell+1}^m \lambda_i \neq 0$ and $\int_0^T \left|\langle \dot{y}(t), u_i\rangle_W\right|^2 \mathrm{d}t \neq 0$ hold. Thus, we infer from Theorems 2.1 and 2.4 the following result.

Theorem 2.6. *Let all hypothesis of Theorems* 1.14, 2.1 *and* 2.4 *be satisfied. If* $\int_0^T \left|\langle \dot{y}(t), u_i\rangle_W\right|^2 \mathrm{d}t \neq 0$, *then there exists a constant* $C > 0$ *and a number* $\bar{n} \in \mathbb{N}$ *such that*

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

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

2.3. Exercises.

2.1) Prove the *Gronwall lemma*: For $T > 0$ let $\eta : [0,T] \to \mathbb{R}$ be a non-negative, differentiable function satisfying

$$\eta'(t) \le \varphi(t)\eta(t) + \psi(t) \quad \text{for all } t \in [0,T],$$

where φ and ψ are real-valued, non-negative, integrable functions on $[0,T]$. Then

$$\eta(t) \le \exp\left(\int_0^t \varphi(s)\,\mathrm{d}s\right)\left(\eta(0) + \int_0^t \psi(s)\,\mathrm{d}s\right) \quad \text{for all } t \in [0,T].$$

In particular, if

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

show that $\eta = 0$ holds in $[0,T]$.

2.2) Show that the operator $\mathcal{P}_n^\ell$ defined in (2.18) is linear, bounded and satisfies $\|\mathcal{P}_n^\ell\|_{L(\mathbb{R}^m)} = 1$.

2.3) Prove that the first-order necessary optimality condition for (2.12) is given by $\tilde{\mathcal{R}}\tilde{u}_i = \tilde{\lambda}_i\tilde{u}_i$, $1 \le i \le \ell$.

2.4) Show that $\tilde{\mathcal{R}}$ is linear, bounded, self-adjoint and non-negative 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.

3. 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 [2].

3.1. The 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

$$J(x,u) = \int_0^T x(t)^T Q x(t) + u(t)^T R u(t)\,\mathrm{d}t + x(T)^T M x(T), \tag{3.1a}$$

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

$$\dot{x}(t) = Ax(t) + Bu(t) \text{ for } t \in (0,T] \quad \text{and} \quad x(0) = x_0. \tag{3.1b}$$

In (3.1a) 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 (3.1b) 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 (3.1) is called the *linear-quadratic regulator problem* (*LQR problem*).

3.2. 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

$$J_t(x, u) = \int_t^T L(x(s), u(s), s)\,\mathrm{d}s + g(x(T)) \tag{3.2a}$$

is minimized subject to the non-linear system dynamics

$$\dot{x}(s) = F(x(s), u(s), s) \text{ for } s \in (0, T] \quad \text{and} \quad x(t) = x_t. \tag{3.2b}$$

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 [1].

Theorem 3.1 (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 3.1 by contradiction. Suppose that there exists a control u^{**} so that

$$\begin{aligned} \int_{t+\Delta t}^T L(x^{**}(s), u^{**}(s), s)\,\mathrm{d}s + g(x^{**}(T)) \\ < \int_{t+\Delta t}^T L(x^*(s), u^*(s), s)\,\mathrm{d}s + g(x^*(T)), \end{aligned} \tag{3.3}$$

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

$$u(s) = \begin{cases} u^*(s) & \text{if } s \in [t, t + \Delta t], \\ u^{**}(s) & \text{if } s \in (t + \Delta t, T]. \end{cases} \tag{3.4}$$

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 (3.3) and (3.4) that

$$\begin{aligned}
&\int_t^T L(x(s), u(s), s)\,\mathrm{d}s + g(x(T)) \\
&= \int_t^{t+\Delta t} L(x^*(s), u^*(s), s)\,\mathrm{d}s + \int_{t+\Delta t}^T L(x^{**}(s), u^{**}(s), s)\,\mathrm{d}s + g(x^{**}(T)) \\
&< \int_t^{t+\Delta t} L(x^*(s), u^*(s), s)\,\mathrm{d}s + \int_{t+\Delta t}^T L(x^*(s), u^*(s), s)\,\mathrm{d}s + g(x^*(T)) \\
&= \int_t^T L(x^*(s), u^*(s), s)\,\mathrm{d}s + g(x^*(T)).
\end{aligned} \tag{3.5}$$

Recall that $u^*(s)$ is optimal for $s \in [t, T]$ by assumption. From (3.5) it follows that the control u given by (3.4) yields a smaller value of the cost functional. This is a contradiction. □

Next we derive the Hamilton-Jacobi-Bellman equation for (3.2). Let $V^* : \mathbb{R}^{m_x} \times [0, T] \to \mathbb{R}$ denote the minimal value function given by

$$\begin{aligned}
&V^*(x_t, t) \\
&= \min_{u:[t,T]\to\mathbb{R}^{m_u}} \Big\{ J_t(x, u) \,\Big|\, \dot{x}(s) = F(x(s), u(s), s),\ s \in (t, T] \text{ and } x(t) = x_t \Big\}
\end{aligned} \tag{3.6}$$

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 (3.6) we conclude

$$\begin{aligned}
&V^*(x_t, t) \\
&= \min_{u:[t,t+\Delta t]\to\mathbb{R}^{m_u}} \Big\{ \int_t^{t+\Delta t} L(x(s), u(s), s)\,\mathrm{d}s + V^*(x(t+\Delta t), t+\Delta t) \,\Big|\, \\
&\qquad\qquad \dot{x}(s) = F(x(s), u(s), s),\ s \in (t, t+\Delta t] \text{ and } x(t) = x_t \Big\}
\end{aligned} \tag{3.7}$$

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 (3.7) 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

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

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

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

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

$$-\frac{\partial V^*}{\partial t}(x_t,t) = \min_{u_t\in\mathbb{R}^{m_u}} \big\{ L(x_t,u_t,t) + \nabla V^*(x_t,t)^T F(x_t,u_t,t)\big\} \tag{3.8a}$$

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

$$V^*(x_t,T) = g(x_t) \tag{3.8b}$$

for all $x_t\in\mathbb{R}^{m_x}$.

To solve (3.8) we proceed in two steps. First we compute a solution u_t to

$$u^*(t) = \operatorname*{argmin}_{u_t\in\mathbb{R}^{m_u}} \big\{ L(x_t,u_t,t) + \nabla V^*(x_t,t)^T F(x_t,u_t,t)\big\}$$

and set

$$\Psi(\nabla V^*(x_t,t),x_t,t) = u^*(t), \tag{3.9}$$

which gives us a control law. Then we insert (3.9) into (3.8a) and solve

$$\begin{aligned} -\frac{\partial V^*}{\partial t}(x_t,t) &= L(x_t,\Psi(\nabla V^*(x_t,t),x_t,t),t) \\ &\qquad + \nabla V^*(x_t,t)^T F(x_t,\Psi(\nabla V^*(x_t,t),x_t,t),t) \end{aligned}$$

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 3.2.
1) In general, it is not possible to solve (3.8) 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. ◊

3.3. 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

$$u^* = -\frac{1}{2} R^{-1} B^T \nabla V^*(x,t). \tag{3.10}$$

For the minimal value function V^* we make the quadratic ansatz

$$V^*(x,t) = x^T P(t) x, \quad P(t) \in \mathbb{R}^{m_x \times m_x} \text{ symmetric.} \tag{3.11}$$

Then, we have $\nabla V^*(x,t) = 2P(t)x$ so that

$$u^* = -R^{-1} B^T P(t) x.$$

Note that

$$\begin{aligned}
\frac{\partial V^*}{\partial t}(x_t,t) &= x_t^T \dot{P}(t) x_t, \\
L(x_t, -R^{-1}B^T P(t) x_t, t) &= x_t^T Q x_t + x_t^T P(t) B R^{-1} B^T P(t) x_t \\
&= x_t^T \big(Q + P(t) B R^{-1} B^T P(t)\big) x_t, \\
F(x_t, -R^{-1}B^T P(t) x_t, t) &= Ax_t - BR^{-1}B^T P(t) x_t = \big(A - BR^{-1}B^T P(t)\big) x_t, \\
\nabla V^*(x_t,t) &= 2P(t) x_t.
\end{aligned}$$

Consequently,

$$\begin{aligned}
-x_t^T \dot{P}(t) x_t &= -\frac{\partial V^*}{\partial t}(x_t,t) \\
&= x_t^T \big(Q + P(t) B R^{-1} B^T P(t)\big) x_t + \big(2P(t)x_t\big)^T \big(A - BR^{-1}B^T P(t)\big) x_t
\end{aligned}$$

for all $x_t \in \mathbb{R}^{m_x}$, which yields

$$\begin{aligned}
&-x_t^T \dot{P}(t) x_t \\
&\quad = x_t^T \big(Q + P(t) B R^{-1} B^T P(t) + 2P(t)A - 2P(t) B R^{-1} B^T P(t)\big) x_t \\
&\quad = x_t^T \big(2P(t)A + Q - P(t) B R^{-1} B^T P(t)\big) x_t
\end{aligned}$$

for all $x_t \in \mathbb{R}^{m_x}$. From $P(t) = P(t)^T$ we deduce that

$$2x_t^T P(t) A x_t = x_t^T P(t) A x_t + x_t^T A^T P(t) x_t = x_t^T \big(A^T P(t) + P(t) A\big) x_t.$$

Using $V^*(x_t,T) = x_t^T P(T) x_t$ and (3.8b) we get

$$-x_t^T \dot{P}(t) x_t = x_t^T \big(A^T P(t) + P(t)A + Q - P(t)BR^{-1}B^T P(t)\big) x_t, \quad t \in [0,T) \tag{3.12a}$$

$$x_t^T P(T) x_t = x_t^T M x_t. \tag{3.12b}$$

Since (3.12) holds for all $x_t \in \mathbb{R}^{m_x}$ we obtain the following *matrix Riccati equation*

$$-\dot{P}(t) = A^T P(t) + P(t)A + Q - P(t)BR^{-1}B^T P(t), \quad t \in [0,T) \tag{3.13a}$$

$$P(T) = M. \tag{3.13b}$$

Finally, the optimal state-feedback is given by

$$u^*(t) = -K(t)x(t) \quad \text{and} \quad K(t) = R^{-1}B^T P(t).$$

Example 3.3. 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 $u^*(t) = -P(t)x(t)$. ◊

3.4. Exercises. Let us consider the one-dimensional heat equation

$$\theta_t(t,x) = \theta_{xx}(t,x) + u(t)\chi(x) \quad \text{for all } (t,x) \in Q = (0,T) \times \Omega, \tag{3.14a}$$

$$\theta_x(t,0) = \theta_x(t,1) = 0 \quad \text{for all } t \in (0,T), \tag{3.14b}$$

$$\theta(0,x) = \theta_0(x) \quad \text{for all } x \in \Omega = (0,1) \subset \mathbb{R}, \tag{3.14c}$$

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.

3.1) Apply a classical finite difference approximation for the spatial variable x (compare Example 1.10) and derive the finite-dimensional initial value problem for the finite difference approximations.

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.

3.3) Formulate the matrix Riccati equation for the discretized quadratic cost functional — see part 3.2) — and the discretized heat equation — see part 3.1).

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?

4. Balanced truncation

Let us consider the linear time-invariant system

$$\dot{x}(t) = Ax(t) + Bu(t) \text{ for } t \in (0,\infty) \quad \text{and} \quad x(0) = x_0, \tag{4.1a}$$

$$y(t) = Cx(t) \qquad \text{for } t \in [0,\infty) \tag{4.1b}$$

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

It is helpful to analyze the linear system (4.1) through the Laplace transform.

Definition 4.1. *Let $f(t)$ be a time-varying vector. Then its* Laplace transform *is defined by*

$$\mathcal{L}[f](s) = \int_0^\infty e^{-st} f(t)\,\mathrm{d}t \quad \textit{for } s \in \mathbb{R}. \tag{4.2}$$

The Laplace transform is defined for those values of s, for which (4.2) *converges.*

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

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

where we have used (4.1b). Note that

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

Therefore, the Laplace transform of the dynamical system (4.1a) yields

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

which gives

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

Thus,

$$\mathcal{L}[y](s) = C\mathcal{L}[x](s) = C(sI - A)^{-1}x(0) + C(sI - A)^{-1}B\mathcal{L}[u](s). \tag{4.3}$$

For $x(0) = 0$ the expression (4.3) reduces to

$$\mathcal{L}[y](s) = G(s)\mathcal{L}[u](s) \tag{4.4}$$

where

$$G(s) = C(sI - A)^{-1}B \tag{4.5}$$

is called the *transfer matrix* of the system.

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

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

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

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

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

Definition 4.2. *The dynamical system* (4.1a) *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* (4.1a) *satisfies* $x(T) = x_T$. *Otherwise,* (A,B) *is said to be* uncontrollable.

Controllability can be verified as stated in the next theorem. For a proof we refer to [14].

Theorem 4.3. *The following claims are equivalent:*

1) (A,B) *are controllable.*
2) *The controllability gramian*

$$W_c(t) = \int_0^t e^{sA}BB^Te^{sA^T}\,\mathrm{d}s$$

is positive definite for every $t > 0$.

3) *The controllability matrix*

$$\mathcal{C} = \begin{bmatrix} B & AB & A^2B & \dots & A^{m_x-1}B \end{bmatrix} \in \mathbb{R}^{m_x \times (m_x m_u)}$$

has full rank.

Definition 4.4. 1) *The unforced system* $\dot{x}(t) = Ax(t)$ *is called* stable, *if the eigenvalues of* A *are in the open left half plane, i.e.,* $\Re\mathrm{e}\lambda < 0$ *for every eigenvalue* λ . *A matrix with this property is said to be* stable *or* Hurwitz.

2) *The dynamical system* (4.1a) *or* (A,B) *are called* stabilizable *if there exists a state-feedback* $u(t) = -Kx(t)$ *so that* $A - BK$ *is stable.*

The next result, which is proved in [14], is a consequence of Theorem 4.3.

Theorem 4.5. *The following claims are equivalent:*

1) (A,B) *are stabilizable.*
2) *The matrix* $[A - \lambda I \;\; B] \in \mathbb{R}^{m_x \times (m_x+m_u)}$ *has full row rank for all* $\lambda \in \mathbb{C}$ *with a negative real part, i.e.,* $\Re\mathrm{e}\lambda < 0$.

Let us now consider the dual notions of observability.

Definition 4.6. *The dynamical system* (4.1) *or* (A,C) *are called* observable *if for any* $t_1 \in (0,T]$, *the initial condition* $x_0 \in \mathbb{R}^{m_x}$ *can be determined from the time history of the input* $u(t)$ *and the output* $y(t)$ *in the interval* $[0,t_1] \subset [0,T]$. *Otherwise, the system or* (A,C) *is said to be* unobservable.

For a proof of the next theorem we refer the reader to [14].

Theorem 4.7. *The following claims are equivalent:*

1) (A,C) *is observable.*
2) *The observability gramian*

$$W_o(t) = \int_0^t e^{sA^T}C^TCe^{sA}\,\mathrm{d}s$$

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*

$$AW_c + W_c A^T + BB^T = 0 \in \mathbb{R}^{n_x \times n_x}, \tag{4.6a}$$

$$A^T W_o + W_o A + C^T C = 0 \in \mathbb{R}^{n_x \times n_x}. \tag{4.6b}$$

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

$$x = \mathcal{T} z \tag{4.7}$$

where $\mathcal{T} \in \mathbb{R}^{n_x \times n_x}$ is a regular matrix. Then we obtain instead of (4.1) the system

$$\dot{z}(t) = \tilde{A} z(t) + \tilde{B} u(t) \text{ for } t \in (0,\infty) \quad \text{and} \quad z(0) = z_0, \tag{4.8a}$$

$$y(t) = \tilde{C} z(t) \qquad \text{for } t \in [0,\infty) \tag{4.8b}$$

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 (4.6a). The controllability gramian $\tilde{W}_c$ for (4.8) satisfies

$$\tilde{A}\tilde{W}_c + \tilde{W}_c \tilde{A}^T + \tilde{B}\tilde{B}^T = 0$$

i.e.,

$$\mathcal{T}^{-1} A \mathcal{T} \tilde{W}_c + \tilde{W}_c \mathcal{T}^T A^T \mathcal{T}^{-T} + \mathcal{T}^{-1} B B^T \mathcal{T}^{-T} = 0. \tag{4.9}$$

Multiplying (4.9) by $\mathcal{T}$ from the left and by $\mathcal{T}^T$ from the right yields

$$A\mathcal{T}\tilde{W}_c\mathcal{T}^T + \mathcal{T}\tilde{W}_c\mathcal{T}^T A^T + BB^T = 0. \tag{4.10}$$

From (4.6a) and (4.10) we infer that $W_c = \mathcal{T}\tilde{W}_c\mathcal{T}^T$ holds. Thus, the coordinate transformation (4.7) 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 (4.6b). The observability gramian $\tilde{W}_o$ for (4.8) satisfies

$$\tilde{A}^T\tilde{W}_o + \tilde{W}_o\tilde{A} + \tilde{C}^T\tilde{C} = 0$$

i.e.,

$$\mathcal{T}^T A^T \mathcal{T}^{-T}\tilde{W}_o + \tilde{W}_o\mathcal{T}^{-1}A\mathcal{T} + \mathcal{T}^T C^T C\mathcal{T} = 0. \tag{4.11}$$

Multiplying (4.9) by $\mathcal{T}^{-T}$ from the left and by $\mathcal{T}^{-1}$ from the right yields

$$A^T\mathcal{T}^{-T}\tilde{W}_o\mathcal{T}^{-1} + \mathcal{T}^{-T}\tilde{W}_o\mathcal{T}^{-1}A + C^T C = 0. \tag{4.12}$$

From (4.6b) and (4.12) we infer that $W_o = \mathcal{T}^{-T}\tilde{W}_o\mathcal{T}^{-1}$ holds. Thus, the coordinate transformation (4.7) 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

$$\mathcal{T}^{-1}W_c\mathcal{T}^{-T} = \mathcal{T}^T W_o\mathcal{T} = \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_{m_x}). \tag{4.13}$$

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 (4.13) 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 (4.8) we only consider the system for the first $\ell \in \{1, \ldots, m_x\}$ components of z:

$$\dot{z}_\ell(t) = \tilde{A}_\ell z_\ell(t) + \tilde{B}_\ell u(t) \text{ for } t \in (0,\infty) \quad \text{and} \quad z_\ell(0) = z_{0\ell}, \tag{4.14a}$$

$$y_\ell(t) = \tilde{C}_\ell z_\ell(t) \qquad \text{for } t \in [0,\infty), \tag{4.14b}$$

where

$$\tilde{A} = \left(\begin{array}{c|c} \tilde{A}_\ell & * \\ \hline * & * \end{array}\right), \quad \tilde{B} = \left(\begin{array}{c} \tilde{B}_\ell \\ \hline * \end{array}\right), \quad \tilde{C} = \left(\begin{array}{c|c} \tilde{C}_\ell & * \end{array}\right), \quad z_{0\ell} = \left(\begin{array}{c} \tilde{z}_{0\ell} \\ \hline * \end{array}\right),$$

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 (4.1) 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 (4.14). 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

$$\|G - G_\ell\| < 2 \sum_{i=\ell+1}^{m_x} \sigma_i.$$

Acknowledgement. The author wants to thank B. Gotthardt, M. Kahlbacher, M. Kanitsar, and M. Müller for their careful reading of the scriptum and their comments which improve the scriptum significantly.

References

[1] R.E. Bellman. The theory of dynamic programming. *Proc. Nat. Acad. Sci.*, USA, 38:716-719, 1952.

[2] P. Dorato, C. Abdallah, and V. Cerone. *Linear-Quadratic Control. An Introduction*. Prentice Hall, Englewood Cliffs, New Jersey 07632, 1995.

[3] P. Holmes, J.L. Lumley, and G. Berkooz. *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge Monographs on Mechanics, Cambridge University Press, 1996.

[4] 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.

[5] T. Kato. *Perturbation Theory for Linear Operators*. Springer-Verlag, Berlin, 1980.

[6] K. Kunisch and S. Volkwein. Control of Burgers' equation by a reduced order approach using proper orthogonal decomposition. *Journal on Optimization Theory and Applications*, 102, 345–371, 1999.

[7] K. Kunisch and S. Volkwein. Galerkin proper orthogonal decomposition methods for parabolic problems. *Numerische Mathematik*, 90:117–148, 2001.

[8] 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.

[9] 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.

[10] B. Noble. *Applied Linear Algebra*. Englewood Cliffs, NJ : Prentice-Hall, 1969.

[11] M. Reed and B. Simon. *Methods of Modern Mathematical Physics I: Functional Analysis*. Academic Press, New York, 1980.

[12] C.W. Rowley. Model reduction for fluids, using balanced proper orthogonal decomposition. *Int. J. on Bifurcation and Chaos*, 15:997-1013, 2005.

[13] L. Sirovich. Turbulence and the dynamics of coherent structures, parts I-III. *Quarterly of Applied Mathematicss*, XLV:561-590, 1987.

[14] K. Zhou, J.C. Doyle, and K. Glover. *Robust and Optimal Control*. Prentice Hall, Upper Saddle River, New Jersey, 07458, 1996.

S. Volkwein, Karl-Franzens-Universität Graz, Institut für Mathematik und Wissenschaftliches Rechnen, Heinrichstrasse 36, A-8010 Graz, Austria

E-mail address: stefan.volkwein@uni-graz.at