MODEL ORDER REDUCTION VIA BALANCED TRUNCATION AND PROPER ORTHOGONAL DECOMPOSITION

by

Charlotte Rodriguez

Contents

Notations and conventions

- Let $(X, \|\cdot\|_X)$ an $(Y, \|\cdot\|_Y)$ be two normed vector spaces. By $\|x\|_X \lesssim \|y\|_Y$ we mean that there exists a constant $C > 0$ such that $\|x\|_X \leq C\|y\|_Y$ for any $x \in X,\ y \in Y$.
- $\mathbb{N} \stackrel{\text{def}}{=} \{1, 2, 3, \ldots\}$, $\mathbb{R}_+ \stackrel{\text{def}}{=} [0, \infty)$, $\mathbb{R}_- \stackrel{\text{def}}{=} (-\infty, 0]$.
- $\mathbb{C}_+ \stackrel{\text{def}}{=} \{s \in \mathbb{C} \colon \Re(s) > 0\}$ denotes the right half-plane.
- $\|\cdot\|$ denotes the operator norm, meaning that if $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ are normed vector spaces and $T \in \mathscr{L}(X, Y)$,

$$\|T\| = \sup_{x \in X \setminus \{0\}} \frac{\|Tx\|_Y}{\|x\|_X}.$$

- The space $\mathbb{C}^n$ is endowed with the inner product and induced norm

$$\langle x\,, y\rangle_{\mathbb{C}^n} \stackrel{\text{def}}{=} \sum_{j=1}^{N} x_j \overline{y_j}, \quad |x| \stackrel{\text{def}}{=} \Big(\sum_{k=1}^{n} |x_k|^2\Big)^{\frac{1}{2}},$$

  for $x = (x_1, \ldots, x_n)^T, y = (y_1, \ldots, y_n)^T$. Here $\overline{y_j}$ denotes the complex conjugate of v_j and $|x_k|$ denotes the modulus of x_k, i.e. $|x_k| = \big(\Re(x_k)^2 + \Im(x_k)^2\big)^{\frac{1}{2}}$.
- $\{e_k\}_{k=1}^n$ denotes the canonical basis of $\mathbb{C}^n$, where $e_k = (0, \ldots, 0, 1, 0, \ldots, 0)^T$, for $k \in \{1, \ldots, n\}$, with 1 lying on the k-th position.
- $\mathbb{C}^{m \times n}$ is the space of matrices with m rows and n columns, having complex components.
- Let $M \in \mathbb{C}^{m \times n}$, the conjugate-transpose of M is denoted by M^*. Similarly, the adjoint of a linear operator T is denoted by T^*.
- $\mathbb{O}_{m,n} \in \mathbb{C}^{n \times m}$ denotes the zero matrix. When $n = m$, $\mathbb{O}_n \stackrel{\text{def}}{=} \mathbb{O}_{n,n}$.
- $I_n \in \mathbb{C}^{n \times n}$ denotes the identity matrix.
- For $n \in \mathbb{N}$, we define the set

$$S_- \stackrel{\text{def}}{=} \{f \colon \mathbb{R} \to \mathbb{C}^n \text{ s.t. } \operatorname{supp}(f) \subset \mathbb{R}_-\}$$

  and

$$S_+ \stackrel{\text{def}}{=} \{f \colon \mathbb{R} \to \mathbb{C}^n \text{ s.t. } \operatorname{supp}(f) \subset \mathbb{R}_+\}.$$

- The space $L^2(\mathbb{R}; \mathbb{C}^n)$ contains the functions $f \colon \mathbb{R} \to \mathbb{C}^n$ satisfying

$$\int_{-\infty}^{+\infty} |f(t)|^2 dt < \infty.$$

It is an Hilbert space when endowed with the inner product

$$\langle f\,,g\rangle_{L^2(\mathbb{R};\mathbb{C}^n)} \stackrel{\text{def}}{=} \int_{-\infty}^{+\infty} \langle f(t)\,,g(t)\rangle_{\mathbb{C}^n} dt.$$

- We define the following subspaces of $L^2(\mathbb{R};\mathbb{C}^n)$:

$$L^2(\mathbb{R}_-;\mathbb{C}^n) \stackrel{\text{def}}{=} L^2(\mathbb{R};\mathbb{C}^n)\cap S_-$$

and

$$L^2(\mathbb{R}_+;\mathbb{C}^n) \stackrel{\text{def}}{=} L^2(\mathbb{R};\mathbb{C}^n)\cap S_+.$$

- We denote the dimension of a vector space X by $\dim(X)$.
- Let X, Y be two vector spaces. For a linear map $T : X \to Y$, we denote the range of T by

$$\mathrm{Ran}(T) \stackrel{\text{def}}{=} \{y \in Y\colon \exists x \in X \text{ such that } y = Tx\}.$$

We denote the rank of T by

$$\mathrm{rank}(T) \stackrel{\text{def}}{=} (T) = \dim(\mathrm{Ran}(T)),$$

and the kernel of T by

$$\ker(T) \stackrel{\text{def}}{=} \{x \in X\colon Tx = 0_Y\},$$

where 0_Y is the null vector of Y.

- Let X be a Hilbert space. An operator $T \in \mathscr{L}(X)$ is called *positive*[1] if

$$\langle Tx\,,x\rangle_X \geq 0, \quad \forall x \in X.$$

The operator T is called *strictly positive* if there exists $m > 0$ such that

$$\langle Tx\,,x\rangle_X \geq m\|x\|_X, \quad \forall x \in X,$$

or equivalently

$$\langle Tx\,,x\rangle_X > 0, \quad x \in X \setminus \{0\}.$$

- For a set E we denote by χ_E the characteristic function of E.
- For a normed vector space $(X, \|\cdot\|_X)$ and a subset E, we denote by $\mathrm{clos}(E)$ the closure of E in X.
- For a subset E of $\mathbb{R}$ or $\mathbb{C}$, we denote its measure by $\mathrm{meas}(E)$.

1. If $T \in \mathscr{L}(X)$ is positive, then it is necessarily self-adjoint. This may me shown using, for any $x, y \in X$, the identity

$$4\langle Ty\,,x\rangle_X = \langle T(x+y)\,,x+y\rangle_X - \langle T(x-y)\,,x-y\rangle_X - i\langle T(x+iy)\,,x+iy\rangle_X + i\langle T(x-iy)\,,x-iy\rangle_X.$$

INTRODUCTION

A variety of complex systems of scientific interest or industrial value are described by mathematical models. These models can be used to simulate the behavior of the processes in question, or to modify their behavior. Examples include heat transfer, fluid dynamics, biological systems and signal propagation in large structures.

The complexity of models, measured in terms of the number of coupled first-order differential equations, may reach the hundreds of thousands. Such models are called *large-scale state-space systems*. In particular, discretization in problems that arise from partial differential equations which evolve in three spatial dimensions can easily lead to one million equations. Simulation of the full model is often not feasible and an approximation is necessary in order to simulate and/or control with a reduced computational cost. The goal of *model reduction* is to replace the given mathematical model by a model that is much smaller, yet still describes certain aspects of the system. In control theory, the aim will be to preserve the input-output behavior of the system. If the approximation error is within a given tolerance, only the smaller system needs to be simulated.

Multiple methods (such as moment matching, balanced truncation, reduced basis and proper orthogonal decomposition) have been discovered and are applicable depending on the properties of the model, i.e. whether it is linear or nonlinear, time-independent or time-varying, stable or unstable, etc. Each comes with its own advantages or disadvantages: preservation or not of important properties of the original system, quantifiability of the approximation error and efficiency of resulting algorithms. We refer to [**5**], which brings together different communities and application domains involved in dimension reduction for simulation and control processes. We refer to [**2**] for the approximation of input-output systems (methods related to the singular value decomposition, Krylov or moment matching concepts, and method related to both ideas); motivating examples are also discussed. We also refer to the survey [**4**] on model order reduction and the references therein.

This work is organized as follows.

In Part I we present the singular value decomposition (SVD). We present this method first, as it is related to both the proper orthogonal decomposition and balanced truncation. We factorize an arbitrary matrix as a product of two

unitary matrices and a rectangular matrix whose only nonzero components are on the main diagonal.

In Part II we present the proper orthogonal decomposition (POD). The aim is to find a subspace of the state space of a system of equations in order to subsequently build a (spatial) Galerkin approximation of the solution. We study both discrete and continuous POD problems, and their relation in the context of evolution and parameter-dependent elliptic equations.

In Part III, we present the balanced truncation method and the balanced proper orthogonal decomposition method, applicable for input-output systems. We present error estimates, algorithms and numerical experiments using the former.

I would like to thank Professor Marius Tucsnak for his patience, help and advice during my studies. I would also like to thank Professor Bernhard Haak for his explanations and his insistance on keeping a clear notation. Finally, I would like to thank Cyril from the BMI for always welcoming us to the library.

PART I. SINGULAR VALUE DECOMPOSITION (SVD)

The singular value decomposition (SVD for short) allows to decompose any matrix $M \in \mathbb{C}^{m\times n}$ as a product of two unitary matrices, and a rectangular matrix whose only nonzero components lie on its main diagonal. We do not ask for M to be square or hermitian.

1. Existence of the SVD

Theorem 1.1. — *Let $M \in \mathbb{C}^{m\times n}$ and $p \overset{def}{=} \min\{m,n\}$. Then, there exist unitary*[2] *matrices $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ such that*

$$M = U\Sigma V^*$$

where $\Sigma \overset{def}{=} \big(\sigma_{ij}\big)_{\substack{1\le i\le m\\ 1\le j\le n}} \in \mathbb{C}^{m\times n}$ satisfies

$$\sigma_{ij} = 0 \quad \forall j \neq i, \quad \text{and} \quad \sigma_{11} \ge \sigma_{22} \ge \ldots \ge \sigma_{pp} \ge 0. \tag{1}$$

In order to simplify the notation, for any $k \in \{1,\ldots,p\}$, we rewrite the diagonal elements of the matrix Σ as

$$\sigma_k \overset{\text{def}}{=} \sigma_{kk}.$$

Before proceeding with the proof, we make the following remarks.

Remark 1.2. — The properties (1) of Σ may be reformulated as follows. When $m > n$ (respectively $m < n$), one has

$$\Sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ & & \\ 0 & & \sigma_n \\ 0 & \ldots & 0 \\ \vdots & & \vdots \\ 0 & \ldots & 0 \end{pmatrix}, \quad \text{respectively } \Sigma = \begin{pmatrix} \sigma_1 & & 0 & 0 & \ldots & 0 \\ & \ddots & & \vdots & & \vdots \\ 0 & & \sigma_m & 0 & \ldots & 0 \end{pmatrix},$$

and when $m = n$, Σ is diagonal.

2. A square matrix $U \in \mathbb{C}^{m\times m}$ is called unitary if $UU^* = U^*U = I_m$, meaning that U is invertible with inverse U^*. Equivalently, U is unitary if its columns form an orthonormal basis of $\mathbb{C}^m$.

Remark 1.3. — Denote by $u_1, \dots, u_m$ and $v_1, \dots, v_n$ respectively, the columns of U and V, meaning

$$U = \begin{pmatrix} | & & | \\ u_1 & \dots & u_m \\ | & & | \end{pmatrix}, \quad V = \begin{pmatrix} | & & | \\ v_1 & \dots & v_n \\ | & & | \end{pmatrix}.$$

The set of columns $\{u_1, \dots, u_m\}$ is an orthonormal basis of $\mathbb{C}^m$, while the set $\{v_1, \dots, v_n\}$ form an orthonormal basis of $\mathbb{C}^n$. Consequently, for any $k \in \{1, \dots, m\}$,

$$U^* u_k = e_k$$

where $\{e_k\}_{k=1}^m$ denotes the canonical basis of $\mathbb{C}^m$, and for any $k \in \{1, \dots, n\}$,

$$V^* v_k = \widetilde{e_k}$$

where $\{\widetilde{e_k}\}_{k=1}^n$ denotes the canonical basis of $\mathbb{C}^n$. This implies that, for any $k \in \{1, \dots, p\}$, the vectors u_k, v_k satisfy the following relations

$$\begin{aligned} M v_k &= U \Sigma \widetilde{e_k} = \sigma_k u_k \\ M^* u_k &= V \Sigma^* e_k = \sigma_k u_k, \end{aligned} \tag{2}$$

otherwise, if $k > p$ then $M v_k = 0$ and $M u_k = 0$. Finally, using (2), notice that for any $k \in \{1, \dots, p\}$, v_k and u_k are respectively eigenvectors of M^*M and MM^* associated with the same eigenvalue σ_k^2. In fact, we will see that the proof proposed here relies on the diagonalization of the hermitian matrix M^*M.

Definition 1.4 **(Singular values/vectors).** — Let the singular value decomposition of $M \in \mathbb{C}^{m \times n}$ be given by $M = U \Sigma V^*$, where the matrices U, D and V are defined as in Theorem 1.1. Let $p \overset{\text{def}}{=} \min(m, n)$. We call the elements $\sigma_1, \sigma_2, \dots, \sigma_p$ the *singular values of* M. The columns U and V are respectively called *left-singular vectors* and *right-singular vectors*.

Proof of Theorem 1.1. — The arguments of the proof follow **[8]**. We can assume, without loss of generality, that the number of rows m is larger than the number of columns n. Indeed, if we had $m \leq n$, then one could use this proof to obtain a singular value decomposition of M^* (which then has more rows than columns) and deduce the existence of a decomposition for M.

Since the matrix $M^*M \in \mathbb{C}^{n\times n}$ is hermitian, it can be diagonalized by a unitary matrix V, meaning

$$M^*M = V \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} V^*,$$

where the diagonal components satisfy $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$ (see [**8**, Theorem 1.9, p.44]). The diagonal elements being nonnegative, we may immediately set $\sigma_k \overset{\text{def}}{=} \sqrt{\lambda_k}$ and define the diagonal matrix D by

$$D \overset{\text{def}}{=} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{pmatrix},$$

hence

$$M^*M = VD^2V^*. \tag{3}$$

Denote by r the number of positive eigenvalues of M^*M. Then, D also writes as

$$D = \begin{pmatrix} D_r & \mathbb{O}_{r,n-r} \\ \mathbb{O}_{n-r,r} & \mathbb{O}_{n-r} \end{pmatrix}, \quad \text{for} \quad D_r \overset{\text{def}}{=} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{pmatrix}.$$

We already have found a unitary matrix V satisfying $M^*M = VD^2V^*$. Our aim is now to find the matrices $\Sigma \in \mathbb{C}^{m\times n}$, $U \in \mathbb{C}^{m\times m}$ having the properties described in Theorem 1.1 and satisfying

$$M = U\Sigma V^*. \tag{4}$$

Before proceeding with the general case, we consider what would happen if all eigenvalues of M^*M were positive. In that case, D would be invertible and we could deduce from (3) that

$$(MVD^{-1})^*MVD^{-1} = I_n.$$

Thus, the columns of $MVD^{-1} \in \mathbb{C}^{m\times n}$ would form an orthonormal set in $\mathbb{C}^m$ and we could set

$$MVD^{-1} \overset{\text{def}}{=} U_1 \;\Leftrightarrow\; M = U_1DV^*.$$

In this case, we would only have to add rows of zeros to D and columns of arbitrary numbers (which we denote by U_2) to U_1, so as to get

$$M = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} D \\ \mathbb{O}_{m-r,n} \end{pmatrix} V^*.$$

The matrix U_2 being arbitrary, we would then be allowed to choose it so that U be unitary (recall that we assumed that $m \geq n$ and that we may complete an orthonormal set of $\mathbb{C}^m$ to obtain an orthonormal basis of this space).

For the general case, that is $r < n$, we use a similar reasoning. However, in place of D we use the matrix $J \in \mathbb{C}^{n\times n}$, which corresponds to the matrix D whose diagonal entries equal to zero have been replaced by one, meaning that

$$J \stackrel{\text{def}}{=} \begin{pmatrix} D_r & \mathbb{O}_{r,n-r} \\ \mathbb{O}_{n-r,r} & I_{n-r} \end{pmatrix}.$$

Contrary to D, the matrix J is invertible, of inverse

$$J^{-1} = \begin{pmatrix} D_r^{-1} & \mathbb{O}_{r,n-r} \\ \mathbb{O}_{n-r,r} & I_{n-r} \end{pmatrix}.$$

Observe that

$$(MVJ^{-1})^* MVJ^{-1} = J^{-1}V^*M^*MVJ^{-1} = J^{-1}\Sigma^2 J^{-1} = \begin{pmatrix} I_r & \mathbb{O}_{r,n-r} \\ \mathbb{O}_{n-r,r} & \mathbb{O}_{n-r} \end{pmatrix}.$$

Thus, the first r columns of $MVJ^{-1} \in \mathbb{C}^{m\times n}$ form an orthonormal set in $\mathbb{C}^m$ and the last $n-r$ columns are equal to zero, meaning

$$MVJ^{-1} = \begin{pmatrix} U_1 & \mathbb{O}_{m,n-r} \end{pmatrix},$$

where $U_1 \in \mathbb{C}^{m\times r}$ contains the first r columns of MVJ^{-1}. Consequently, we obtain the following decomposition for M

$$\begin{aligned} M &= \begin{pmatrix} U_1 & \mathbb{O}_{m,n-r} \end{pmatrix} \begin{pmatrix} D_r & \mathbb{O}_{r,n-r} \\ \mathbb{O}_{n-r,r} & I_{n-r} \end{pmatrix} V^* \\ &= \begin{pmatrix} U_1 & \mathbb{O}_{m,n-r} \end{pmatrix} \begin{pmatrix} D_r & \mathbb{O}_{r,n-r} \\ \mathbb{O}_{n-r,r} & \mathbb{O}_{n-r} \end{pmatrix} V^*, \end{aligned}$$

where we only replaced the bottom right block of the second matrix by zeros. This equality still holds when we increase the number of columns of the first

matrix and the number of rows of the second matrix by adding zeros, hence

$$M = \begin{pmatrix} U_1 & \mathbb{O}_{m,m-r} \end{pmatrix} \begin{pmatrix} D_r & \mathbb{O}_{r,n-r} \\ \mathbb{O}_{m-r,r} & \mathbb{O}_{m-r} \end{pmatrix} V^*.$$

For any matrix $U_2 \in \mathbb{C}^{m,m-r}$, M is also equal to

$$M = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \underbrace{\begin{pmatrix} D_r & \mathbb{O}_{r,n-r} \\ \mathbb{O}_{m-r,r} & \mathbb{O}_{m-r,n-r} \end{pmatrix}}_{\stackrel{\text{def}}{=} \Sigma \in \mathbb{C}^{m \times n}} V^*,$$

so it remains to choose U_2 such that $U = \begin{pmatrix} U_1 & U_2 \end{pmatrix}$ is unitary. □

The following corollary adds details to the singular value decomposition.

Corollary 1.5. — *Let the singular value decomposition of $M \in \mathbb{C}^{m \times n}$ be given by $M = U\Sigma V^*$, where the matrices $U \stackrel{def}{=} (U_1\ U_2), V \stackrel{def}{=} (V_1\ V_2)$ and Σ are defined as Theorem 1.1. Then,*

$$r \stackrel{def}{=} rank(M^*M) = rank(MM^*) = rank(M) = rank(M^*)$$

and

$$M = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} D_r & \mathbb{O}_{r,n-r} \\ \mathbb{O}_{m-r,r} & \mathbb{O}_{m-r,n-r} \end{pmatrix} \begin{pmatrix} V_1^* \\ V_2^* \end{pmatrix} = U_1 D_r V_1^*$$

where $U_1 \in \mathbb{C}^{m,r}$, $U_2 \in \mathbb{C}^{m,m-r}$, $V_1 \in \mathbb{C}^{n,r}$, $V_2 \in \mathbb{C}^{n,n-r}$ and $D_r \in \mathbb{C}^{r \times r}$ is a diagonal matrix whose entries on the diagonal are the positive singular values of M ordered decreasingly, meaning that

$$D_r \stackrel{def}{=} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{pmatrix}.$$

Notice that since U and V are unitary, $U_1^*U_1 = I_r$ and $U_2^*U_2 = I_{n-r}$. There are several ways to prove this Corollary using the SVD, here we present one of them.

Proof. — The singular value decomposition of M yields

$$\begin{aligned} M &= U\Sigma V^* \\ M^* &= V\Sigma^* U^* \\ M^*M &= V\Sigma^*\Sigma V^* \\ MM^* &= U\Sigma\Sigma^* U^*, \end{aligned}$$

and $\Sigma^*\Sigma \in \mathbb{C}^{n\times n}$ and $\Sigma\Sigma^* \in \mathbb{C}^{m\times m}$ are both square diagonal matrices whose diagonal entries are the squared singular values of M. Denote by r the number of positive singular values of M.

We will show that r is equal to the rank of M^*M. The matrix $\Sigma^*\Sigma$ also writes as

$$\Sigma^*\Sigma = \begin{pmatrix} D_r^2 & \mathbb{O}_{r,n-r} \\ \mathbb{O}_{n-r,r} & \mathbb{O}_{n-r} \end{pmatrix}.$$

Since V is bijective, $\mathrm{Ran}(V^*) = V^*(\mathbb{C}^n) = \mathbb{C}^n$ and

$$\mathrm{Ran}(M^*M) = (V\Sigma^*\Sigma V^*)(\mathbb{C}^n) = (V\Sigma^*\Sigma)(\mathbb{C}^n) = \mathrm{Ran}(V\Sigma^*\Sigma).$$

Denoting the columns of $\Sigma^*\Sigma$ by c_k, we observe that

$$\{\Sigma^*\Sigma x \colon x \in \mathbb{C}^n\} = \Big\{ \sum_{k=1}^{n} x_k c_k \colon (x_1, \dots, x_n)^T \in \mathbb{C}^n \Big\} = \mathrm{span}\{c_k \colon 1 \le k \le r\},$$

where $\{c_k \colon 1 \le k \le r\}$ is independent in $\mathbb{C}^n$, hence $\mathrm{rank}(\Sigma^*\Sigma) = r$. Since $\mathrm{rank}(V\Sigma^*\Sigma) \le \mathrm{rank}(\Sigma^*\Sigma)$, we deduce that $\mathrm{rank}(M^*M) \le r$. On the other hand, the number of positive eigenvalues cannot exceed $\mathrm{rank}(M^*M)$ since for $k \in \{1, \dots, r\}$ the vector v_k rewrites as $v_k = M^*M\frac{v_k}{\sigma_k^2}$, and consequently the following inclusion holds

$$\mathrm{span}\{v_k : 1 \le k \le r\} \subset \mathrm{Ran}(M^*M).$$

Thus, $\mathrm{rank}(M^*M) = r$.

We may show that the rank of M, M^* and MM^* also equal r using similar arguments, hence

$$r = \mathrm{rank}(M^*M) = \mathrm{rank}(MM^*) = \mathrm{rank}(M) = \mathrm{rank}(M^*) \le p. \qquad \square$$

2. Singular values and matrix norm

A matrix $M \in \mathbb{C}^{m\times n}$, as a bounded linear operator from $\mathbb{C}^n$ into $\mathbb{C}^m$, has the norm

$$\|M\| \overset{\text{def}}{=} \sup_{x\in\mathbb{C}^n\setminus\{0\}} \frac{|Mx|}{|x|}.$$

Here, we will see that $\|M\|$ is equal to the largest singular value of M.

Definition 2.1. — For a matrix $M \in \mathbb{C}^{m\times n}$, we denote by $\overline{\sigma}(M)$ the largest singular value σ_1 of M.

Proposition 2.2. — *Let $M \in \mathbb{C}^{m\times n}$, then $\|M\| = \overline{\sigma}(M)$.*

Proof. — Let x be an arbitrary nonzero element of $\mathbb{C}^n$. The set $\{v_k\}_{k=1}^n$ is an orthonormal basis in $\mathbb{C}^n$, thus

$$|Mx|^2 = \langle Mx\,, Mx\rangle_{\mathbb{C}^n} = \langle M^*Mx\,, x\rangle_{\mathbb{C}^n} = \Big\langle M^*M\Big(\sum_{k=1}^n \langle x\,, v_k\rangle_{\mathbb{C}^n} v_k\Big)\,, x\Big\rangle_{\mathbb{C}^n}.$$

Since v_k are eigenvectors of M^*M associated with eigenvalues σ_k^2, we get

$$\begin{aligned} |Mx|^2 &= \sum_{k=1}^n \langle x\,, v_k\rangle_{\mathbb{C}^n} \langle M^*Mv_k\,, x\rangle_{\mathbb{C}^n} \\ &= \sum_{k=1}^n \sigma_k^2 |\langle x\,, v_k\rangle_{\mathbb{C}^n}|^2 \\ &\le \sigma_1^2 \sum_{k=1}^n |\langle x\,, v_k\rangle_{\mathbb{C}^n}|^2 \quad = \ \sigma_1^2 |x|^2. \end{aligned}$$

Hence $\|M\| \le \sigma_1$. Moreover, since the vector v_1 (which is of norm one) satisfies

$$|Mv_1| = |\sigma_1 u_1| = \sigma_1 |u_1| = \sigma_1,$$

we can conclude that $\|M\| = \sigma_1 = \overline{\sigma}(M)$. □

PART II. PROPER ORTHOGONAL DECOMPOSITION (POD)

The proper orthogonal decomposition is known, in other areas, as the Karlhunen-Loève decomposition, and is closely connected to the principal component analysis and the singular value decomposition.

The reduced-order approach is based on projecting the system onto subspaces consisting of basis elements that contain characteristics of the expected solution. This is in contrast to, for example the finite element method, where the basis elements of the subspaces do not relate to the physical properties of the system that they approximate.

For example, we may consider an evolution problem for which a POD basis $\{\psi_k\}_{k=1}^{\ell}$ has been computed, the solution is then approximated by

$$z(t) \approx \sum_{k=1}^{\ell} \langle z(t)\,, \psi_k \rangle_X \psi_k.$$

The basis elements express characteristics of the expected solution and approximate the given data in a least-square optimal sense. The key idea of the POD is to reduce a large number of interdependent vectors (called snapshots) to a much smaller number of independent vectors, while retaining as much variation as possible in the original vectors. The POD method may be applied to infinite dimensional systems as well.

Here we study the discrete proper orthogonal decomposition, then we consider the continuous POD and its relation to the discrete POD in the context of parameter-dependent elliptic PDEs and evolution PDEs.

We henceforth denote by X a complex separable Hilbert space endowed with an inner product $\langle \cdot\,, \cdot \rangle_X$ and the induced norm $\| \cdot \|_X$.

3. Discrete POD

3.1. The problem and the governing operators. — We consider N elements of X denoted $z_1, z_2, \ldots, z_N$ and called *snapshots*. We define the closed subspace of X (endowed with $\langle \cdot\,, \cdot \rangle_X$)

$$\mathcal{V}_N \stackrel{\text{def}}{=} \operatorname{span}\{z_j : 1 \le j \le N\},$$

whose dimension is denoted by $d(N) < \infty$. Let $\alpha_1, \alpha_2, \ldots, \alpha_N$ be positive weights. We fix a priori the integer $\ell \le d(N)$. The following notation will be used throughout the sections on discrete and continuous POD.

Definition 3.1. — Let $\mathcal{B} \overset{\text{def}}{=} \{\phi_k\}_{k=1}^{\ell}$ be an orthonormal set in X. We denote by $P^{\mathcal{B}}$ the orthogonal projection on the subspace of X spanned by $\mathcal{B}$, meaning that for any $x \in X$,

$$P^{\mathcal{B}} x = \sum_{k=1}^{\ell} \langle x \,, \phi_k \rangle_X \phi_k.$$

As a consequence, for any $x \in X$, Pythagoras' theorem yields $\|P^{\mathcal{B}} x\|_X^2 = \sum_{k=1}^{\ell} |\langle x \,, \phi_k \rangle_X|^2$. We may now give a definition for the *discrete POD basis of rank* ℓ.

Definition 3.2 **(Discrete POD basis).** — The *discrete POD basis of rank* ℓ is the set $\boldsymbol{\beta} \overset{\text{def}}{=} \{\psi_k\}_{k=1}^{\ell}$ solution to

$$(\mathrm{P}_{N,1}) \qquad \begin{cases} \arg\min \Big\{ \displaystyle\sum_{j=1}^{N} \alpha_j \|z_j - P^{\mathcal{B}} z_j\|_X^2 \colon \mathcal{B} \overset{\text{def}}{=} \{\phi_k\}_{k=1}^{\ell} \subset \mathcal{V}_N \Big\} \\ \text{subject to } \big\langle \phi_i \,, \phi_j \big\rangle_X = \delta_{i,j} \quad \text{for } 1 \leq i, j \leq \ell \end{cases}$$

This could also be reformulated as

$$\sum_{j=1}^{N} \alpha_j \|z_j - P^{\boldsymbol{\beta}} z_j\|_X^2 = \min \Big\{ \sum_{j=1}^{N} \alpha_j \|z_j - P^{\mathcal{B}} z_j\|_X^2 \colon \mathcal{B} \text{ orthonormal in } \mathcal{V}_N \Big\}.$$

In other words, $\boldsymbol{\beta}$ is an orthonormal set in $\mathcal{V}_N$ that minimizes the mean squared error between the elements $\{z_j \colon 1 \leq j \leq N\}$ and their corresponding orthogonal projections on subspaces of $\mathcal{V}_N$.

Following the arguments given by **[14]** and **[22]**, we prove the existence of the POD basis and see different ways of computing this basis. First, we introduce and study operators that will appear in this section.

Definition 3.3 **($\mathcal{Y}_N$).** — Let $\mathcal{Y}_N \colon \mathbb{C}^N \to X$ be the linear operator defined by

$$\mathcal{Y}_N v \overset{\text{def}}{=} \sum_{j=1}^{N} \sqrt{\alpha_j} v_j z_j.$$

Using the Cauchy-Schwarz inequality, observe that $\mathcal{Y}_N$ is bounded, since

$$\big\|\mathcal{Y}_N v\big\|_X = \Big\| \sum_{j=1}^{N} \sqrt{\alpha_j} v_j z_j \Big\|_X \leq \underbrace{\Big(\sum_{j=1}^{N} \alpha_j \|z_j\|_X^2 \Big)^{\frac{1}{2}}}_{<\infty} \Big(\sum_{j=1}^{N} |v_j|^2 \Big)^{\frac{1}{2}} \lesssim \|v\|_{\mathbb{C}^N},$$

holds for any $v \in \mathbb{C}^N$. Moreover, its adjoint $\mathcal{Y}_N^*$ consequently belongs to $\mathscr{L}(X, \mathbb{C}^N)$ and is given by

$$\mathcal{Y}_N^* x = \begin{pmatrix} \sqrt{\alpha_1}\langle x\,, z_1\rangle_X \\ \vdots \\ \sqrt{\alpha_N}\langle x\,, z_N\rangle_X \end{pmatrix}.$$

This follows from observing that

$$\langle \mathcal{Y}_N v\,, x\rangle_X = \sum_{j=1}^{N} v_j \overline{\sqrt{\alpha_j}\langle x\,, z_j\rangle}_X = \langle v\,, \mathcal{Y}_N^* x\rangle_{\mathbb{C}^N}$$

holds for any $v \in \mathbb{C}^N$ and $w \in X$.

The following Lemma will allow us relate the rank of $\mathcal{Y}_N$ to the rank of operators defined bellow.

Lemma 3.4. — *Let $\mathcal{H}_1, \mathcal{H}_2$ be separable Hilbert spaces and let $T \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ be such that* $\operatorname{rank}(T) < \infty$. *Then,*

$$\operatorname{rank}(T) = \operatorname{rank}(T^*) = \operatorname{rank}(T^*T) = \operatorname{rank}(TT^*).$$

Proof. — Set $r \overset{\text{def}}{=} \operatorname{rank}(T)$. We begin by showing that T and its adjoint have the same rank. Denote by Π the orthogonal projection from $\mathcal{H}_2$ onto $\operatorname{Ran}(T)$. Since Π is the identity operator on $\operatorname{Ran}(T)$, we have $\Pi T = T$, and Π being self-adjoint, $T^*\Pi = T^*$. Moreover, $\operatorname{Ran}(\Pi) = \operatorname{Ran}(T)$, hence $\operatorname{rank}(\Pi) = r$ and

$$\operatorname{Ran}(T^*) = \operatorname{Ran}(T^*\Pi) = T^*\big(\operatorname{Ran}(\Pi)\big) = T^*\big(\operatorname{span}\{\pi_k \colon 1 \le k \le r\}\big),$$

for a basis $\{\pi_k \colon 1 \le k \le r\}$ of $\operatorname{Ran}(\Pi)$. Consequently,

$$\operatorname{Ran}(T^*) \subset \operatorname{span}\big\{T^*\pi_k \colon 1 \le k \le r\big\}$$

hence $\operatorname{rank}(T^*) \le r = \operatorname{rank}(T)$. Applying this result to T^* instead of T yields $\operatorname{rank}(T^{**}) \le \operatorname{rank}(T^*)$, thus

$$\operatorname{rank}(T) \le \operatorname{rank}(T^*).$$

Hence, $\operatorname{rank}(T) = \operatorname{rank}(T^*)$. Next, we show that TT^* is of rank r. We rewrite

$$\operatorname{Ran}(TT^*) = T\big(\operatorname{Ran}(T^*)\big)$$

and denote by $\{\varphi_k\}_{k=1}^r$ an orthonormal basis of $\operatorname{Ran}(T^*)$. We will show that $\{T\varphi_k\}_{k=1}^r$ is independent in $\mathcal{H}_2$. Let $a_1, \dots, a_r$ be arbitrary complex numbers

such that

$$\sum_{k=1}^{r} a_k T\varphi_k = 0$$

holds in $\mathcal{H}_2$. Then,

$$\begin{aligned} & \Big\langle \sum_{k=1}^{r} a_k T\varphi_k \,, y \Big\rangle_{\mathcal{H}_2} = 0, \quad \forall y \in \mathcal{H}_2 \\ \iff \quad & \sum_{k=1}^{r} a_k \langle \varphi_k \,, T^* y \rangle_{\mathcal{H}_1} = 0, \quad \forall y \in \mathcal{H}_2 \\ \iff \quad & \sum_{k=1}^{r} a_k \langle \varphi_k \,, x \rangle_{\mathcal{H}_1} = 0, \quad \forall x \in \mathrm{Ran}(T^*). \end{aligned}$$

Choosing $x = \varphi_j$ for $j \in \{1, \ldots, r\}$ shows that $a_1 = \ldots = a_r = 0$. Hence $\{T\varphi_k\}_{k=1}^r$ is independent in $\mathcal{H}_2$ and, since it is a subset of $\mathrm{Ran}(TT^*)$, we deduce that $r \leq \mathrm{rank}(TT^*)$. As $\mathrm{Ran}(TT^*) \subset \mathrm{Ran(T)}$, we also know that $\mathrm{rank}(TT^*) \leq r$, hence $\mathrm{rank}(TT^*) = r$. Applying this result to T^* instead of T yields $\mathrm{rank}(T^*T^{**}) = r$, thus

$$\mathrm{rank}(T^*T) = r. \qquad \square$$

Definition 3.5 **($\mathcal{S}_N$).** — Let $\mathcal{S}_N \colon X \longrightarrow X$ be the linear operator defined by

$$\mathcal{S}_N w \stackrel{\mathrm{def}}{=} \sum_{j=1}^{N} \alpha_j \langle w \,, z_j \rangle_X z_j.$$

Observe that $\mathcal{S}_N$ can be rewritten as $\mathcal{S}_N = \mathcal{Y}_N \mathcal{Y}_N^*$, since

$$\mathcal{S}_N w = \sum_{j=1}^{N} \alpha_j \langle w \,, z_j \rangle_X z_j = \sum_{j=1}^{N} \sqrt{\alpha_j} \big(\mathcal{Y}_N^* w\big)_j z_j = \mathcal{Y}_N \mathcal{Y}_N^* w$$

holds for any $w \in X$. Moreover, $\mathcal{S}_N$ is bounded (as a composition of bounded operators) and self-adjoint (as $(\mathcal{Y}_N \mathcal{Y}_N^*)^* = \mathcal{Y}_N^{**} \mathcal{Y}_N^* = \mathcal{Y}_N \mathcal{Y}_N^*$). Observe that $\mathcal{S}_N$ is positive, either because of its self-adjointness (as $\langle \mathcal{S}_N w \,, w \rangle_X = \langle \mathcal{Y}_N \mathcal{Y}_N^* w \,, w \rangle_X = \|\mathcal{Y}_N^* w\|_{\mathbb{C}^N}^2 \geq 0$ holds for any $w \in X$), or by the computation

$$\langle \mathcal{S}_N w \,, w \rangle_X = \Big\langle \sum_{j=1}^{N} \alpha_j \langle w \,, z_j \rangle_X z_j \,, w \Big\rangle_X = \sum_{j=1}^{N} \alpha_j |\langle w \,, z_j \rangle_X|^2 \ \geq 0, \quad \forall w \in X.$$

Since $\mathcal{S}_N$ has values in $\mathcal{V}_N$, its rank is finite, hence $\mathcal{S}_N$ is compact. Moreover, since $\mathcal{S}_N = \mathcal{Y}_N\mathcal{Y}_N^*$ and

$$\mathrm{Im}(\mathcal{Y}_N) = \Big\{ \sum_{j=1}^{N} \sqrt{\alpha_j} v_j z_j \colon (v_1, \dots, v_N)^T \in \mathbb{C}^N \Big\} = \mathrm{span}\{z_j \colon 1 \leq j \leq N\},$$

we have $\mathrm{rank}(\mathcal{Y}_N) = d(N)$. As a consequence, the Lemma 3.4 yields that $\mathrm{rank}(\mathcal{S}_N) = d(N)$.

Definition 3.6 **($\mathcal{K}_N$).** — Let $\mathcal{K}_N$ be the linear operator defined by $\mathcal{K}_N \stackrel{\text{def}}{=} \mathcal{Y}_N^*\mathcal{Y}_N$.

This operator lies in $\mathscr{L}(\mathbb{C}^N)$, is self-adjoint and compact (since its rank is finite). Observe that $\mathcal{K}_N$ is positive, either by its self-adjointness, or by computing, for any $v \in \mathbb{C}^N$,

$$\langle \mathcal{K}_N v \, , v \rangle_{\mathbb{C}^N} = \sum_{n=1}^{N} \Big(\sum_{j=1}^{N} \sqrt{\alpha_n \alpha_j} v_j \langle z_j \, , z_n \rangle_X \Big) \overline{v_n} = \Big\| \sum_{j=1}^{N} \sqrt{\alpha_j} v_j z_j \Big\|_X^2 \geq 0.$$

Moreover, $\mathcal{K}_N$ may be represented by the hermitian matrix defined bellow, whose rows are indexed by i and columns indexed by j,

$$\Big(\sqrt{\alpha_i \alpha_j} \overline{\langle z_i \, , z_j \rangle}_X \Big)_{1 \leq i,j \leq N} \in \mathbb{C}^{N \times N}$$

since

$$\begin{aligned} \mathcal{K}_N v &= \begin{pmatrix} \sqrt{\alpha_1} \langle \mathcal{Y}_N v \, , z_1 \rangle_X \\ \vdots \\ \sqrt{\alpha_N} \langle \mathcal{Y}_N v \, , z_N \rangle_X \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{N} \sqrt{\alpha_1 \alpha_j} v_j \langle z_j \, , z_1 \rangle_X \\ \vdots \\ \sum_{j=1}^{N} \sqrt{\alpha_N \alpha_j} v_j \langle v_j \, , z_N \rangle_X \end{pmatrix} \\ &= \Big(\sqrt{\alpha_i \alpha_j} \langle z_j \, , z_i \rangle_X \Big)_{1 \leq i,j \leq N} \begin{pmatrix} v_1 \\ \vdots \\ v_N \end{pmatrix} \end{aligned}$$

holds for any $v \in \mathbb{C}^N$.

3.2. Existence of the discrete POD-basis. — We now show that the Definition 3.2 makes sense, i.e. that there exists a discrete POD basis of rank ℓ. To this end, we first have the following result which allows us to reformulate the problem.

Lemma 3.7. — *The problem* $(\mathrm{P}_{N,1})$ *has the same solutions as*

$$(\mathrm{P}_{N,2}) \qquad \begin{cases} \arg\min \Big\{ -\displaystyle\sum_{j=1}^{N} \alpha_j \|P^{\mathcal{B}} z_j\|_X^2 : \mathcal{B} \overset{def}{=} \{\phi_k\}_{k=1}^{\ell} \subset \mathcal{V}_N \Big\} \\ \text{subject to } \big\langle \phi_i \, , \phi_j \big\rangle_X = \delta_{i,j} \quad \text{for } 1 \leq i,j \leq \ell. \end{cases}$$

Proof. — If $\mathcal{B} \overset{\text{def}}{=} \{\phi_k\}_{k=1}^{\ell}$ is an orthonormal subset of X, we can write, for any $j \in \{1, \ldots, N\}$,

$$\begin{aligned} \|z_j - P^{\mathcal{B}} z_j\|_X^2 \; &= \Big\langle z_j - \sum_{k=1}^{\ell} \langle z_j \, , \phi_k \rangle_X \phi_k \, , \; z_j - \sum_{m=1}^{\ell} \langle z_j \, , \phi_m \rangle_X \phi_m \Big\rangle_X \\ &= \|z_j\|_X^2 - \sum_{m=1}^{\ell} \overline{\langle z_j \, , \phi_m \rangle}_X \langle z_j \, , \phi_m \rangle_X - \sum_{k=1}^{\ell} \langle z_j \, , \phi_k \rangle_X \langle \phi_k \, , z_j \rangle_X \\ &\quad + \sum_{k=1}^{\ell} \sum_{m=1}^{\ell} \langle z_j \, , \phi_k \rangle_X \overline{\langle z_j \, , \phi_m \rangle}_X \langle \phi_k \, , \phi_m \rangle_X \\ &= \|z_j\|_X^2 - \sum_{k=1}^{\ell} |\langle z_j \, , \phi_k \rangle_X|^2 \; = \; \|z_j\|_X^2 - \|P^{\mathcal{B}} z_j\|_X^2. \end{aligned}$$

Consequently,

$$\sum_{j=1}^{N} \alpha_j \|z_j - P^{\mathcal{B}} z_j\|_X^2 = \sum_{j=1}^{N} \alpha_j \|z_j\|_X^2 - \sum_{j=1}^{N} \alpha_j \|P^{\mathcal{B}} z_j\|_X^2,$$

and since $\sum_{j=1}^{N} \alpha_j \|z_j\|_X^2$ does not depend on $\{\phi_k\}_{k=1}^{\ell}$, any orthonormal set minimizes the right-hand side of the above equation if and only if it also minimizes the term

$$-\sum_{j=1}^{N} \alpha_j \|P^{\mathcal{B}} z_j\|_X^2. \qquad \square$$

The following Proposition reveals a property that is necessarily satisfied by a discrete POD basis.

Proposition 3.8. — *Assume that* $\boldsymbol{\beta} \overset{def}{=} \{\psi_k\}_{k=1}^{\ell}$ *is a discrete POD basis of rank* ℓ*. Then the elements of* $\boldsymbol{\beta}$ *necessarily satisfy the following eigenvalue problem*

$$(\mathrm{Eig}_{1,N}) \qquad \mathcal{S}_N \psi_k = \lambda_k \psi_k \quad \text{for} \quad 1 \leq k \leq \ell.$$

Proof. — We assume that $\boldsymbol{\beta} \overset{\text{def}}{=} \{\psi_k\}_{k=1}^{\ell}$ is a discrete POD basis of rank ℓ. First we rewrite $(\mathrm{P}_{N,2})$ as minimization problem with equality constrains. For $k \in \{1, \dots, \ell\}$, define the maps $c_k \colon X^\ell \to \mathbb{C}^{\ell-k+1}$ by

$$c_k(\varphi_1, \dots, \varphi_\ell) \overset{\text{def}}{=} \begin{pmatrix} \langle \varphi_k \, , \varphi_k \rangle_X - 1 \\ \langle \varphi_{k+1} \, , \varphi_k \rangle_X \\ \langle \varphi_{k+2} \, , \varphi_k \rangle_X \\ \vdots \\ \langle \varphi_\ell \, , \varphi_k \rangle_X \end{pmatrix},$$

and define the map $J \colon X^\ell \to \mathbb{R}$ by

$$J(\varphi_1, \dots, \varphi_\ell) \overset{\text{def}}{=} -\sum_{j=1}^{N} \alpha_j \sum_{k=1}^{\ell} |\langle z_j \, , \varphi_k \rangle_X|^2,$$

we rewrite $(\mathrm{P}_{N,2})$ as

$$(\mathrm{P}_{N,3}) \qquad \begin{cases} \arg\min \left\{ J(\phi_1, \dots, \phi_\ell) \colon \{\phi_k\}_{k=1}^{\ell} \subset \mathcal{V}_N \right\} \\ \text{subject to } c_k(\phi_1, \dots, \phi_\ell) = 0 \quad \text{for } 1 \leq k \leq \ell. \end{cases}$$

For this problem, the Lagrangian $L \colon X^\ell \times \mathbb{C}^\ell \times \mathbb{C}^{\ell-1} \times \ldots \times \mathbb{C} \to \mathbb{C}$ is defined by

$$L(\varphi_1, \dots, \varphi_\ell, \eta_1, \dots, \eta_l) \overset{\text{def}}{=} J(\varphi_1, \dots, \varphi_\ell) + \sum_{k=1}^{\ell} \left\langle c_k(\varphi_1, \dots, \varphi_\ell) \, , \eta_k \right\rangle_{\mathbb{C}^{\ell-k+1}},$$

where $\eta_k \overset{\text{def}}{=} (\eta_k^1, \ \eta_k^2, \ \dots, \ \eta_k^{\ell-k+1})^T$. Since $\boldsymbol{\beta}$ is a solution of $(\mathrm{P}_{N,3})$, the first order optimality condition asserts that there exists Lagrange multipliers $\{\eta_k\}_{k=1}^{\ell} \subset \mathbb{C}^\ell \times \mathbb{C}^{\ell-1} \times \ldots \times \mathbb{C}$ such that $\boldsymbol{\beta}$ is solution to

$$(\mathrm{OC}_N) \qquad \begin{cases} \dfrac{\partial L}{\partial \psi_m}(\psi_1, \dots, \psi_\ell, \eta_1, \dots, \eta_\ell) = 0 \quad \text{in } X \\ \text{subject to } e_m(\psi_1, \dots, \psi_\ell) = 0 \quad \text{in } \mathbb{C}^{\ell-m+1} \end{cases} \qquad \text{for } 1 \leq m \leq \ell$$

The first line of (OC_N) rewrites as

$$\frac{d}{d\varepsilon} L\big(\psi_1, \dots, (\psi_m + \varepsilon \mathbf{v}), \dots, \psi_\ell, \eta_1, \dots, \eta_l\big)\Big|_{\varepsilon=0} = 0, \quad \forall \mathbf{v} \in X, \ 1 \leq m \leq \ell.$$

Equivalently, we may decompose L by gathering its components the following may

$$\begin{aligned}
\frac{d}{d\varepsilon}\Big[&- \Big(\sum_{j=1}^{N}\alpha_j \sum_{\substack{k=1\\k\neq m}}^{\ell} |\langle z_j\,,\psi_k\rangle_X|^2\Big) + \Big(\sum_{j=1}^{N}\alpha_j |\langle z_j\,,\psi_m+\varepsilon\mathbf{v}\rangle_X|^2\Big)\\
&+ \Big(\sum_{\substack{k=1\\k\neq m}}^{\ell}\big(\langle\psi_k\,,\psi_k\rangle_X-1\big)\overline{\eta_k^1}\Big) + \big(\langle\psi_m+\varepsilon\mathbf{v}\,,\psi_m+\varepsilon\mathbf{v}\rangle_X-1\big)\overline{\eta_m^1}\\
&+ \Big(\sum_{k=1}^{m-1}\sum_{\substack{q=k+1\\q\neq m}}^{\ell}\langle\psi_q\,,\psi_k\rangle_X\overline{\eta_k^{q-k+1}}\Big) + \Big(\sum_{k=1}^{m-1}\langle\psi_m+\varepsilon\mathbf{v}\,,\psi_k\rangle_X\overline{\eta_k^{m-k+1}}\Big)\\
&+ \Big(\sum_{q=m+1}^{\ell}\langle\psi_q\,,\psi_m+\varepsilon\mathbf{v}\rangle_X\overline{\eta_m^{q-m+1}}\Big)\\
&+ \Big(\sum_{k=m+1}^{\ell}\sum_{q=k+1}^{\ell}\langle\psi_q\,,\psi_k\rangle_X\overline{\eta_k^{q-k+1}}\Big)\Big]\Big|_{\varepsilon=0} = 0,
\end{aligned}$$

where the first line concerns the functional J while the other lines concern the inner products between e_k and η_k (the second line concerns the first component of η_k and the third, fourth and fifth lines concern the remaining components of η_k). In the right-hand side of this equation we can, equivalently, remove from the sums the terms not containing ε, thus the equation reduces to

$$\begin{aligned}
\frac{d}{d\varepsilon}\Big[&\Big(\sum_{j=1}^{N}\alpha_j\langle z_j\,,\psi_m+\varepsilon\mathbf{v}\rangle_X\langle\psi_m+\varepsilon\mathbf{v}\,,z_j\rangle_X\Big) + \big(\langle\psi_m+\varepsilon\mathbf{v}\,,\psi_m+\varepsilon\mathbf{v}\rangle_X-1\big)\overline{\eta_m^1}\\
&+ \Big(\sum_{k=1}^{m-1}\langle\psi_m+\varepsilon\mathbf{v}\,,\psi_k\rangle_X\overline{\eta_k^{m-k+1}}\Big) + \Big(\sum_{q=m+1}^{\ell}\langle\psi_q\,,\psi_m+\varepsilon\mathbf{v}\rangle_X\overline{\eta_m^{q-m+1}}\Big)\Big]\Big|_{\varepsilon=0} = 0,
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
-\sum_{j=1}^{N}\alpha_j\big(\langle z_j\,,\psi_m\rangle_X\langle\mathbf{v}\,,z_j\rangle_X + \langle\psi_m\,,z_j\rangle_X\langle z_j\,,\mathbf{v}\rangle_X\big) + \big(\langle\psi_m\,,\mathbf{v}\rangle_X + \langle\mathbf{v}\,,\psi_m\rangle_X\big)\overline{\eta_m^1}\\
+\sum_{k=1}^{m-1}\langle\mathbf{v}\,,\psi_k\rangle_X\overline{\eta_k^{m-k+1}} + \sum_{q=m+1}^{\ell}\langle\psi_q\,,\mathbf{v}\rangle_X\overline{\eta_m^{q-m+1}} = 0.
\end{aligned}$$

Using the sesquilinearity of the inner product and recognizing the term $\mathcal{S}_N\psi_m$, we obtain for any $\mathbf{v} \in X$,

$$\begin{aligned}-2\Re\langle \mathcal{S}_N\psi_m\,,\mathbf{v}\rangle_X + \langle\overline{\eta_m^1}\psi_m\,,\mathbf{v}\rangle_X + \langle\mathbf{v}\,,\eta_m^1\psi_m\rangle_X \\ + \Big\langle \mathbf{v}\,,\sum_{k=1}^{m-1}\eta_k^{m-k+1}\psi_k\Big\rangle_X + \Big\langle\sum_{q=m+1}^{\ell}\overline{\eta_m^{q-m+1}}\psi_q\,,\mathbf{v}\Big\rangle_X = 0,\end{aligned}$$

which is equivalent to having

$$\begin{aligned}\Re\Big\langle -2\mathcal{S}_N\psi_m + (\eta_m^1+\overline{\eta_m^1})\psi_m + \sum_{k=1}^{m-1}\eta_k^{m-k+1}\psi_k \\ + \sum_{k=2}^{\ell-m+1}\overline{\eta_m^k}\psi_{k+m-1}\,,\ \mathbf{v}\Big\rangle_X = 0\end{aligned} \tag{5}$$

and

$$\Im\Big\langle(\overline{\eta_m^1}-\eta_m^1)\psi_m - \sum_{k=1}^{m-1}\eta_k^{m-k+1}\psi_k + \sum_{q=m+1}^{\ell}\overline{\eta_m^{q-m+1}}\psi_q\,,\ \mathbf{v}\Big\rangle_X = 0. \tag{6}$$

For any $\mathbf{v} \in X$, the vector $i\mathbf{v}$ also lies in X, hence

$$\begin{aligned}0 &= \Im\Big\langle(\overline{\eta_m^1}-\eta_m^1)\psi_m - \sum_{k=1}^{m-1}\eta_k^{m-k+1}\psi_k + \sum_{q=m+1}^{\ell}\overline{\eta_m^{q-m+1}}\psi_q\,,\ i\mathbf{v}\Big\rangle_X \\ &= \Re\Big\langle(\overline{\eta_m^1}-\eta_m^1)\psi_m - \sum_{k=1}^{m-1}\eta_k^{m-k+1}\psi_k + \sum_{q=m+1}^{\ell}\overline{\eta_m^{q-m+1}}\psi_q\,,\ \mathbf{v}\Big\rangle_X.\end{aligned} \tag{7}$$

The equations (6) and (7) imply that

$$\Big\langle(\overline{\eta_m^1}-\eta_m^1)\psi_m - \sum_{k=1}^{m-1}\eta_k^{m-k+1}\psi_k + \sum_{q=m+1}^{\ell}\overline{\eta_m^{q-m+1}}\psi_q\,,\ \mathbf{v}\Big\rangle_X, \quad \forall\mathbf{v}\in X$$

meaning that

$$(\overline{\eta_m^1}-\eta_m^1)\psi_m - \sum_{k=1}^{m-1}\eta_k^{m-k+1}\psi_k + \sum_{q=m+1}^{\ell}\overline{\eta_m^{q-m+1}}\psi_q = 0$$

holds in X. Since the set $\{\psi_k\}_{k=1}^{\ell}$ is independent in X, the coefficients themselves are equal to zeros:

$$\begin{cases}\eta_k^{m-k+1} = 0, & \text{for } 1 \le k \le (m+1)\\ \eta_m^{q-m+1} = 0, & \text{for } (m+1) \le q \le \ell\\ \overline{\eta_m^1} = \eta_m^1,\end{cases}$$

the last equation meaning that η_m^1 is real. Injecting this information in (5) yields

$$\Re\big\langle -2\mathcal{S}_N\psi_m + 2\eta_m^1\psi_m\,,\ \mathbf{v}\big\rangle_X = 0$$

Choosing $\mathbf{v} = -2\mathcal{S}_N\psi_m + 2\eta_m^1\psi_m$, we deduce that

$$\mathcal{S}_N\psi_m = \eta_m^1\psi_m$$

holds in X. Thus, a solution $\{\psi_k\}_{k=1}^{\ell}$ of $(\mathrm{P}_{N,3})$ necessarily satisfies $(\mathrm{Eig}_{1,N})$. □

Before giving the existence Theorem for the discrete POD basis we state the following Theorem from functional analysis which will be used both in this section and in the section on continuous POD. The proof can be found for example in [**17**, Theorem VI.16, p.203]

Theorem 3.9 **(Hilbert-Schmidt).** — *Let X be a separable Hilbert space, and T a self-adjoint compact operator on H. Then, there is an orthonormal basis $\{\varphi_k\}_{k=1}^{\infty}$ of X so that $T\varphi_k = \lambda_k\varphi_k$ and $\lim_{k\to\infty}\lambda_k = 0$.*

Theorem 3.10 **(Existence, discrete POD).** — *There exists eigenvalues $\{\lambda_k\}_{k=1}^{\infty}$ of $\mathcal{S}_N$ and associated orthonormal eigenvectors $\{\psi_k\}_{k=1}^{\infty}$ satisfying*

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{d(N)} > 0 = \lambda_{d(N)+1} = \lambda_{d(N)+2} = \ldots$$

For $\ell \leq d(N)$, the first ℓ eigenvectors $\boldsymbol{\beta} \overset{def}{=} \{\psi_k\}_{k=1}^{\ell}$ form a discrete POD basis of rank ℓ, and

$$\sum_{j=1}^{N}\alpha_j\|z_j - P^{\boldsymbol{\beta}}z_j\|_X^2 = \sum_{k=\ell+1}^{d(N)}\lambda_k.$$

Proof. — Consider now the problem $(\mathrm{Eig}_{1,N})$. We have seen that $\mathcal{S}_N$ is a self-adjoint compact operator from the separable Hilbert space X to itself, consequently the Hilbert-Schmidt theorem asserts that there exists an orthonormal basis $\{\psi_k\}_{k=1}^{\infty}$ for X such that

$$\mathcal{S}_N\psi_k = \lambda_k\psi_k \quad \text{with} \quad \lambda_k \xrightarrow[k\to\infty]{} 0. \tag{8}$$

Since the eigenvalues converge to zero, we may assume without loss of generality that they are ordered decreasingly. First we show that

$$\lambda_1 \geq \ldots \geq \lambda_{d(N)} > 0 \quad \text{and} \quad \lambda_k = 0, \text{ for } k > d(N). \tag{9}$$

Denote the number of positive eigenvalues by p, meaning that

$$\lambda_1 \geq \ldots \geq \lambda_p > 0 = \lambda_{p+1} = \lambda_{p+2} = \ldots \tag{10}$$

On the one hand, for any $k \in \{1, \dots, p\}$, the eigenvector ψ_k rewrites as

$$\psi_k = S\Big(\frac{\psi_k}{\lambda_k}\Big) \in \mathrm{Ran}(\mathcal{S}_N)$$

hence $\mathrm{span}\{\psi_k\}_{k=1}^p \subset \mathrm{Ran}(\mathcal{S}_N)$. On the other hand, take $y \in \mathrm{Ran}(\mathcal{S}_N)$, then $y = \mathcal{S}_N x$ for some $x \in X$. Since $\{\psi_k\}_{k=1}^\infty$ is an orthonormal basis of X, the vector x is the limit in X of a sequence $(x_m) \subset \mathrm{span}\big(\{\psi_k\}_{k=1}^\infty\big)$. The map $\mathcal{S}_N$ being bounded,

$$\mathcal{S}_N x_m \xrightarrow[m\to\infty]{} \mathcal{S}_N x \quad \text{in } X.$$

However for any m, the vector $\mathcal{S}_N x_m$ belongs to $\mathrm{span}\{\psi_k\}_{k=1}^p$ which is closed (since finite dimensional), hence $\mathcal{S}_N x \in \mathrm{span}\{\psi_k\}_{k=1}^p$. As a consequence, $\mathrm{Ran}(\mathcal{S}_N) \subset \mathrm{span}\{\psi_k\}_{k=1}^p$. Thus,

$$\mathrm{Ran}(\mathcal{S}_N) = \mathrm{span}\{\psi_k\}_{k=1}^p$$

and

$$\mathrm{rank}(\mathcal{S}_N) = p$$

the eigenvectors $\{\psi_k\}_{k=1}^\infty$ being independent in X. We have seen in the Section 3.1, that the rank of $\mathcal{S}_N$ is also equal to the dimension of the subspace $\mathcal{V}_N$, which is denoted by $d(N)$. As a consequence, $p = d(N)$ and

$$\mathcal{V}_N = \mathcal{S}_N(X) = \mathrm{span}\{\psi_k\}_{k=1}^{d(N)}.$$

Rewriting of the eigenvalues $\{\lambda_k\}_{k=1}^\infty$ as

$$\begin{aligned} \lambda_k &= \langle \lambda_k \psi_k \, , \psi_k \rangle_X \\ &= \langle \mathcal{S}_N \psi_k \, , \varphi_k \rangle_X \\ &= \Big\langle \sum_{j=1}^N \alpha_j \langle \psi_k \, , z_j \rangle_X z_j \, , \psi_k \Big\rangle_X \\ &= \sum_{j=1}^N \alpha_j |\langle \psi_k \, , z_j \rangle_X|^2, \end{aligned} \tag{11}$$

will be useful in what follows. We will now show that the set

$$\boldsymbol{\beta} \overset{\text{def}}{=} \{\psi_k\}_{k=1}^\ell$$

is a discrete POD basis. Since we already know that $\boldsymbol{\beta}$ is an orthonormal set in X, for any $k \in \{1, \dots, \ell\}$ the constraints $c_k(\psi_1, \dots, \psi_\ell) = 0$ are satisfied.

Hence it remains to check that for any other orthonormal set $\mathcal{B} \overset{\text{def}}{=} \{\phi_k\}_{k=1}^{\ell}$ in $\mathcal{V}_N$,

$$-\sum_{j=1}^{N} \alpha_j \|P^{\mathcal{B}} z_j\|_X \geq -\sum_{j=1}^{N} \alpha_j \|P^{\beta} z_j\|_X$$

We rewrite

$$-\sum_{j=1}^{N} \alpha_j \|P^{\mathcal{B}} z_j\|_X^2 = -\sum_{j=1}^{N} \alpha_j \sum_{k=1}^{\ell} |\langle z_j \,, \phi_k \rangle_X|^2 = \sum_{k=1}^{\ell} \Big(-\sum_{j=1}^{N} \alpha_j |\langle z_j \,, \phi_k \rangle_X|^2 \Big),$$

and study the term $\sum_{j=1}^{N} \alpha_j |\langle z_j \,, \phi_k \rangle_X|^2$ of the above equation. Let $k \in \{1, \ldots, \ell\}$, since ϕ_k lies in $\mathcal{V}_N$, it may be rewritten as

$$\phi_k = \sum_{m=1}^{d(N)} \langle \phi_k, \psi_m \rangle_X \psi_m,$$

and its norm, which is equal to one, rewrites as

$$\|\phi_k\|_X^2 = \sum_{m=1}^{d(N)} |\langle \phi_k, \psi_m \rangle_X|^2.$$

Using this identity and the fact that $\{\psi_k\}_{k=1}^{\ell}$ are eigenvectors of $\mathcal{S}_N$, we compute

$$\begin{aligned} \sum_{j=1}^{N} \alpha_j |\langle z_j \,, \phi_k \rangle_X|^2 &= \Big\langle \underbrace{\sum_{j=1}^{N} \alpha_j \langle \phi_k \,, z_j \rangle_X z_j}_{=\mathcal{S}_N \phi_k} \,, \phi_k \Big\rangle_X \\ &= \Big\langle \sum_{m=1}^{d(N)} \langle \phi_k \,, \psi_m \rangle_X \mathcal{S}_N \psi_m \,, \phi_k \Big\rangle_X = \sum_{m=1}^{d(N)} \lambda_m |\langle \phi_k \,, \psi_m \rangle_X|^2. \end{aligned}$$

We want to bound $-\sum_{j=1}^{N} \alpha_j \sum_{k=1}^{\ell} |\langle z_j \,, \phi_k \rangle_X|^2$ from bellow by a similar term involving the vectors ψ_k instead of the vectors ϕ_k. For this purpose, we the

identity (11) and the fact that ϕ_k is of norm one:

$$
\begin{aligned}
-\sum_{j=1}^{N}\alpha_j|\langle z_j\,,\phi_k\rangle_X|^2 &= -\lambda_\ell\|\phi_k\|_X-\sum_{m=1}^{d(N)}\lambda_m|\langle\phi_k\,,\psi_m\rangle_X|^2+\lambda_\ell\|\phi_k\|_X\\
&= -\lambda_\ell-\Big[\sum_{m=1}^{\ell}\lambda_m|\langle\phi_k\,,\psi_m\rangle_X|^2\Big]-\Big[\sum_{m=\ell+1}^{d(N)}\lambda_m|\langle\phi_k\,,\psi_m\rangle_X|^2\Big]\\
&\quad+\lambda_\ell\Big(\Big[\sum_{m=1}^{\ell}|\langle\phi_k,\psi_m\rangle_X|^2\Big]+\Big[\sum_{m=\ell+1}^{d(N)}|\langle\phi_k,\psi_m\rangle_X|^2\Big]\Big)\\
&= -\lambda_\ell+\Big[\sum_{m=1}^{\ell}(\lambda_\ell-\lambda_m)|\langle\phi_k\,,\psi_m\rangle_X|^2\Big]\\
&\quad+\underbrace{\Big[\sum_{m=\ell+1}^{d(N)}(\lambda_\ell-\lambda_m)|\langle\phi_k\,,\psi_m\rangle_X|^2\Big]}_{\geq\,0}.
\end{aligned}
$$

The set $\{\phi_k\}_{k=1}^{\ell}$ being orthonormal in X, we may use Bessel's inequality $\sum_{k=1}^{\ell}|\langle\psi_k\,,\phi_k\rangle_X|^2\leq\|\psi_k\|_X$ (see [**19**, Chapter 4, Corollary p.84]). The vectors $\{\psi_m\}_{m=1}^{\ell}$ being of norm one, we deduce the following lower bound:

$$
\begin{aligned}
-\sum_{j=1}^{N}\alpha_j\|P^{\mathcal{B}}z_j\|_X &= -\sum_{j=1}^{N}\alpha_j\sum_{k=1}^{\ell}|\langle z_j\,,\phi_k\rangle_X|^2\\
&\geq\Big(-\sum_{k=1}^{\ell}\lambda_\ell\Big)+\sum_{m=1}^{\ell}\underbrace{(\lambda_\ell-\lambda_m)}_{\leq 0}\sum_{k=1}^{\ell}|\langle\phi_k\,,\psi_m\rangle_X|^2\\
&\geq\Big(-\sum_{k=1}^{\ell}\lambda_\ell\Big)+\sum_{m=1}^{\ell}(\lambda_\ell-\lambda_m)\;=\;-\sum_{k=1}^{\ell}\lambda_k.
\end{aligned}
$$

Hence, from (11), we deduce that

$$
-\sum_{j=1}^{N}\alpha_j\|P^{\mathcal{B}}z_j\|_X\geq-\sum_{j=1}^{N}\alpha_j\sum_{k=1}^{\ell}|\langle\psi_k\,,z_j\rangle_X|^2\;=\;-\sum_{j=1}^{N}\alpha_j\|P^{\boldsymbol{\beta}}z_j\|_X.
$$

Thus, $\boldsymbol{\beta}$ is discrete POD basis of rank ℓ, furthermore

$$
-\sum_{j=1}^{N}\alpha_j\|P^{\boldsymbol{\beta}}z_j\|_X=-\sum_{k=1}^{\ell}\lambda_k.
$$

At the beginning of the section, we have seen that the problems ($\mathrm{P}_{N,1}$) and ($\mathrm{P}_{N,2}$) have the same solutions and that

$$\sum_{j=1}^{N} \alpha_j \|z_j - P^{\boldsymbol{\beta}} z_j\|_X^2 = \sum_{j=1}^{N} \alpha_j \|z_j\|_X^2 - \sum_{j=1}^{N} \alpha_j \|P^{\boldsymbol{\beta}} z_j\|_X$$

(see Lemma 3.7). Since z_j belongs to $\mathrm{span}\{\psi_k\}_{k=1}^{d(N)}$, its norm rewrites as $\|z_j\|_X^2 = \sum_{n=1}^{d(N)} |\langle z_j \,, \psi_k \rangle_X|^2$. Consequently, using again the identity (11), we deduce that

$$\sum_{j=1}^{N} \alpha_j \|z_j - P^{\boldsymbol{\beta}} z_j\|_X^2 = \Big(\sum_{n=1}^{d(N)} \lambda_n \Big) - \Big(\sum_{k=1}^{\ell} \lambda_k \Big) = \sum_{m=\ell+1}^{d(N)} \lambda_m,$$

which concludes the proof. □

3.3. Computing the discrete POD basis. — We know, from Theorem 3.10, the eigenvectors associated to the first ℓ largest eigenvalues of $\mathcal{S}_N$ form a discrete POD basis. Moreover, choosing $\ell \leq d(N)$ ensures that these eigenvalues are positive. Here, we present another way of computing the discrete POD basis $\boldsymbol{\beta}$, than solving ($\mathrm{Eig}_{1,N}$), which is sometimes called *the method of snapshots*. We will be led to consider the following eigenvalue problem

$$\mathcal{K}_N v_k = \lambda_k v_k, \quad \text{for } 1 \leq k \leq \ell, \tag{$\mathrm{Eig}_{N,2}$}$$

where we denote $v_k = \big(v_k^1,\ v_k^2,\ \dots\ ,\ v_k^N\big)^T$.

Lemma 3.11. — *The operators $\mathcal{S}_N$ and $\mathcal{K}_N$ have the same nonzero eigenvalues, with the same multiplicities.*

Proof. — Recall that $\mathcal{S}_N = \mathcal{Y}_N \mathcal{Y}_N^*$ and $\mathcal{K}_N = \mathcal{Y}_N^* \mathcal{Y}_N$. Let $\lambda > 0$ be an eigenvalue of $\mathcal{S}_N$ and φ an associated eigenvector, meaning that

$$\mathcal{Y}_N \mathcal{Y}_N^* \varphi = \lambda \varphi.$$

Applying $\mathcal{Y}_N^*$ on both sides yields

$$\mathcal{K}_N \mathcal{Y}_N^* \varphi = \lambda \mathcal{Y}_N^* \varphi.$$

Since $\mathcal{Y}_N^* \varphi$ cannot be equal to zero without implying the following contradiction

$$\varphi = \frac{1}{\lambda} \mathcal{Y}_N \mathcal{Y}_N^* \varphi = 0, \tag{12}$$

we deduce that $\mathcal{Y}_N^* \varphi$ is an eigenvector of $\mathcal{K}_N$ associated to λ. Since $\mathcal{S}_N$ is compact, $\lambda > 0$ has a finite multiplicity. Denote by m the multiplicity of λ and let $\{\varphi_k\}_{k=1}^m$ be an orthonormal basis of $\ker(\lambda I - \mathcal{S}_N)$. We now know that

$\{\mathcal{Y}_N^*\varphi_k\}_{k=1}^m$ are eigenvectors of $\mathcal{K}_N$ associated to λ. We will show that they are independent in $\mathbb{C}^N$. For some $a_1, a_2, \ldots, a_m \in \mathbb{C}$, assume that

$$\sum_{k=1}^{m} a_k \mathcal{Y}_N^* \varphi_k = 0.$$

This means that

$$\Big\langle \sum_{k=1}^{m} a_k \mathcal{Y}_N^* \varphi_k \, , v \Big\rangle_X = 0$$

holds for any $v \in \mathbb{C}^N$, or equivalently,

$$\sum_{k=1}^{m} a_k \langle \varphi_k \, , \mathcal{Y}_N v \rangle_{\mathbb{C}^N} = 0.$$

Hence, for any $x \in \mathrm{Ran}(\mathcal{Y}_N)$, we have

$$\sum_{k=1}^{m} a_k \langle \varphi_k \, , x \rangle_{\mathbb{C}^N} = 0.$$

However, we know by (12) that $\{\varphi_k\}_{k=1}^m \subset \mathrm{Ran}(\mathcal{Y}_N)$. Consequently, the equation above holds for $x \in \{\varphi_k\}_{k=1}^m$ and having chosen $\{\varphi_k\}_{k=1}^m$ orthonormal, we deduce that $a_1 = \ldots = a_m = 0$. Hence, λ is an eigenvalue of $\mathcal{K}_N$ of multiplicity at least m. We may follow similar arguments to show that any positive eigenvalue of $\mathcal{K}_N$ of multiplicity m is also an eigenvalue of $\mathcal{S}_N$ of multiplicity at least m. As a consequence both operators have the same positive eigenvalues with the same multiplicities. □

Remark 3.12. — Hence, the first ℓ largest eigenvalues of $\mathcal{S}_N$ are equal to the first ℓ largest eigenvalues of $\mathcal{K}_N$.

Theorem 3.13. — *A discrete POD basis may be determined by following the steps:*

1. *Solve the problem* $(\mathrm{Eig}_{N,2})$ *to determine the positive eigenvalues* $\{\lambda_k\}_{k=1}^{\ell}$ *and the coefficients* $\{v_k^j \colon 1 \le k \le \ell, \ 1 \le j \le N\}$,
2. *For* $k \in \{1, \ldots, \ell\}$, *set*

$$\psi_k \overset{def}{=} \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{N} \sqrt{\alpha_j} v_k^j z_j.$$

Then, $\boldsymbol{\beta} \overset{def}{=} \{\psi_k\}_{k=1}^{\ell}$ *is a discrete POD basis of rank* ℓ.

Proof. — For $k \in \{1, \dots, \ell\}$, the eigenvectors ψ_k of $\mathcal{S}_N$ have the form

$$\psi_k = \frac{1}{\lambda_k}\mathcal{S}_N\psi_k = \frac{1}{\lambda_k}\sum_{j=1}^{N}\alpha_j\langle\psi_k\,,z_j\rangle_X z_j.$$

Whence, for $k \in \{1, \dots, \ell\}$, we will look for the vectors $\{\psi_k\}_{k=1}^{\ell}$ in the form

$$\psi_k = \Theta_k\sum_{j=1}^{N}\sqrt{\alpha_j}v_k^j z_j, \tag{13}$$

and we will look for conditions on v_k sufficient to imply that $\{\psi_k\}_{k=1}^{\ell}$ is a discrete POD basis (we know that it would be sufficient to show that $\{\psi_k\}_{k=1}^{\ell}$ are orthonormal eigenvectors of $\mathcal{S}_N$ associated to its ℓ largest eigenvalues). The constants $\{\Theta_k\}_{k=1}^{\ell} \subset \mathbb{R}$ will be determined so that $\|\psi_k\|_X = 1$. First, for $k \in \{1, \dots, \ell\}$, injecting (13) in the problem $(\mathrm{Eig}_{1,N})$ yields

$$\sum_{j=1}^{N}\alpha_j\Big\langle\Theta_k\sum_{m=1}^{N}\sqrt{\alpha_m}v_k^m z_m\,,z_j\Big\rangle_X z_j = \lambda_k\Theta_k\sum_{j=1}^{N}\sqrt{\alpha_j}v_k^j z_j$$

or equivalently

$$\Theta_k\sum_{j=1}^{N}\sqrt{\alpha_j}\Big(\sqrt{\alpha_j}\sum_{m=1}^{N}\sqrt{\alpha_m}\langle z_m, z_j\rangle_X v_k^m - \lambda_k v_k^m\Big)z_j = 0.$$

A sufficient condition for ψ_k to satisfy this equation is that

$$\sum_{m=1}^{N}\sqrt{\alpha_j\alpha_m}\overline{\langle z_j\,,z_m\rangle}_X v_k^m = \lambda_k v_k^m, \quad \text{for } 1 \le j \le N,$$

which may be rewritten, using the operator $\mathcal{K}_N$ introduced before, as

$$\mathcal{K}_N v_k = \lambda_k v_k.$$

Consequently, we look for $\{\psi_k\}_{k=1}^{\ell}$ satisfying the problem $(\mathrm{Eig}_{N,2})$. Since $\mathcal{K}_N$ is self-adjoint and compact, it admits N orthonormal eigenvectors $\{v_k\}_{k=1}^{N} \subset \mathbb{C}^N$ with nonnegative associated eigenvalues $\{\lambda_k\}_{k=1}^{\ell}$ that we may assume ordered decreasingly. Thus, if we use these eigenvectors and eigenvalues to define ψ_k for $k \in \{1, \dots, \ell\}$, the set $\{\psi_k\}_{k=1}^{\ell}$ must satisfy $(\mathrm{Eig}_{1,N})$. Lemma 3.11 ensures that $\{\lambda_k\}_{k=1}^{\ell}$ are the ℓ largest eigenvalues of $\mathcal{S}_N$. The set $\{v_k\}_{k=1}^{N}$ being orthonormal

in $\mathbb{C}^N$, so are the vectors ψ_k, since

$$\begin{aligned}\langle \psi_k \, , \psi_m \rangle_X &= \Big\langle \Theta_k \sum_{j=1}^{N} \sqrt{\alpha_j} v_k^j z_j \, , \Theta_m \sum_{n=1}^{N} \sqrt{\alpha_n} v_m^n z_n \Big\rangle_X \\ &= \Theta_k \Theta_m \sum_{n=1}^{N} \underbrace{\Big(\sum_{j=1}^{N} \sqrt{\alpha_n \alpha_j} \overline{\langle z_n \, , z_j \rangle}_X v_k^j \Big)}_{= \big(\mathcal{K}_N v_k \big)_n} \overline{v_m^n} \\ &= \Theta_k \Theta_m \langle \mathcal{K}_N v_k \, , v_m \rangle_{\mathbb{C}^N} \\ &= \Theta_k \Theta_m \lambda_k \langle v_m \, , v_k \rangle_{\mathbb{C}^N} \ = \ \delta_{k,m},\end{aligned}$$

if we set $\Theta_k \stackrel{\text{def}}{=} \frac{1}{\sqrt{\lambda_k}}$ for $k \in \{1, \ldots, \ell\}$. □

According to [**13**, p.10], no general a priori rules are available to choose ℓ. It is rather based on heuristic considerations and the observation of the ratio

$$\frac{\sum_{k=1}^{\ell} \lambda_k}{\sum_{k=1}^{d(N)} \lambda_k},$$

which, by (11), is also equal to

$$\frac{\sum_{k=1}^{\ell} \lambda_k}{\sum_{j=1}^{N} \alpha_k \|z_j\|_X^2}.$$

3.4. The case of a finite dimensional state space X. — Assume that the dimension n of X is finite and is greater than the number N of elements z_k. When $X = \mathbb{C}^n$, the operator $\mathcal{Y}_N \colon \mathbb{C}^N \longrightarrow \mathbb{C}^n$ may be represented by the matrix in $\mathbb{C}^{n \times N}$ whose columns are the elements $\big\{ \sqrt{\alpha_j} z_j \big\}_{j=1}^{N}$, since

$$\mathcal{Y}_N v = \sum_{j=1}^{N} \sqrt{\alpha_j} z_j v_j = \begin{pmatrix} | & | & & | \\ \sqrt{\alpha_1} z_1 & \sqrt{\alpha_2} z_2 & \ldots & \sqrt{\alpha_N} z_N \\ | & | & & | \end{pmatrix} v$$

holds for any $v \in \mathbb{C}^N$. let the singular value decomposition of $\mathcal{Y}_N$ be given by

$$\mathcal{Y}_N = U \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_N \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{pmatrix} V^*,$$

where

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{d(N)} > \sigma_{d(N)+1} = \ldots = \sigma_N = 0$$

(see section I). Since $U \in \mathbb{C}^{n\times n}$ and $V \in \mathbb{C}^{N\times N}$ are unitary, their columns denoted by $\{u_k\}_{k=1}^n$ and $\{v_k\}_{k=1}^N$ respectively, satisfy

$$\mathcal{Y}_N v_k = \sigma_k u_k, \quad \mathcal{Y}_N^* u_k = \sigma_k v_k \tag{14}$$

and

$$\mathcal{S}_N u_k = \mathcal{Y}_N \mathcal{Y}_N^* u_k = \sigma_k^2 u_k, \quad \mathcal{K}_N v_k = \mathcal{Y}_N^* \mathcal{Y}_N v_k = \sigma_k^2 v_k,$$

for any $k \in \{1, \ldots, N\}$. In this context, $\mathcal{S}_N \in \mathbb{C}^{n\times n}$ and $\mathcal{K}_N \in \mathbb{C}^{N\times N}$. If mN and pN can be chosen much smaller than than n, then solving the eigenvalue problem $\mathcal{K}_N v_k = \sigma_k^2 v_k$ will have a lower computational cost. Recall that ℓ is chosen smaller than $d(N) = \operatorname{rank}(\mathcal{Y}_N)$. Using (14), observe that once $\{v_k\}_{k=1}^\ell$ have been computed, the eigenvectors u_k of $\mathcal{S}_N$ associated with its ℓ largest eigenvalues σ_k^2 are given by

$$u_k = \frac{1}{\sigma_k} \mathcal{Y}_N v_k,$$

as it is also stated in Theorem 3.13. Then, $\boldsymbol{\beta} \overset{\text{def}}{=} \{u_k\}_{k=1}^\ell$ is a discrete POD basis of rank ℓ.

4. Continuous POD

In this section, we link the discrete POD problem to the so-called *continuous POD problem*. This link is first studied in the context of the model reduction of evolution equations, then for parameter-dependent elliptic equations.

4.1. Evolution equations. — Assume that $z \in C([0,T];X)$ is be the solution of an evolution equation. For example, we may consider the following initial/boundary-value problem, for Ω an open bounded subset of $\mathbb{R}^n$, with $\partial\Omega$ smooth, and $T > 0$

$$\begin{cases} \dfrac{\partial}{\partial t} z + Lz = f & \text{in } (0,T) \times \Omega \\ u = 0 & \text{on } [0,T] \times \partial\Omega \\ u = g & \text{on } \{t = 0\} \times \Omega, \end{cases}$$

where L denotes for each time t a second-order partial differential operator. See for example [**11**, Theorem 3, p.356 and Theorem 5, p.360-361] for information on the existence and regularity of the solution.

Let the interval $[0,T]$ be divided in $N-1$ subintervals by defining N time instances $\{t_j\}_{j=1}^N \subset [0,T]$. If the time instances are equidistant, and $h_t \overset{\text{def}}{=} \frac{T}{N-1}$ is the distance between two consecutive time instances, the interval has the form

Let the vectors $\{z_j\}_{j=1}^N$ be defined by

$$z_j \overset{\text{def}}{=} z(t_j).$$

In this context, studying the *continuous POD problem* and its relation to the discrete POD problem helps to understand how to choose the time instances t_j and the coefficients α_j.

4.1.1. *The problem and the governing operators.* — Let $\mathcal{V}$ be the closed subspace of X, endowed with the inner product $\langle \cdot\,, \cdot \rangle_X$, defined by

$$\mathcal{V} \overset{\text{def}}{=} \operatorname{span}\{z(t) \colon t \in [0,T]\},$$

of dimension $d \leq +\infty$. For an orthonormal basis $\{\psi_k\}_{k=1}^d$ of $\mathcal{V}$, we can rewrite $z(t)$, for $t \in [0,T]$, as

$$z(t) = \sum_{k=1}^{d} \langle z(t)\,, \psi_k \rangle_X \psi_k. \tag{15}$$

***Definition 4.1* (Continuous POD basis).** — The *continuous POD basis of rank* ℓ is the set $\boldsymbol{\beta}^{\infty} \overset{\text{def}}{=} \{\psi_k\}_{k=1}^{\ell}$ solution to

$$(\mathrm{P}_1) \qquad \begin{cases} \arg\min \Big\{ \displaystyle\int_0^T \big\| z(t) - P^{\mathcal{B}} z(t) \big\|_X^2 dt \,:\, \mathcal{B} \overset{\text{def}}{=} \{\phi_k\}_{k=1}^{\ell} \subset \mathcal{V} \Big\} \\ \text{subject to } \big\langle \phi_i \,,\, \phi_j \big\rangle_X = \delta_{i,j} \quad \text{for } 1 \leq i,j \leq \ell. \end{cases}$$

In other words, $\boldsymbol{\beta}^{\infty}$ is the orthonormal set (in $\mathcal{V}$) which minimizes the mean squared error between the elements $z(t)$, for $t \in [0,T]$, and their corresponding orthogonal projection on subspaces of $\mathcal{V}$.

As before, we first define some operators which will be used in what follows.

***Definition 4.2* ($\mathcal{Y}$).** — Let $\mathcal{Y}\colon L^2(0,T;\mathbb{C}) \longrightarrow X$ be the linear operator defined by

$$\mathcal{Y} v \overset{\text{def}}{=} \int_0^T v(t) z(t) dt.$$

Using properties of the Bochner integral (see [**7**, Theorem 4, p.46]) and the Cauchy-Schwarz inequality, we observe that $\mathcal{Y}$ is bounded, as

$$\big\| \mathcal{Y} v \big\|_X \leq \int_0^T |v(t)| \|z(t)\|_X dt \leq \|v\|_{L^2(0,T;\mathbb{C})} \underbrace{\|z\|_{L^2(0,T;X)}}_{<\infty}$$

holds for any $v \in L^2(0,T;X)$. It's adjoint $\mathcal{Y}^*$ consequently lies in $\mathscr{L}(X, L^2(0,T;\mathbb{C}))$ and is given by

$$\mathcal{Y}^* v(t) = \langle w \,,\, z(t) \rangle_X, \quad \forall t \in [0,T].$$

In effect, for any $v \in L^2(0,T;\mathbb{C})$, $w \in X$,

$$\langle \mathcal{Y} v \,,\, w \rangle_X = \Big\langle \int_0^T v(t) z(t) dt \,,\, w \Big\rangle_X = \int_0^T v(t) \langle z(t) \,,\, w \rangle_X dt = \langle v \,,\, \mathcal{Y}^* w \rangle_{L^2(0,T;\mathbb{C})}.$$

For the second equality above, the integral and inner product may switch since the map $\mathscr{T}\colon \phi \longmapsto \langle \phi \,,\, w \rangle_X$ belongs to $\mathscr{L}(X,\mathbb{C})$ (as a consequence, $\mathscr{T}$ is closed and $\mathscr{T}\big(v(\cdot) z(\cdot)\big) = v(\cdot) \big\langle z(\cdot) \,,\, w \big\rangle_X$ lies in $L^2(0,T;\mathbb{C})$), see [**7**, Theorem 6, p.47]).

Lemma 4.3. — *The operator $\mathcal{Y}^*$ is compact from X to $L^2(0,T;\mathbb{C})$.*

Proof. — Denote by B_X the unit open ball in X, meaning

$$B_X = \{x \in X \colon \|x\|_X < 1\}.$$

We will show that the closure of $\mathcal{Y}^* B_X$ in $L^2(0,T;\mathbb{C})$ is a compact set, by using a criterion given by Kolmogorov for compact sets in L^p spaces (see [**1**, Theorem 31.20, p.279]). The set $\mathrm{clos}(\mathcal{Y}^* B_X)$ being closed and bounded (since $\mathcal{Y}^*$ is a

bounded operator), the criterion in question states that this set is compact for the L^2 norm if and only if for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\|f - f_h\|_{L^2(0,T;\mathbb{C})} < \varepsilon \text{ holds } \forall f \in \text{clos}(\mathcal{Y}^* B_X) \text{ and } \forall\, 0 < h < \delta,$$

where f_h is defined by

$$f_h(t) \overset{\text{def}}{=} \frac{1}{2h} \int_{t-h}^{t+h} f(s)ds,$$

for any element f of $L^2(0,T;\mathbb{C})$, considering that $f(t) = 0$ for $t \notin [0,T]$. Moreover, the inequality

$$\|f_h\|_{L^2(0,T;\mathbb{C})} \leq \|f\|_{L^2(0,T;\mathbb{C})} \tag{16}$$

holds for any $f \in L^2(0,T;\mathbb{C})$ (see [**1**, Lemma 31.19, p.268]). Let $\varepsilon > 0$ and $f \in \text{clos}(\mathcal{Y}^* B_X)$. There exists $f^\varepsilon \in \mathcal{Y}^* B_X$ such that

$$\|f - f^\varepsilon\|_{L^2(0,T;\mathbb{C})} < \varepsilon.$$

moreover, $f^\varepsilon = \mathcal{Y}^* g^\varepsilon$ for some $g^\varepsilon \in B_X$. We begin by showing that $f^\varepsilon - f_h^\varepsilon$ can be arbitrary small for h small enough. For $h > 0$,

$$\begin{aligned}
\left\|f^\varepsilon - f_h^\varepsilon\right\|^2_{L^2(0,T;\mathbb{C})} &= \int_0^T \left| f^\varepsilon(t) - \frac{1}{2h} \int_{t-h}^{t+h} f^\varepsilon(s)ds \right|^2 dt \\
&= \int_0^T \left| \langle g^\varepsilon, z(t)\rangle_X - \frac{1}{2h} \int_{t-h}^{t+h} \langle g^\varepsilon, z(s)\rangle_X ds \right|^2 dt
\end{aligned}$$

Integral and inner product switch for the same reason as before, hence

$$\begin{aligned}
\left\|f^\varepsilon - f_h^\varepsilon\right\|^2_{L^2(0,T;\mathbb{C})} &= \int_0^T \left| \langle g^\varepsilon, z(t)\rangle_X - \frac{1}{2h} \Big\langle g^\varepsilon, \int_{t-h}^{t+h} z(s)ds \Big\rangle_X \right|^2 dt \\
&= \int_0^T \left| \Big\langle g^\varepsilon, \frac{1}{2h} \int_{t-h}^{t+h} (z(t) - z(s))ds \Big\rangle_X \right|^2 dt \\
&\leq T \|g^\varepsilon\|_X^2 \max_{t\in[0,T]} \left\| \frac{1}{2h} \int_{t-h}^{t+h} (z(t) - z(s))ds \right\|_X^2 \\
&\leq T \max_{t\in[0,T]} \Big(\frac{1}{2h} \int_{t-h}^{t+h} \|z(t) - z(s)\|_X ds \Big)^2 \\
&\leq T \max_{t\in[0,T]} \Big(\max_{s\in[t-h,t+h]} \|z(t) - z(s)\|_X^2 \Big)
\end{aligned}$$

As z is continuous from the compact set $[0,T]$ to X, it is uniformly continuous. Consequently there exists $\delta > 0$ (notice that δ is independent of f) such that for any $t, s \in [0,T]$,

$$-\delta \leq t - s \leq \delta \quad \Rightarrow \quad \|z(t) - z(s)\|_X < \varepsilon.$$

As a consequence, for any $0 < h < \delta$, if $-h \leq t-s \leq h$, then $\|z(t)-z(s)\|_X < \varepsilon$. We can also write this as

$$\max_{t\in[0,T]} \max_{-h\leq t-s\leq h} \|z(t)-z(s)\|_X \leq \varepsilon, \quad \forall 0 < h < \delta.$$

Going back to estimating $f - f_h$, in the light of (16) and the previous estimate, we see that

$$\begin{aligned}\|f - f_h\|_X &\leq \|f - f^\varepsilon\|_X + \|f^\varepsilon - f_h^\varepsilon\|_X + \|(f^\varepsilon - f)_h\|_X \\ &\leq 2\|f - f^\varepsilon\|_X + \|f^\varepsilon - f_h^\varepsilon\|_X \\ &\leq 2\varepsilon + \sqrt{T}\varepsilon\end{aligned}$$

holds for any $0 < h < \delta$. □

***Definition 4.4* ($\mathcal{S}$).** — Let $\mathcal{S}\colon X \to X$ be the linear operator defined by

$$\mathcal{S}w \stackrel{\text{def}}{=} \int_0^T \langle w\,, z(t)\rangle_X z(t)dt.$$

This operator can be decomposed as $\mathcal{S} = \mathcal{Y}\mathcal{Y}^*$, since

$$\mathcal{S}w = \int_0^T \langle w\,, z(t)\rangle_X z(t)dt = \int_0^T (\mathcal{Y}^*w)(t)z(t)dt = \mathcal{Y}\mathcal{Y}^*w$$

holds for any $w \in X$. Moreover, $\mathcal{S}$ is bounded (composition of two bounded operators), self-adjoint (since $(\mathcal{Y}\mathcal{Y}^*)^* = \mathcal{Y}^{**}\mathcal{Y}^* = \mathcal{Y}\mathcal{Y}^*$), positive since

$$\langle \mathcal{S}w\,, w\rangle_X = \langle \mathcal{Y}\mathcal{Y}^*w\,, w\rangle_X = \langle \mathcal{Y}^*w\,, \mathcal{Y}^*w\rangle_{L^2(0,T;\mathbb{C})} = \|\mathcal{Y}^*w\|_X \geq 0,$$

and compact (composition of a bounded operator and a compact operator).

***Definition 4.5* ($\mathcal{K}$).** — Let $\mathcal{K}\colon L^2(0,T;\mathbb{C}) \to L^2(0,T;\mathbb{C})$ be the linear operator defined by $\mathcal{K} \stackrel{\text{def}}{=} \mathcal{Y}^*\mathcal{Y}$.

This operator is also bounded, self-adjoint, positive and compact (by the same arguments as for $\mathcal{S}$), and may be rewritten as

$$\mathcal{K}v(t) = \langle \mathcal{Y}v\,, z(t)\rangle_X = \Big\langle \int_0^T v(s)z(s)ds\,, z(t)\Big\rangle_X = \int_0^T \langle z(s)\,, z(t)\rangle_X v(s)ds.$$

Remark 4.6. — There are other ways of showing that $\mathcal{S}$ and $\mathcal{K}$ are compact operators. For example, one could first show that $\mathcal{K} = \mathcal{Y}^*\mathcal{Y}$ is compact by using that $z \in L^2(0,T;\mathbb{C})$ (see [**23**, Example 2 and Proof, p.277-278]). Secondly, one may deduce from this result that $\mathcal{Y}$ is also compact (hence $\mathcal{S}$ and $\mathcal{K}$ are compact).

In effect, we may show that the image by $\mathcal{Y}$ of any bounded sequence in X, admits a converging subsequence. Assuming that (x_k) is a bounded sequence

in $L^2(0,T;\mathbb{C})$, the compacity of $\mathcal{Y}^*\mathcal{Y}$ implies the existence of a subsequence $\big(x_{\varphi(k)}\big)$ such that $\big(\mathcal{Y}^*\mathcal{Y}x_{\varphi(k)}\big)$ converges in $L^2(0,T;\mathbb{C})$. We denote my M the bound

$$M \stackrel{\text{def}}{=} \sup_{k\in\mathbb{N}} \|x_k\|_{L^2(0,T;\mathbb{C})}.$$

The Cauchy-Schwarz inequality yields, for any $k,j\in\mathbb{N}$,

$$\begin{aligned}\|\mathcal{Y}x_{\varphi(k)} - \mathcal{Y}x_{\varphi(j)}\|_X &= \big\langle \mathcal{Y}^*\mathcal{Y}x_{\varphi(k)} - \mathcal{Y}^*\mathcal{Y}x_{\varphi(j)}\,, x_{\varphi(k)} - x_{\varphi(j)}\big\rangle_{L^2(0,T;\mathbb{C})}\\ &\le \big\|\mathcal{Y}^*\mathcal{Y}x_{\varphi(k)} - \mathcal{Y}^*\mathcal{Y}x_{\varphi(j)}\big\|_{L^2(0,T;\mathbb{C})}\big\|x_{\varphi(k)} - x_{\varphi(j)}\big\|_{L^2(0,T;\mathbb{C})}\\ &\le 2M\big\|\mathcal{Y}^*\mathcal{Y}x_{\varphi(k)} - \mathcal{Y}^*\mathcal{Y}x_{\varphi(j)}\big\|_{L^2(0,T;\mathbb{C})} \xrightarrow[k,j\to\infty]{} 0.\end{aligned}$$

Hence, $\big(\mathcal{Y}x_{\varphi(k)}\big)$ is a Cauchy sequence of the Hilbert space X, and consequently converges in X.

4.1.2. *Existence of the continuous POD basis.* — Similarly to the discrete POD, (P_1) has the same solutions as the problem

$$(\mathrm{P}_2)\qquad \begin{cases} \arg\min\Big\{ -\displaystyle\int_0^T \|P^{\mathcal{B}}z(t)\|_X^2 dt \;:\; \mathcal{B}\stackrel{\text{def}}{=}\{\phi_k\}_{k=1}^{\ell}\subset\mathcal{V}\Big\}\\ \text{subject to } \big\langle \phi_i\,,\phi_j\big\rangle_X = \delta_{i,j} \quad \text{for } 1\le i,j\le \ell, \end{cases}$$

since, for any set $\mathcal{B}\stackrel{\text{def}}{=}\{\phi_k\}_{k=1}^{\ell}$ orthonormal in X, the equality

$$\|z(t) - P^{\mathcal{B}}z(t)\|_X^2 = \|z(t)\|_X^2 - \|P^{\mathcal{B}}z(t)\|_X^2$$

holds for any $t\in[0,T]$.

Proposition 4.7. — *Assume that $\boldsymbol{\beta}^{\infty}\stackrel{def}{=}\{\psi_k\}_{k=1}^{\ell}$ is a continuous POD basis of rank ℓ. Then, the components of $\boldsymbol{\beta}^{\infty}$ satisfy the following eigenvalue problem*

$$(\mathrm{Eig})\qquad \mathcal{S}\psi_k = \lambda_k\psi_k, \quad \textit{for } 1\le k\le \ell.$$

Proof. — For $k\in\{1,\dots,\ell\}$, define the maps e_k as in the proof of Proposition 3.8, in order to rewrite (P_2) as

$$(\mathrm{P}_3)\qquad \begin{cases} \arg\min\Big\{ -\displaystyle\int_0^T \big\|P^{\mathcal{B}}z(t)\big\|_X^2 dt \;:\; \mathcal{B}\stackrel{\text{def}}{=}\{\phi_k\}_{k=1}^{\ell}\subset\mathcal{V}\Big\}\\ \text{subject to } e_k(\psi_1,\dots,\psi_\ell) = 0, \quad \text{for } 1\le k\le \ell. \end{cases}$$

For this optimality problem, we consider the Lagrangian $L\colon X^\ell \times \mathbb{C}^\ell \times \mathbb{C}^{\ell-1} \times \ldots \times \mathbb{C} \to \mathbb{C}$ defined by

$$L(\varphi_1, \ldots, \varphi_\ell, \eta_1, \ldots, \eta_l) \stackrel{\text{def}}{=} - \int_0^T \sum_{k=1}^{\ell} |\langle z(t)\,, \varphi_k \rangle_X|^2 dt + \sum_{k=1}^{\ell} \big\langle e_k(\varphi_1, \ldots, \varphi_\ell)\,, \eta_k \big\rangle_{\mathbb{C}^{l-k+1}},$$

where $\eta_k \stackrel{\text{def}}{=} (\eta_k^1, \eta_k^2, \ldots, \eta_k^{l-k+1})^T$. If $\boldsymbol{\beta}^\infty \stackrel{\text{def}}{=} \{\psi_k\}_{k=1}^\ell$ is a solution of (P_3), then the first order optimality condition for this problem yields the existence of Lagrange multipliers $\{\eta_k\}_{k=1}^\ell \subset \mathbb{C}^\ell \times \mathbb{C}^{\ell-1} \times \ldots \times \mathbb{C}$ such that

$$\text{(OC)} \qquad \begin{cases} \dfrac{\partial L}{\partial \psi_m}(\psi_1, \ldots, \psi_\ell, \eta_1, \ldots, \eta_\ell) = 0 \quad \text{in } X \\ \text{subject to } e_m(\psi_1, \ldots, \psi_\ell) = 0 \quad \text{in } \mathbb{C}^{\ell-m+1} \end{cases} \qquad \text{for } 1 \le m \le \ell.$$

The first line of (OC) can be rewritten as

$$\frac{d}{d\varepsilon} L\big(\psi_1, \ldots, (\psi_m + \varepsilon \mathbf{v}), \ldots, \psi_\ell, \eta_1, \ldots, \eta_l\big)\Big|_{\varepsilon=0} = 0, \quad \forall \mathbf{v} \in X,\ 1 \le m \le \ell,$$

meaning that

$$\begin{aligned} \frac{d}{d\varepsilon}\Big[&- \Big(\int_0^T \sum_{\substack{k=1 \\ k\neq m}}^{\ell} |\langle z(t)\,, \psi_k \rangle_X|^2 dt \Big) + \Big(\int_0^T |\langle z(t)\,, \psi_m + \varepsilon \mathbf{v} \rangle_X|^2 dt \Big) \\ &+ \Big(\sum_{\substack{k=1 \\ k\neq m}}^{\ell} \big(\langle \psi_k\,, \psi_k \rangle_X - 1\big)\overline{\eta_k^1} \Big) + \big(\langle \psi_m + \varepsilon \mathbf{v}\,, \psi_m + \varepsilon \mathbf{v} \rangle_X - 1\big)\overline{\eta_m^1} \\ &+ \Big(\sum_{k=1}^{m-1} \sum_{\substack{q=k+1 \\ q\neq m}}^{\ell} \langle \psi_q\,, \psi_k \rangle_X \overline{\eta_k^{q-k+1}} \Big) + \Big(\sum_{k=1}^{m-1} \langle \psi_m + \varepsilon \mathbf{v}\,, \psi_k \rangle_X \overline{\eta_k^{m-k+1}} \Big) \\ &+ \Big(\sum_{q=m+1}^{\ell} \langle \psi_q\,, \psi_m + \varepsilon \mathbf{v} \rangle_X \overline{\eta_m^{q-m+1}} \Big) \\ &+ \Big(\sum_{k=m+1}^{\ell} \sum_{q=k+1}^{\ell} \langle \psi_q\,, \psi_k \rangle_X \overline{\eta_k^{q-k+1}} \Big) \Big]\Big|_{\varepsilon=0} = 0, \end{aligned} \tag{17}$$

The computations do not change from the discrete POD apart from those involving the first line of the above equation, for which we have

$$\begin{aligned}
&\frac{d}{d\varepsilon}\Big[-\Big(\int_0^T \sum_{\substack{k=1\\k\neq m}}^{\ell} |\langle z(t)\,,\psi_k\rangle_X|^2 dt\Big) + \Big(\int_0^T |\langle z(t)\,,\psi_m+\varepsilon\mathbf{v}\rangle_X|^2 dt\Big)\Big]\Big|_{\varepsilon=0}\\
&= \frac{d}{d\varepsilon}\Big[-\int_0^T \langle z(t)\,,\psi_m+\varepsilon\mathbf{v}\rangle_X\langle\psi_m+\varepsilon\mathbf{v}\,,z(t)\rangle_X dt\Big]\Big|_{\varepsilon=0}\\
&= \Big[-\int_0^T \frac{d}{d\varepsilon}\Big(\langle z(t)\,,\psi_m+\varepsilon\mathbf{v}\rangle_X\langle\psi_m+\varepsilon\mathbf{v}\,,z(t)\rangle_X\Big) dt\Big]\Big|_{\varepsilon=0}\\
&= \Big[-\int_0^T \Big(\langle z(t)\,,\mathbf{v}\rangle_X\langle\psi_m\,,z(t)\rangle_X + \langle z(t)\,,\psi_m\rangle_X\langle\mathbf{v}\,,z(t)\rangle_X + 2\varepsilon|\langle z(t)\,,\mathbf{v}\rangle_X|^2\Big) dt\Big]\Big|_{\varepsilon=0}\\
&= -\int_0^T \Big(\langle z(t)\,,\mathbf{v}\rangle_X\langle\psi_m\,,z(t)\rangle_X + \langle z(t)\,,\psi_m\rangle_X\langle\mathbf{v}\,,z(t)\rangle_X\Big) dt\\
&= 2\Re\Big(-\int_0^T \langle z(t)\,,\mathbf{v}\rangle_X\langle\psi_m\,,z(t)\rangle_X dt\Big).
\end{aligned}$$

Once again properties of the Bochner integral, and the fact that z lies in $L^2(0,T;X)$, allow the switch between integral and inner product, hence

$$\begin{aligned}
&\frac{d}{d\varepsilon}\Big[-\Big(\int_0^T \sum_{\substack{k=1\\k\neq m}}^{\ell} |\langle z(t)\,,\psi_k\rangle_X|^2 dt\Big) + \Big(\int_0^T |\langle z(t)\,,\psi_m+\varepsilon\mathbf{v}\rangle_X|^2 dt\Big)\Big]\Big|_{\varepsilon=0}\\
&\qquad = -2\Re\Big\langle \int_0^T \langle\psi_m\,,z(t)\rangle_X z(t) dt\,,\mathbf{v}\Big\rangle_X \;=\; -2\Re\langle\mathcal{S}\psi_m\,,\mathbf{v}\rangle_X.
\end{aligned}$$

Consequently (17) is equivalent to

$$\begin{aligned}
-2\Re\langle\mathcal{S}\psi_m\,,\mathbf{v}\rangle_X + \langle\mathbf{v}\,,\eta_m^1\psi_m\rangle_X + \Big\langle\mathbf{v}\,,\sum_{k=1}^{m-1}\eta_k^{m-k+1}\psi_k\Big\rangle_X\\
+\Big\langle\sum_{q=m+1}^{\ell}\overline{\eta_m^{q-m+1}}\psi_q\,,\mathbf{v}\Big\rangle_X = 0,
\end{aligned}$$

or, equivalently

$$\Re\Big\langle -2\mathcal{S}\psi_m + (\eta_m^1+\overline{\eta_m^1})\psi_m + \sum_{k=1}^{m-1}\eta_k^{m-k+1}\psi_k + \sum_{k=2}^{\ell-m+1}\overline{\eta_m^k\psi_{k+m-1}}\,,\;\mathbf{v}\Big\rangle_X = 0$$

and

$$\Im\Big\langle (\overline{\eta_m^1} - \eta_m^1)\psi_m - \sum_{k=1}^{m-1} \eta_k^{m-k+1}\psi_k + \sum_{q=m+1}^{\ell} \eta_m^{q-m+1}\psi_q \,,\ \mathbf{v}\Big\rangle_X = 0.$$

The same arguments as for discrete POD yield

$$\mathcal{S}\psi_m = \eta_m^1 \psi_m.$$

Thus, a solution $\boldsymbol{\beta}^\infty$ of (P_2) necessarily satisfies the problem (Eig). □

***Theorem 4.8* (Existence, continuous POD).** — *There exists eigenvalues* $\{\lambda_k\}_{k=1}^\infty$ *of* $\mathcal{S}$ *and associated orthonormal eigenvectors* $\{\psi_k\}_{k=1}^\infty$ *satisfying*

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d > 0 = \lambda_{d+1} = \lambda_{d+2} = \ldots, \quad \text{if } d < \infty,$$

and

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k \xrightarrow[k\to\infty]{} 0, \quad \text{if } d = +\infty.$$

For $\ell \leq d$*, the first* ℓ *eigenvectors* $\boldsymbol{\beta}^\infty \overset{def}{=} \{\psi_k\}_{k=1}^\ell$ *form a continuous POD basis of rank* ℓ*, and*

$$\int_0^T \left\| z(t) - P^{\boldsymbol{\beta}^\infty} z(t) \right\|_X^2 dt = \sum_{k=\ell+1}^{d} \lambda_k.$$

Proof. — The operator $\mathcal{S}$ is a self-adjoint compact operator from the separable Hilbert space X to itself, hence the Hilbert-Schmidt theorem asserts that there exists an orthonormal basis $\{\psi_k\}_{k=1}^\infty$ of X, such that

$$\mathcal{S}\psi_k = \lambda_k \psi_k \quad \text{and} \quad \lambda_k \xrightarrow[k\to\infty]{} 0. \tag{18}$$

and we may assume that the eigenvalues $\{\lambda_k\}_{k=1}^\infty$ are ordered decreasingly. We want to show that the set $\boldsymbol{\beta}^\infty \overset{\text{def}}{=} \{\psi_k\}_{k=1}^\ell$ is a continuous POD basis. The equation (18) allows to deduce the following identity for any $k \in \{1, \ldots, \ell\}$:

$$\begin{aligned} \lambda_k &= \big\langle \lambda_k \psi_k \,, \psi_k \big\rangle_X \\ &= \Big\langle \int_0^T \big\langle \psi_k \,, z(t) \big\rangle_X z(t) \,, \psi_k \Big\rangle_X \\ &= \int_0^T |\langle z(t) \,, \psi_k \rangle_X|^2, \end{aligned} \tag{19}$$

which also implies that

$$-\int_0^T \sum_{k=1}^{\ell} |\langle z(t) \,, \psi_k \rangle_X|^2 = -\sum_{k=1}^{\ell} \lambda_k.$$

Since we already know that $\boldsymbol{\beta}^\infty$ it is an orthonormal set in X, we only have to check that the inequality

$$-\int_0^T \sum_{k=1}^{\ell} |\langle z(t)\,,\phi_k\rangle_X|^2 dt \geq -\sum_{k=1}^{\ell} \lambda_k$$

holds for any orthonormal set $\{\phi_k\}_{k=1}^{\ell}$ in $\mathcal{V}$. The right-hand side of the above inequality rewrites as

$$-\int_0^T \sum_{k=1}^{\ell} |\langle z(t)\,,\phi_k\rangle_X|^2 dt = \sum_{k=1}^{\ell} \Big(-\int_0^T |\langle z(t)\,,\phi_k\rangle_X|^2 dt\Big)$$

and we first consider the term inside the parenthesis. Let $k \in \{1,\dots,\ell\}$, since ϕ_k lies in X, it may be rewritten as

$$\phi_k = \sum_{m=1}^{\infty} \langle \phi_k, \psi_m\rangle_X \psi_m. \tag{20}$$

Using (20) and the fact that $\{\psi_k\}_{k=1}^{\ell}$ are eigenvectors, we compute

$$\begin{aligned}
\int_0^T |\langle z(t)\,,\phi_k\rangle_X|^2 dt &= \int_0^T \big\langle \langle \phi_k\,, z(t)\rangle_X z(t)\,,\phi_k\big\rangle_X dt \\
&= \Big\langle \underbrace{\int_0^T \langle \phi_k\,, z(t)\rangle_X z(t)\, dt}_{=\mathcal{S}\phi_k}\,,\phi_k\Big\rangle_X \\
&= \Big\langle \mathcal{S}\sum_{m=1}^{\infty} \langle \phi_k\,,\psi_m\rangle_X \psi_m\,,\phi_k\Big\rangle_X \\
&= \Big\langle \sum_{m=1}^{\infty} \langle \phi_k\,,\psi_m\rangle_X \mathcal{S}\psi_m\,,\phi_k\Big\rangle_X \\
&= \sum_{m=1}^{\infty} \lambda_m |\langle \phi_k\,,\psi_m\rangle_X|^2.
\end{aligned}$$

As for the discrete POD, we make use of the identity (19) to obtain

$$\begin{aligned}
&-\int_0^T |\langle z(t)\,,\phi_k\rangle_X|^2 dt \\
&\qquad = -\lambda_\ell \|\phi_k\|_X - \sum_{m=1}^{\infty} \lambda_m |\langle \phi_k\,,\psi_m\rangle_X|^2 + \lambda_\ell \sum_{m=1}^{\infty} |\langle \phi_k, \psi_m\rangle_X|^2
\end{aligned}$$

$$= -\lambda_\ell + \sum_{m=1}^{\ell} (\lambda_\ell - \lambda_m) |\langle \phi_k \, , \psi_m \rangle_X|^2 + \underbrace{\sum_{m=\ell+1}^{\infty} (\lambda_\ell - \lambda_m) |\langle \phi_k \, , \psi_m \rangle_X|^2}_{\geq 0}.$$

As for the discrete POD, Bessel's inequality and the fact that the vectors $\{\psi_m\}_{m=1}^{\infty}$ are of norm one allows to deduce the following lower bound:

$$\begin{aligned} -\int_0^T \sum_{k=1}^{\ell} |\langle z(t) \, , \phi_k \rangle_X|^2 dt &\geq \Big(-\sum_{k=1}^{\ell} \lambda_\ell \Big) + \sum_{m=1}^{\ell} \underbrace{(\lambda_\ell - \lambda_m)}_{\leq 0} \underbrace{\sum_{k=1}^{\ell} |\langle \phi_k \, , \psi_m \rangle_X|^2}_{\leq \|\psi_m\|_X} \\ &\geq \Big(-\sum_{k=1}^{\ell} \lambda_\ell \Big) + \sum_{m=1}^{\ell} (\lambda_\ell - \lambda_m) \\ &= -\sum_{k=1}^{\ell} \lambda_k. \end{aligned}$$

Thus $\boldsymbol{\beta}^\infty$ is the POD basis, and

$$\begin{aligned} \int_0^T \big\| z(t) - \sum_{k=1}^{\ell} \langle z(t) \, , \psi_k \rangle_X \psi_k \big\|_X^2 dt &= \int_0^T \|z(t)\|_X^2 dt - \int_0^T \sum_{k=1}^{\ell} |\langle z(t) \, , \psi_k \rangle_X|^2 \\ &= \int_0^T \sum_{k=1}^{\infty} |\langle z(t) \, , \psi_k \rangle_X|^2 dt - \int_0^T \sum_{k=1}^{\ell} |\langle z(t) \, , \psi_k \rangle_X|^2 \\ &= \Big(\sum_{m=1}^{\infty} \lambda_m \Big) - \Big(\sum_{k=1}^{\ell} \lambda_k \Big) \\ &= \sum_{m=\ell+1}^{\infty} \lambda_m. \end{aligned}$$

□

The operator $\mathcal{K}$ is not studied in this part as it was done for $\mathcal{K}_N$ in the Section 3.3, however more information on this may be found for example in [**14**, p.498].

4.1.3. *Relating continuous and discrete POD.* — The proposition and the theorem that follow will permit to understand how the time instances t_j and the weights α_j can be chosen.

Proposition 4.9. — *Assume that z belongs to $H^1(0,T;X)$, and choose the time instances*

$$t_j \overset{def}{=} (j-1)\mathrm{h}_t, \quad \textit{where} \;\; \mathrm{h}_t \overset{def}{=} \frac{T}{N-1},$$

and the weights

$$\alpha_j = \begin{cases} \dfrac{\mathrm{h}_t}{2} & \text{if } j = 1 \text{ or } j = N, \\ \mathrm{h}_t & \text{if } j \in \{2, \dots, N-1\}. \end{cases}$$

Then the operator $\mathcal{S}_N$ converges uniformly to $\mathcal{S}$, meaning

$$\lim_{N\to\infty} \|\mathcal{S}_N - \mathcal{S}\| = 0.$$

Remark 4.10. — We know from [**11**, Section 5.9, Theorem 2, p.286], that $z \in H^1(0,T;X)$ implies that $z \in C([0,T];X)$.

Proof. — We follow the arguments of proof given by [**13**, Lemma 2.16, p.18]. We want to show that

$$\lim_{N\to\infty} \sup_{x\in X\setminus\{0\}} \frac{\|\mathcal{S}_N x - \mathcal{S}x\|_X}{\|x\|_X} = 0 \tag{21}$$

Let $x \in X$ and define the continuous function $f\colon [0,T] \to X$, which depends on x, by

$$f(t) \overset{\text{def}}{=} \langle x\,, z(t)\rangle_X z(t)$$

We can then rewrite $\mathcal{S}_N$ and $\mathcal{S}$ as

$$\mathcal{S}x = \int_0^T f(t)dt = \sum_{k=1}^{N-1} \int_{t_k}^{t_{k+1}} f(t)dt$$

and

$$\mathcal{S}_N x = \sum_{j=1}^{N} \alpha_j f(t_j) = \sum_{k=1}^{N-1} \frac{\Delta t}{2}\big(f(t_k) + f(t_{k+1})\big).$$

Then, we observe that their difference reads as

$$\|\mathcal{S}_N x - \mathcal{S}x\|_X = \Big\| \sum_{k=1}^{N-1} \Big(\frac{\Delta t}{2}\big(f(t_k) + f(t_{k+1})\big) - \int_{t_k}^{t_{k+1}} f(t)dt \Big) \Big\|_X.$$

As f is continuous, it also lies in $L^2(0,T;X)$ since

$$\|f\|^2_{L^2(0,T;X)} = \int_0^T \|f(t)\|_X^2 dt \le \Big(\max_{t\in[0,T]} \|z(t)\|_X^4 \Big) T\|x\|_X^2 < \infty.$$

The weak derivative of f is given by

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} f(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\Big(\big\langle x\,, z(\cdot)\big\rangle_X \Big)(t) z(t) + \big\langle x\,, z(t)\big\rangle_X \frac{\mathrm{d}}{\mathrm{d}t} z(t) \\ &= \big\langle x\,, \frac{\mathrm{d}}{\mathrm{d}t} z(t)\big\rangle_X z(t) + \big\langle x\,, z(t)\big\rangle_X \frac{\mathrm{d}}{\mathrm{d}t} z(t). \end{aligned}$$

Then, we have

$$\begin{aligned}\Big\|\frac{\mathrm{d}}{\mathrm{d}t}f\Big\|^2_{L^2(0,T;X)} &= \int_0^T \Big\|\big\langle x\,,\frac{\mathrm{d}}{\mathrm{d}t}z(t)\big\rangle_X z(t) + \big\langle x\,,z(t)\big\rangle_X \frac{\mathrm{d}}{\mathrm{d}t}z(t)\Big\|^2_X dt\\ &\le 4\int_0^T \Big\|\frac{\mathrm{d}}{\mathrm{d}t}z(t)\Big\|^2_X \|z(t)\|^2_X dt \|x\|^2_X\\ &\le \underbrace{4\Big(\max_{t\in[0,T]}\|z(t)\|^2_X\Big)\Big\|\frac{\mathrm{d}}{\mathrm{d}t}z\Big\|^2_{L^2(0,T;X)}}_{\stackrel{\text{def}}{=}C_z}\|x\|^2_X.\end{aligned}$$

For $t\in[t_k,t_{k+1}]$, we know by [**11**, Section 5.9, Theorem 2, p.286] that

$$f(t)=f(t_k)+\int_{t_k}^t \frac{\mathrm{d}}{\mathrm{d}t}f(s)ds,\quad f(t)=f(t_{k+1})+\int_{t_{k+1}}^t \frac{\mathrm{d}}{\mathrm{d}t}f(s)ds.$$

Thus, we may compute

$$\begin{aligned}\int_{t_k}^{t_{k+1}} f(t)dt &= \frac{1}{2}\int_{t_k}^{t_{k+1}} f(t)dt + \frac{1}{2}\int_{t_k}^{t_{k+1}} f(t)dt\\ &= \frac{1}{2}\int_{t_k}^{t_{k+1}}\Big(f(t_k)+\int_{t_k}^t \frac{d}{dt}f(s)ds\Big)dt + \frac{1}{2}\int_{t_k}^{t_{k+1}}\Big(f(t_{k+1})+\int_{t_{k+1}}^t \frac{d}{dt}f(s)ds\Big)dt\\ &= \frac{\Delta t}{2}\big(f(t_k)+f(t_{k+1})\big)+\frac{1}{2}\int_{t_k}^{t_{k+1}}\int_{t_k}^t \frac{\mathrm{d}}{\mathrm{d}t}f(s)dsdt + \frac{1}{2}\int_{t_k}^{t_{k+1}}\int_{t_{k+1}}^t \frac{\mathrm{d}}{\mathrm{d}t}f(s)dsdt.\end{aligned}$$

Consequently,

$$\begin{aligned}\|\mathcal{S}_N x-\mathcal{S}x\|_X &= \frac{1}{2}\Big\|\sum_{k=1}^{N-1}\int_{t_k}^{t_{k+1}}\int_{t_k}^t \frac{\mathrm{d}}{\mathrm{d}t}f(s)dsdt + \int_{t_k}^{t_{k+1}}\int_{t_{k+1}}^t \frac{\mathrm{d}}{\mathrm{d}t}f(s)dsdt\Big\|_X\\ &\le \frac{1}{2}\sum_{k=1}^{N-1}\Big(\Big\|\int_{t_k}^{t_{k+1}}\int_{t_k}^t \frac{\mathrm{d}}{\mathrm{d}t}f(s)dsdt\Big\|_X + \Big\|\int_{t_k}^{t_{k+1}}\int_{t_{k+1}}^t \frac{\mathrm{d}}{\mathrm{d}t}f(s)dsdt\Big\|_X\Big)\\ &\le \frac{1}{2}\sum_{k=1}^{N-1}\Big(\int_{t_k}^{t_{k+1}}\Big\|\int_{t_k}^t \frac{\mathrm{d}}{\mathrm{d}t}f(s)ds\Big\|_X dt + \int_{t_k}^{t_{k+1}}\Big\|\int_{t_{k+1}}^t \frac{\mathrm{d}}{\mathrm{d}t}f(s)ds\Big\|_X dt\Big),\end{aligned}$$

since the functions $t\longmapsto\int_{t_k}^t \frac{\mathrm{d}}{\mathrm{d}t}f(s)ds$ and $t\longmapsto\int_{t_{k+1}}^t \frac{\mathrm{d}}{\mathrm{d}t}f(s)ds$ belong to $L^2(0,T;X)\subset L^1(0,T;X)$. Moreover, using the Cauchy-Schwarz inequality

(twice), we get

$$\begin{aligned}\int_{t_k}^{t_{k+1}} \Big\| \int_{t_k}^{t} \frac{\mathrm{d}}{\mathrm{d}t} f(s)ds \Big\|_X dt &\leq \underbrace{\Big(\int_{t_k}^{t_{k+1}} 1dt \Big)^{\frac{1}{2}}}_{=\sqrt{\Delta t}} \Big(\int_{t_k}^{t_{k+1}} \Big\| \int_{t_k}^{t} \frac{\mathrm{d}}{\mathrm{d}t} f(s)ds \Big\|_X^2 dt \Big)^{\frac{1}{2}} \\ &\leq \sqrt{\mathrm{h}_t} \Big(\int_{t_k}^{t_{k+1}} \Big(\int_{t_k}^{t} \Big\| \frac{\mathrm{d}}{\mathrm{d}t} f(s) \Big\|_X ds \Big)^2 dt \Big)^{\frac{1}{2}} \\ &\leq \sqrt{\mathrm{h}_t} \Big(\int_{t_k}^{t_{k+1}} \underbrace{\Big(\int_{t_k}^{t} 1ds \Big)}_{\leq \mathrm{h}_t} \underbrace{\Big(\int_{t_k}^{t} \Big\| \frac{\mathrm{d}}{\mathrm{d}t} f(s) \Big\|_X^2 ds \Big)}_{\leq \left\| \frac{\mathrm{d}}{\mathrm{d}t} f \right\|^2_{L^2(0,T;X)}} dt \Big)^{\frac{1}{2}} \\ &\leq \mathrm{h}_t \sqrt{\mathrm{h}_t}\ C_z\ \|x\|_X.\end{aligned}$$

The same computations yield

$$\int_{t_k}^{t_{k+1}} \Big\| \int_{t_{k+1}}^{t} \frac{\mathrm{d}}{\mathrm{d}t} f(s)ds \Big\|_X dt \leq \mathrm{h}_t \sqrt{\mathrm{h}_t}\ C_z\ \|x\|_X.$$

Thus we obtain the following estimate

$$\|\mathcal{S}_N x - \mathcal{S}x\|_X \leq \underbrace{(N-1)\mathrm{h}_t}_{=T} \sqrt{\mathrm{h}_t}\ C_z\ \|x\|_X,$$

and equivalently

$$\frac{\|\mathcal{S}_N x - \mathcal{S}x\|_X}{\|x\|_X} \leq T\sqrt{\mathrm{h}_t} C_z,$$

which implies that (21) holds since C_z is independent of x, and $N \to \infty$ if and only if $\mathrm{h}_t \to 0$. □

Theorem 4.11. — *Assume that $z \in H^1(0,T;X)$ and the weights α_j are such that*

$$\lim_{N\to\infty} \|\mathcal{S}_N - \mathcal{S}\| = 0$$

holds. Let $\{\psi_k^N\}_{k=1}^\infty$, $\{\lambda_k^N\}_{k=1}^\infty$ and $\{\psi_k^\infty\}_{k=1}^\infty$, $\{\lambda_k^\infty\}_{k=1}^\infty$ be the eigenvectors and eigenvalues respectively defined in Theorem 3.10 and Theorem 4.8. Choose ℓ so that $\lambda_\ell^\infty \neq \lambda_{\ell+1}^\infty$ Then,

$$\lim_{N\to\infty} \lambda_k^N = \lambda_k^\infty, \qquad \lim_{N\to\infty} \psi_k^N = \psi_k^\infty$$

holds for any $k \in \{1,\dots,\ell\}$. Moreover,

$$\lim_{N\to\infty} \Big(\sum_{k=\ell+1}^{d(N)} \lambda_k^N \Big) = \sum_{m=\ell+1}^{d} \lambda_m^\infty.$$

For a proof, we refer to [**13**, Theorem 2.17, p.20-21] and [**14**, Section 3.2, p.497-500].

4.2. Parameter-dependent elliptic equations. — For this section, we follow the problem setting given by [**10**]. Denote by $\mathcal{I}$ a closed bounded interval of $\mathbb{R}$. For a parameter $\theta \in \mathcal{I}$ and $f \in X'$, consider the variational problem

$$\begin{cases} \text{Find } y \in X \text{ such that} \\ \mathfrak{a}(y, \phi\,; \theta) = \big\langle f\,, \phi \big\rangle_{X',X} \quad \forall \phi \in X \end{cases}$$

where $\mathfrak{a}(\cdot, \cdot\ ; \theta)$ is a continuous, X-elliptic, sesquilinear form. From Lax-Milgram theorem, we can deduce that there exists a unique solution $y(\cdot\,; \theta)$ for this problem. We define the function $z(\theta) \overset{\text{def}}{=} y(\cdot\,; \theta)$.
Moreover, we assume that there exists $\gamma > 0$, independent of θ, such that

$$\mathfrak{a}(\phi, \phi\,; \theta) \geq \gamma \|\phi\|_X^2,$$

and that the form a might be rewritten as

$$\mathfrak{a}(\varphi, \phi; \theta) = \sum_{q=1}^{Q} \mathfrak{a}^q(\varphi, \phi) g^q(\theta),$$

where $Q \in \mathbb{N}$, and for $q \in \{1, \ldots, Q\}$, the form $\mathfrak{a}^q$ is continuous sesquilinear and the function g^q is Lipschitz-continuous. In this context, we have the following Lemma.

Lemma 4.12. — *The function $z\colon \theta \longmapsto z(\theta)$ is Lipschitz-continuous from $\mathcal{I}$ to X.*

Proof. — We follow the arguments of [**10**, Lemma 2.1] in this simplified context. Let $\theta_1, \theta_2 \in \mathcal{I}$, then $z(\theta_1)$ and $z(\theta_2)$ satisfy

$$\begin{cases} \mathfrak{a}(z(\theta_1), \phi\,; \theta_1) = \big\langle f\,, \phi \big\rangle_{X',X} \\ \mathfrak{a}(z(\theta_2), \phi\,; \theta_2) = \big\langle f\,, \phi \big\rangle_{X',X} \end{cases}$$

for any $\phi \in X$, and consequently,

$$\begin{aligned} \mathfrak{a}\big(z(\theta_1) - z(\theta_2), \phi\,; \theta_1\big) &= \big\langle f\,, \phi \big\rangle_{X',X} - \mathfrak{a}(z(\theta_2), \phi\,; \theta_1) \\ &= \mathfrak{a}(z(\theta_2), \phi\,; \theta_2) - \mathfrak{a}(z(\theta_2), \phi\,; \theta_1) \\ &= \sum_{q=1}^{Q} \mathfrak{a}^q(z(\theta_2), \phi)\big(g^q(\theta_2) - g^q(\theta_1)\big). \end{aligned}$$

Choose $\phi = z(\theta_1) - z(\theta_2)$. On the one hand ellipticity, the hypothesis implies that

$$\|z(\theta_1) - z(\theta_2)\|_X^2 \leq \frac{1}{\gamma}\mathfrak{a}\big(z(\theta_1) - z(\theta_2), z(\theta_1) - z(\theta_2)\,;\theta_1\big).$$

On the other , we can use the hypothesis made on $\mathfrak{a}^q$ and g^q, to estimate

$$\begin{aligned}\mathfrak{a}\big(z(\theta_1) - z(\theta_2), z(\theta_1) - z(\theta_2)\,;\theta_1\big)&\\ &\leq \sum_{q=1}^{Q} \big|\mathfrak{a}^q(z(\theta_2), z(\theta_1) - z(\theta_2))\big|\big|g^q(\theta_2) - g^q(\theta_1)\big|\\ &\lesssim \|z(\theta_2)\|_X\|z(\theta_1) - z(\theta_2)\|_X|\theta_1 - \theta_2|.\end{aligned}$$

Thus we obtain $\|z(\theta_1) - z(\theta_2)\|_X \lesssim \|z(\theta_2)\|_X|\theta_1 - \theta_2|$. Using again the ellipticity hypothesis and the equation satisfied by $z(\theta)$, we can estimate its norm independently of the parameter θ, since

$$\|z(\theta)\|_X^2 \leq \frac{1}{\gamma}\mathfrak{a}(z(\theta), z(\theta)\,;\theta) = \frac{1}{\gamma}\big\langle z(\theta)\,, f\big\rangle_{X',X} \leq \frac{\|f\|_{X'}}{\gamma}\|z(\theta)\|_X$$

implies that $\|z(\theta)\|_X \leq \frac{\|f\|_{X'}}{\gamma}$. Consequently, z is Lipschitz continuous

$$\|z(\theta_1) - z(\theta_2)\|_X \lesssim |\theta_1 - \theta_2|. \qquad \square$$

Remark 4.13. — Since z is Lipschitz continuous, it belongs to $H^1(\mathcal{I};X)$. To show this, we use the same reasoning as [**11**, Section 5.8, Theorem 4, p.279]. Being continuous, $z \in L^2(\mathcal{I};X)$ with

$$\|z\|_{L^2(\mathcal{I};X)}^2 = \int_{\mathcal{I}} \|z(\theta)\|_X^2 d\theta \leq \mathrm{meas}(\mathcal{I})\Big(\max_{\theta\in\mathcal{I}}\|z(\theta)\|_X^2\Big) < \infty.$$

Moreover, for $h > 0$, using the change of variable $\theta_1 = \theta_2 - h$, we observe that

$$\sup_{\theta_1\neq\theta_2\in\mathcal{I}} \Big\|\frac{z(\theta_1) - z(\theta_2)}{\theta_1 - \theta_2}\Big\|_X < \infty,$$

is equivalent to

$$\sup_{\theta_2\in\mathcal{I},\ h>0} \Big\|\frac{z(\theta_2 - h) - z(\theta_2)}{-h}\Big\|_X < \infty.$$

Then, for any $h > 0$ define the maps $g_h\colon \mathcal{I} \to X$ by

$$g_h(\theta) \stackrel{\text{def}}{=} \frac{1}{-h}\big(z(\theta - h) - z(\theta)\big),$$

These maps satisfy

$$\sup_{h>0,\ \theta\in\mathcal{I}} \|g_h(\theta)\|_X < \infty,$$

which implies that $\{g_h\colon h > 0\}$ is bounded in $L^2(\mathcal{I};X)$. Since $L^2(\mathcal{I};X)$ is a Hilbert space, we deduce from the Banach-Alaoglu theorem that there exists

$\widetilde{g} \in L^2(\mathcal{I};X)$ and a sequence $(h_k)_{k\geq 1} \subset (0,+\infty[$ such that $g_{h_k} \rightharpoonup \widetilde{g}$ weakly in $L^2(\mathcal{I};X)$. Using this information, we can show that $\widetilde{g}$ is the weak derivative of z. Indeed, for any $\phi \in C_c^\infty(\mathcal{I},\mathbb{R})$ we have

$$\begin{aligned}\int_{\mathcal{I}} \Big\langle z(\theta)\,, \frac{\mathrm{d}}{\mathrm{d}\theta}\phi(\theta)\Big\rangle_X d\theta &= \int_{\mathcal{I}} \Big\langle z(\theta)\,, \lim_{k\to\infty} \frac{1}{h_k}\big(\phi(\theta+h_k)-\phi(\theta)\big)\Big\rangle_X d\theta \\ &= \lim_{k\to\infty} \int_{\mathcal{I}} \Big\langle z(\theta)\,, \frac{1}{h_k}\big(\phi(\theta+h_k)-\phi(\theta)\big)\Big\rangle_X d\theta \\ &= \lim_{k\to\infty} \Big[\int_{\mathcal{I}} \Big\langle z(\theta)\,, \frac{\phi(\theta+h_k)}{h_k}\Big\rangle_X d\theta - \int_{\mathcal{I}} \Big\langle z(\theta)\,, \frac{\phi(\theta)}{h_k}\Big\rangle_X d\theta\Big].\end{aligned}$$

Changing $\theta + h_k$ by s in the first integral yields

$$\begin{aligned}\int_{\mathcal{I}} \Big\langle z(\theta)\,, \frac{\mathrm{d}}{\mathrm{d}\theta}\phi(\theta)\Big\rangle_X d\theta &= \lim_{k\to\infty} \Big[\int_{\mathcal{I}} \Big\langle z(s-h_k)\,, \frac{\phi(s)}{h_k}\Big\rangle_X ds - \int_{\mathcal{I}} \Big\langle z(\theta)\,, \frac{\phi(\theta)}{h_k}\Big\rangle_X d\theta\Big] \\ &= -\lim_{k\to\infty} \int_{\mathcal{I}} \big\langle g_{h_k}(\theta)\,, \phi(\theta)\big\rangle_X d\theta \\ &= -\int_{\mathcal{I}} \big\langle \widetilde{g}(\theta)\,, \phi(\theta)\big\rangle_X d\theta,\end{aligned}$$

since g_{h_k} converges weakly to $\widetilde{g}$. We can conclude that $u \in H^1(\mathcal{I};X)$, since $\frac{\mathrm{d}}{\mathrm{d}\theta} z = \widetilde{g} \in L^2(\mathcal{I};X)$.

In this context, we could define the discrete POD problem, and link it to the continuous POD problem, using the same reasoning as for evolution equations. We may divide the interval $\mathcal{I}$ in $N-1$ subintervals by defining N parameters $\{\theta_j\}_{j=1}^N \subset \mathcal{I}$. For equidistant parameters, $h_t \overset{\text{def}}{=} \frac{T}{N-1}$ and we obtain:

Define the vectors $\{z_j\}_{j=1}^N$ by $z_j \overset{\text{def}}{=} z(\theta_j)$. The discrete POD problem may be posed, for the set $\mathcal{V}_N \overset{\text{def}}{=} \operatorname{span}\{z_k\}_{k=1}^N$, as

$$(\mathrm{P}_N) \qquad \begin{cases} \arg\min \Big\{ \displaystyle\sum_{j=1}^N \alpha_j \|z_j - P^{\mathcal{B}} z_j\|_X^2 : \mathcal{B} \overset{\text{def}}{=} \{\phi_k\}_{k=1}^\ell \subset \mathcal{V}_N \Big\} \\ \text{subject to } \big\langle \phi_i\,, \phi_j\big\rangle_X = \delta_{i,j}, \quad \text{for } 1 \leq i,j \leq \ell, \end{cases}$$

and the continuous POD problem, for $\mathcal{V} \stackrel{\text{def}}{=} \text{span}\{z(\theta) \colon \theta \in \mathcal{I}\}$, is given by

$$\text{(P)} \qquad \begin{cases} \arg\min \left\{ \displaystyle\int_{\mathcal{I}} \left\| z(\theta) - P^{\mathcal{B}} z(\theta) \right\|_X^2 d\theta \, : \, \mathcal{B} \stackrel{\text{def}}{=} \{\phi_k\}_{k=1}^{\ell} \subset \mathcal{V}_N \right\} \\ \text{subject to } \left\langle \phi_i \, , \phi_j \right\rangle_X = \delta_{i,j}, \quad \text{for } 1 \leq i, j \leq \ell. \end{cases}$$

Since $z \in H^1(\mathcal{I}; X)$, a POD basis exists both for the continuous POD problem and for the discrete POD problem, and Theorem 4.11 holds.

PART III. BALANCED TRUNCATION AND BALANCED POD

In this part, we mainly follow [**8**], [**12**] and [**20**] to introduce the balanced truncation method, as well as [**18**] for a modification of this method called *balanced proper orthogonal decomposition* (BPOD for short).

5. The problem and the governing operators

We consider the following linear finite dimensional system

$$(\mathbf{S}):\ \begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}z(t) = Az(t) + Bu(t), \quad z(0) = z_0 \\ y(t) = Cz(t) + Du(t) \end{cases}$$

where $A \in \mathbb{C}^{n\times n}$ is a Hurwitz matrix (meaning that all eigenvalues of A have negative real parts), $B \in \mathbb{C}^{n\times m}$, $C \in \mathbb{C}^{p\times n}$ and $D \in \mathbb{C}^{p\times m}$. For any $t \geq 0$, $z(t)$ lies in the *state space* $\mathbb{C}^n$, $u(t)$ lies in the *input space* $\mathbb{C}^m$ and $y(t)$ lies in the *output space* $\mathbb{C}^p$.

Remark 5.1. — Since A is Hurwitz, there exist constants $M > 1$ and $q < 0$ such that

$$\left\|e^{At}\right\| \leq Me^{qt}.$$

For a proof, see [**21**, Proposition 1.3.3, p.9]. In what follows, we will use this property to justify the convergence of some integrals and the boundedness of some operators.

Definition 5.2 **(McMillan degree or order).** — The *McMillan degree* or *order* of a system is the dimension of its state space.

Here, $(\mathbf{S})$ has order n. Our aim is to find a system $(\mathbf{S}_\ell)$ of order $\ell < n$, which approximates $(\mathbf{S})$. By *approximate*, we mean that the connection between inputs and outputs in $(\mathbf{S}_\ell)$ is close to the connection between inputs and outputs in $(\mathbf{S})$. We will latter define how these systems are compared. The approximate system studied here will be obtained by the *balanced truncation method*, and will have the form

$$(\mathbf{S}_\ell):\ \begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}z_\ell(t) = A_\ell z_\ell(t) + B_\ell u(t), \quad z(0) = z_0 \\ y(t) = C_\ell z_\ell(t) + Du(t). \end{cases}$$

As indicated by the notation, the matrix D will remain unchanged by the method, while the matrices A, B and C will be replaced by $A_\ell \in \mathbb{C}^{\ell\times\ell}$, $B_\ell \in \mathbb{C}^{\ell\times m}$ and $C_\ell \in \mathbb{C}^{p\times\ell}$.

5.1. Relation between inputs and outputs. — Here we study two ways of observing the relation between inputs and outputs in (**S**).

5.1.1. *In the time domain.* — For a given function $u \in L^2(\mathbb{R}_+;\mathbb{C}^m)$, the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}z(t) = Az(t) + Bu(t), \quad z(0) = z_0$$

has a unique solution $z \in C(0,+\infty;\mathbb{C}^n)$ given by

$$z(t) = e^{tA}z_0 + \int_0^t e^{(t-\tau)A}Bu(\tau)d\tau, \quad t \in \mathbb{R}_+.$$

The output is consequently given by

$$y(t) = Ce^{tA}z_0 + \int_0^t Ce^{(t-\tau)A}Bu(\tau)d\tau + Du(t), \quad t \in \mathbb{R}_+.$$

When the initial condition is $z_0 = 0$, this relation provides a link between the input functions u and the output functions y, for any $t \in \mathbb{R}_+$

$$\begin{aligned} y(t) &= \int_0^t Ce^{(t-\tau)A}Bu(\tau)d\tau + Du(t) \\ &= \int_0^{+\infty} Ce^{\tau A}Bu(t-\tau)d\tau + Du(t). \end{aligned} \tag{22}$$

In order to deduce the last equality, one has to use that the support of u is included in $\mathbb{R}_+$.

5.1.2. *In the frequency domain.* — Let $s \in \mathbb{C}$ be such that $(sI - A)^{-1}$ is well defined. In particular we can choose any s with a nonnegative real part, since A is Hurwitz.

***Definition 5.3* (Transfer function).** — The linear map $G(s)\colon \mathbb{C}^m \longmapsto \mathbb{C}^p$ defined by

$$G(s) \overset{\text{def}}{=} C(sI_n - A)^{-1}B + D \overset{\text{def}}{=} \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]$$

is called the *transfer function*.

The transfer function maps the Laplace transform of input functions to the Laplace transform of output functions. This relation appears when we apply the Laplace transform to the system (S) with initial condition $z(0) = 0$. The system is transformed into

$$\begin{cases} s(\mathcal{L}z)(s) - z(0) = A(\mathcal{L}z)(s) + B(\mathcal{L}u)(s), \quad z(0) = 0 \\ (\mathcal{L}y)(s) = C(\mathcal{L}z)(s) + D(\mathcal{L}u)(s). \end{cases}$$

This yields the identity $(\mathcal{L}z)(s) = (sI_n - A)^{-1}B(\mathcal{L}u)(s)$, consequently

$$\begin{aligned} (\mathcal{L}y)(s) &= \big(C(sI_n - A)^{-1}B\big)(\mathcal{L}u)(s) + D(\mathcal{L}u)(s) \\ &= G(s)(\mathcal{L}u)(s) \\ &= \big[\mathcal{M}_G(\mathcal{L}u)\big](s), \end{aligned}$$

where we denote by $\mathcal{M}_G$ the multiplication operator by G.

Another way to derive the same relation is to compute the Laplace transform of the output function y given by (22) for some input function $u \in L^2(\mathbb{R}_+; \mathbb{C}^m)$. Let $s \in \mathbb{C}_+$ be such that $\mathcal{L}u$ is well defined,

$$\begin{aligned} \mathcal{L}(y)(s) &= \lim_{t\to\infty} \int_0^t e^{-s\tau} y(\tau) d\tau + D(\mathcal{L}u)(s) \\ &= \lim_{t\to\infty} \int_0^t e^{-s\tau} \Big(\int_0^{+\infty} Ce^{\theta A} B u(\tau - \theta) d\theta \Big) d\tau + D(\mathcal{L}u)(s) \end{aligned}$$

We may apply the Fubini theorem since the function f defined by

$$f(\tau, \theta) \stackrel{\text{def}}{=} e^{-s\tau} Ce^{\theta A} B u(\tau - \theta)$$

belongs to $L^1(\mathbb{R}_+ \times \mathbb{R}_+; \mathbb{C}^m)$ with

$$\|f\|_{L^1(\mathbb{R}_+\times\mathbb{R}_+;\mathbb{C}^m)} \le \|C\|\|B\| \Big(\int_0^{+\infty} e^{-\Re(s)\tau} d\tau \Big) \|u\|_{L^2(\mathbb{R}_+;\mathbb{C}^m)} \Big(\int_0^{+\infty} e^{tA} dt \Big)^{\frac{1}{2}}.$$

Hence,

$$\begin{aligned} \mathcal{L}(y)(s) &= \lim_{t\to\infty} \int_0^{+\infty} Ce^{\theta A} B \Big(\int_0^t e^{-s\tau} u(\tau - \theta) d\tau \Big) d\theta + D(\mathcal{L}u)(s) \\ &= \lim_{t\to\infty} \int_0^{+\infty} e^{-s\theta} Ce^{\theta A} B \Big(\int_0^{t-\theta} e^{-s\tau} u(\tau) d\tau \Big) d\theta + D(\mathcal{L}u)(s) \end{aligned}$$

We define

$$w_t(\theta) \stackrel{\text{def}}{=} e^{-s\theta} Ce^{\theta A} B \int_0^{t-\theta} e^{-s\tau} u(\tau) d\tau.$$

For almost every $\theta \in \mathbb{R}_+$, $w_t(\theta)$ tends to $e^{-s\theta}Ce^{\theta A}B(\mathcal{L}u)(s)$ as t goes to $+\infty$. Moreover, for almost every $\theta, t \in \mathbb{R}_+$,

$$\begin{aligned}|w_t(\theta)| &\leq e^{-\Re(s)\theta}\|Ce^{\theta A}B\| \int_0^{t-\theta} e^{-\Re(s)\tau}|u(\tau)|d\tau \\ &\leq e^{-\Re(s)\theta}\|Ce^{\theta A}B\| \left\|e^{\Re(s)\cdot}\right\|_{L^2(\mathbb{R}_+;\mathbb{R})} \|u\|_{L^2(\mathbb{R}_+;\mathbb{C}^m)},\end{aligned}$$

where the right-hand side is an integrable function of r. Thus, applying Lebesgue's dominated convergence theorem yields

$$\begin{aligned}\mathcal{L}(y)(s) &= \int_0^{\infty} e^{-s\theta}Ce^{\theta A}B(\mathcal{L}u)(s)d\theta + D(\mathcal{L}u)(s) \\ &= \big[\mathcal{L}(Ce^{\cdot A}B) + D\big](s)(\mathcal{L}u)(s).\end{aligned}$$

Finally, we may show that $\big[\mathcal{L}(Ce^{\cdot A}B) + D\big](s) = G(s)$ by computing, for $s \in \mathbb{C}_+$,

$$\begin{aligned}\big[\mathcal{L}(Ce^{\cdot A}B)\big](s) &= \lim_{t\to\infty} \int_0^t e^{-s\tau}Ce^{\tau A}Bd\tau \\ &= C \lim_{t\to\infty} \int_0^t e^{-s\tau I_n}e^{\tau A}d\tau B \\ &= C(A - sI_n)^{-1} \lim_{t\to\infty} \int_0^t (A - sI_n)e^{\tau(A-sI_n)}d\tau B \\ &= C(A - sI_n)^{-1}\Big(\lim_{t\to\infty} e^{t(A-sI_n)} - e^{0(A-sI_n)}\Big)B\end{aligned}$$

Since A is Hurwitz, $\lim_{t\to\infty} e^{t(A-sI_n)} = \mathbb{O}_n$ in $\mathscr{L}(\mathbb{C}^n)$, which yields

$$\big[\mathcal{L}(Ce^{\cdot A}B)\big](s) = C(A - sI_n)^{-1}(\mathbb{O}_n - I_n)B = C(sI_n - A)^{-1}B.$$

Remark 5.4. — The transfer function G is analytic on $\mathbb{C}_+$, and also continuous and bounded on the closure of $\mathbb{C}_+$. Indeed, since the eigenvalues of A belong to the left half-plane, the closure of the right half-plane is included in the resolvent set of A. This implies that the resolvent $s \mapsto (sI_n - A)^{-1}$ is analytic on $\mathbb{C}_+$ and continuous on closure($\mathbb{C}_+$), hence so is $s \mapsto G(s)$. Moreover, the transfer function is *rational* in the sense that each of its components are rational functions of s, that is, functions f of the form

$$f(s) = \frac{P(s)}{Q(s)},$$

where P and Q are polynomials. Indeed, let s be such that $(sI_n - A)^{-1}$ is well defined, then

$$(sI_n - A)^{-1} = \frac{\mathrm{adj}(sI_n - A)}{\det(sI_n - A)}.$$

Here, adj denotes the *adjugate matrix*[3]. The components of $adj(sI_n - A)$ are $(\pm)$ determinants of submatrices of $(sI_n - A)$ formed by deleting a row and a column. Thus these components are polynomials of degree less than n. The determinant of $(sI_n - A)$ is a polynomial of degree n. Consequently, every component $f_{ij}(s)$ of $G(s)$ is a rational function whose denominator polynomial has a larger degree than the numerator polynomial, hence

$$\lim_{|s|\to\infty} f_{ij}(s) = 0.$$

Furthermore, we may deduce from theses remarks that G is also bounded on $\mathrm{clos}(\mathbb{C}_+)$. Indeed, fix s and consider the largest singular value $\overline{\sigma}(G(s))$ of $G(s)$ which satisfies, for some normalized vectors $u(s) \stackrel{\text{def}}{=} (u_1(s), \ldots, u_p(s))^T \in \mathbb{C}^p$ and $v(s) \stackrel{\text{def}}{=} (v_1(s), \ldots, v_m(s))^T \in \mathbb{C}^m$,

$$G(s)u(s) = \overline{\sigma}\big(G(s)\big)v(s).$$

For $j \in \{1, \ldots, m\}$, denote by $G_j(s)$ the columns of $G(s)$. We can see that $\overline{\sigma}(G(s))$ tends to zero as $|s|$ goes to $+\infty$ since, on the one hand,

$$\overline{\sigma}(G(s)) = \overline{\sigma}(G(s))|u(s)| = |G(s)v(s)| = \Big|\sum_{j=1}^{m} G_j(s)v_j(s)\Big|$$

and on the other hand, the triangular inequality and the fact that $v(s)$ is of unit norm yield

$$\overline{\sigma}(G(s)) \leq \sum_{j=1}^{m} |G_j(s)| = \sum_{j=1}^{m} \Big(\sum_{i=1}^{p} |f_{ij}(s)|^2\Big)^{\frac{1}{2}} \xrightarrow[|s|\to\infty]{} 0.$$

Knowing that the transfer function is also continuous on $\mathrm{clos}(\mathbb{C}_+)$, we deduce that it bounded on this set.

Following [**15**, Section 1.4, p.13-14] we define the Hardy space of bounded analytic matrix-valued functions. The norm of this space will be used to quantify the error made when approximating the system $(\mathbf{S})$ by $(\mathbf{S}_\ell)$.

3. For a matrix M, the adjugate matrix of M is the transpose of its cofactor matrix

Definition 5.5. — The Hardy space $H^\infty(\mathbb{C}_+;\mathbb{C}^{p\times m})$ is the following set of matrix-valued functions

$$\big\{F\colon \mathbb{C}_+ \to \mathbb{C}^{p\times m} \text{ analytic, s.t. } \|F\|_{H^\infty_{p\times m}} < \infty\big\}$$

where the $H^\infty_{p\times m}$ norm is

$$\|F\|_{H^\infty_{p\times m}} \stackrel{\text{def}}{=} \sup_{s\in\mathbb{C}_+} \|F(s)\|.$$

Proposition 5.6. — *If $F\colon \mathbb{C}_+ \to \mathbb{C}^{p\times m}$ is analytic on $\mathbb{C}_+$, and is also continuous and bounded on the closure of $\mathbb{C}_+$, then*

$$\|F\|_{H^\infty_{p\times m}} = \sup_{w\in\mathbb{R}} \overline{\sigma}\big(M(iw)\big).$$

A sketch of a proof is given in the Appendix.

***Definition 5.7* (H^∞-error).** — Let G and G_ℓ be the transfer functions of $(\mathbf{S})$ and $(\mathbf{S}_\ell)$ respectively. The *H^∞-error of approximation* is defined as

$$\mathrm{err}(G, G_\ell) \stackrel{\text{def}}{=} \|G - G_\ell\|_{H^\infty_{p\times m}}.$$

We will see below, in Proposition 5.16, that this definition implies that we are asking for the size of the worst output error $y - y_\ell$ to be kept small over all normalized inputs u.

We now introduce operators that are linked to one another and will be used in what follows.

5.2. Input map, output map and Gramians. — We begin with the definition the *input map* and the *output map*, which will be used to define or rewrite several quantities and operators.

***Definition 5.8* (Output map).** — The linear map $\Psi\colon \mathbb{C}^n \to L^2(\mathbb{R}_+;\mathbb{C}^p)$ defined by

$$(\Psi x)(t) = \begin{cases} Ce^{tA}x, & t \geq 0 \\ 0, & t < 0 \end{cases}$$

is called the *output map* of $(\mathbf{S})$.

Since A is Hurwitz, Ψ is well defined and bounded. Indeed, for any $x \in \mathbb{C}^n$,

$$\|\Psi x\|^2_{L^2(0,+\infty,\mathbb{C})} = \int_0^{+\infty} |Ce^{tA}x|^2 dt \leq \|C\|^2 \int_0^{+\infty} \|e^{tA}\|^2_{\mathscr{L}(\mathbb{C}^n)} |x|^2 \ < \infty.$$

For nonnegative t, we can interpret Ψx as the output function of $(\mathbf{S})$, with a zero input function, and the initial condition x, meaning the output of the following system:

$$(\mathbf{S}_{\mathrm{o}})\colon \begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}z(t) = Az(t), \quad z(0) = x \\ y(t) = Cz(t). \end{cases}$$

***Definition 5.9* (Input map).** — The linear map $\Phi\colon L^2(\mathbb{R}_+;\mathbb{C}^m) \to \mathbb{C}^n$ defined by

$$\Phi u = \int_0^{+\infty} e^{\tau A} B u(\tau) d\tau.$$

is called the *input map* of $(\mathbf{S})$.

Since A is Hurwitz, Φ is well defined and bounded. Indeed, for any $u \in L^2(\mathbb{R}_+;\mathbb{C}^m)$,

$$|\Phi u| \leq \|B\| \int_0^{+\infty} \|e^{\tau A}\| |u(\tau)| d\tau \leq \|B\| \Big(\int_0^{+\infty} \|e^{\tau A}\|^2 d\tau \Big)^{\frac{1}{2}} \|u\|_{L^2(\mathbb{R}_+;\mathbb{C}^m)}.$$

In order to interpret Φu, we need to introduce the following operator which maps an element of $L^2(\mathbb{R}_+;\mathbb{C}^m)$ to an element of $L^2(\mathbb{R}_-;\mathbb{C}^m)$.

***Definition 5.10* ($\boldsymbol{\mathfrak{R}}$).** — Let $\boldsymbol{\mathfrak{R}}\colon L^2(\mathbb{R};\mathbb{C}^m) \to L^2(\mathbb{R};\mathbb{C}^m)$ be the linear operator defined by $(\boldsymbol{\mathfrak{R}}u)(t) \stackrel{\text{def}}{=} u(-t)$.

Let $u \in L^2(\mathbb{R}_+;\mathbb{C}^m)$, then the function Φu is also the solution to the following differential equation, evaluated at time $t = 0$:

$$(\mathbf{S}_{\mathrm{c}}) \ : \ \begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}z^\infty(t) = Az^\infty(t) + B(\boldsymbol{\mathfrak{R}}u)(t), \quad t \in \mathbb{R}_- \\ \lim\limits_{\tau \to -\infty} z^\infty(\tau) = 0. \end{cases}$$

Let us give the meaning to $(\mathbf{S}_{\mathrm{c}})$. The solution z^∞ to $(\mathbf{S}_{\mathrm{c}})$ is seen as the limit, as τ tends to $-\infty$, of the family of solutions $(z^\tau)_{\tau \leq 0}$ to the systems (defined for any $\tau \leq 0$)

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}z^\tau(t) = Az^\tau(t) + B(\boldsymbol{\mathfrak{R}}u)(t), \quad t \in [\tau, 0] \\ z^\tau(\tau) = \eta_\tau, \end{cases}$$

where $\lim_{\tau \to -\infty} \eta_\tau = 0$. The solution of the above system is given by

$$z^\tau(t) = e^{(t-\tau)A}\eta_\tau + \int_\tau^t e^{(t-\theta)A} B(\boldsymbol{\mathfrak{R}}u)(\theta) d\theta, \quad \text{for } t \in [\tau, 0].$$

For any $t \in [\tau, 0]$, $z^\tau(t)$ tends to

$$z^\infty(t) \stackrel{\text{def}}{=} \int_{-\infty}^{t} e^{(t-\theta)A} B(\mathfrak{R}u)(\theta)d\theta$$

as τ goes to $-\infty$. Evaluating at $t = 0$ yields

$$\begin{aligned} z^\infty(0) &= \int_{-\infty}^{0} e^{-\theta A} B(\mathfrak{R}\, u)(\theta)d\theta \\ &= \int_{0}^{+\infty} e^{\theta A} (\mathfrak{R}\, u)(-\theta)d\theta \\ &= \int_{0}^{+\infty} e^{\theta A} u(\theta)d\theta \;=\; \Phi u. \end{aligned}$$

***Definition 5.11* (Observability Gramian).** — The linear map $\mathcal{Q}\colon \mathbb{C}^n \to \mathbb{C}^n$ defined by

$$\mathcal{Q} \stackrel{\text{def}}{=} \Psi^*\Psi,$$

is called the *observability Gramian* of (**S**).

This operator is bounded and self-adjoint, and may also be rewritten as

$$\mathcal{Q}\colon x \longmapsto \int_0^\infty e^{tA^*} C^* C e^{tA} x dt.$$

Indeed, the adjoint of Ψ may be found by computing, for $x \in \mathbb{C}^n$ and $y \in L^2(0, +\infty; \mathbb{C}^p)$,

$$\begin{aligned} \big\langle \Psi x\,, y \big\rangle_{L^2(0,+\infty;\mathbb{C}^p)} &= \int_0^\infty \langle C e^{tA} x\,, y(t)\rangle_{\mathbb{C}^p} dt \\ &= \int_0^\infty \langle x\,, e^{tA^*} C^* y(t)\rangle_{\mathbb{C}^n} dt \\ &= \Big\langle x\,, \underbrace{\int_0^\infty e^{tA^*} C^* y(t) dt}_{=\,\Psi^* y} \Big\rangle_{\mathbb{C}^n}. \end{aligned}$$

Using this formula, we observe that

$$\Psi^*\Psi x = \int_0^\infty e^{tA^*} C^* (\Psi x)(t) dt = \int_0^\infty e^{tA^*} C^* C e^{tA} x dt$$

holds for any $x \in \mathbb{C}^n$.

***Definition 5.12* (Controllability Gramian).** — The linear map $\mathcal{R}\colon \mathbb{C}^n \to \mathbb{C}^n$ defined by

$$\mathcal{R} \stackrel{\text{def}}{=} \Phi\Phi^*,$$

is called the *controllability operator* of (**S**).

This operator is bounded and self-adjoint, and may also be rewritten as

$$\mathcal{R}\colon x \longmapsto \int_0^\infty e^{tA}BB^*e^{tA^*}x dt.$$

Indeed, the adjoint of Φ may be found by computing, for $u \in L^2(\mathbb{R}_+;\mathbb{C}^m)$ and $x \in \mathbb{C}^n$

$$\begin{aligned}\langle \Phi u\,, x\rangle_{\mathbb{C}^n} &= \Big\langle \int_0^{+\infty} e^{\tau A}Bu(\tau)d\tau\,, x\Big\rangle_{\mathbb{C}^n}\\ &= \int_0^{+\infty} \langle e^{\tau A}Bu(\tau)\,, x\rangle_{\mathbb{C}^n} d\tau\\ &= \int_0^{+\infty} \langle u(\tau)\,, B^*e^{\tau A^*}x\rangle_{\mathbb{C}^m} d\tau\\ &= \big\langle u\,, B^*e^{A^*\cdot}x\big\rangle_{L^2(\mathbb{R}_+;\mathbb{C}^m)}.\end{aligned}$$

Then, using the formula $\Phi^*x = B^*e^{A^*\cdot}x$ for the adjoint, we observe that

$$\Phi\Phi^*x = \int_0^{+\infty} e^{\tau A}B(\Phi^*x)(\tau)d\tau = \int_0^{+\infty} e^{\tau A}BB^*e^{\tau A^*}x d\tau$$

holds for any $x \in \mathbb{C}^n$.

We state two results involving the Gramians that will be useful for what follows. Proofs may be found in [**21**, Proposition 1.5.13 and Proposition 1.5.14, p.18].

Proposition 5.13. — *If A is Hurwitz, then the observability and controllability Gramians $\mathcal{Q}$ and $\mathcal{R}$ are the unique solutions in $\mathscr{L}(\mathbb{C}^n)$ to the Lyapunov equations*

$$A\mathcal{R} + \mathcal{R}A^* + BB^* = 0 \qquad \textit{and} \qquad \mathcal{Q}A + A^*\mathcal{Q} + C^*C = 0.$$

Proposition 5.14. — *If A is Hurwitz, then (A, B) is controllable if and only if $\mathcal{R}$ is strictly positive, (A, C) is observable if and only if $\mathcal{Q}$ is strictly positive.*

5.3. Hankel operator and relations between norms. — Before introducing more operators, we make the following assumption valid for the remaining of part III.

Assumption 5.1. — *We assume without loss of generality that the matrix D is equal to zero. This will neither affect the balanced truncation method since we will see that this method does not affect or depend on D, nor will it affect the*

H^∞-error since the transfer function of $(\mathbf{S}_\ell)$ is $G_\ell(s) = C_\ell(\cdot I_n - A_\ell)^{-1}B_\ell + D$ and we have

$$\begin{aligned}\big\|G(s) - G_\ell(s)\big\|_{H^\infty_{p\times m}} &= \big\|C(\cdot I_n - A)^{-1}B + D - \big(C_\ell(\cdot I_n - A_\ell)^{-1}B_\ell + D\big)\big\|_{H^\infty_{p\times m}} \\ &= \big\|C(\cdot I_n - A)^{-1}B - C_\ell(\cdot I_n - A_\ell)^{-1}B_\ell\big\|_{H^\infty_{p\times m}}.\end{aligned}$$

Hence, from now onwards, $(\mathbf{S})$ and $(\mathbf{S}_\ell)$ denote the systems

$$(\mathbf{S})\colon \begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}z(t) = Az(t) + Bu(t), \quad z(0) = z_0 \\ y(t) = Cz(t), \end{cases}$$

of transfer function $G(s) = C(sI_n - A)^{-1}B$, and

$$(\mathbf{S}_\ell)\colon \begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}z_\ell(t) = A_\ell z_\ell(t) + B_\ell u(t), \quad z(0) = z_0 \\ y(t) = C_\ell z_\ell(t), \end{cases}$$

of transfer function $G_\ell(s) \overset{def}{=} C_\ell(sI_n - A_\ell)^{-1}B_\ell$.

Notation 5.1. — *To lighten the notation, we will denote by g the following matrix valued function*

$$g(t) \overset{def}{=} Ce^{tA}B.$$

This function is also known as the *impulse response* of the system $(\mathbf{S})$ defined above. Notice that G is also given by $G(s) = (\mathcal{L}g)(s)$ (see section 5.1.2).

We may now introduce the operator γ_g which will help us to compare the H^∞-norm of G and the norm of another operator, studied bellow, called the *Hankel operator*.

Definition 5.15 (γ_g). — For g the impulse response of $(\mathbf{S})$, let $\gamma_g\colon L^2(\mathbb{R};\mathbb{C}^m) \to L^2(\mathbb{R};\mathbb{C}^p)$ be the linear operator defined by

$$(\gamma_g v)(t) \overset{\text{def}}{=} \int_{-\infty}^{t} g(t-\tau)v(\tau)d\tau.$$

Proposition 5.16. — *The transfer function G and the operator γ_g satisfy*

$$\|G\|_{H^\infty_{p\times m}} = \|\mathcal{M}_G\| = \|\gamma_g\|.$$

where $\mathcal{M}_G\colon L^2(i\mathbb{R},\mathbb{C}^m) \to L^2(i\mathbb{R},\mathbb{C}^p)$ is the multiplication operator by G, defined by $(\mathcal{M}_G u)(iw) \overset{def}{=} G(iw)u(iw)$.

Proof. — Take some $u \in L^2(i\mathbb{R}, \mathbb{C}^m)$, we observe that

$$\begin{aligned}
\|\mathcal{M}_G u\|^2_{L^2(i\mathbb{R},\mathbb{C}^p)} &= \int_{-\infty}^{+\infty} |\mathcal{M}_G u(iw)|^2 dw \\
&= \int_{-\infty}^{+\infty} |G(iw)u(iw)|^2 dw \\
&\leq \int_{-\infty}^{+\infty} \|G(iw)\|^2_{\mathscr{L}(\mathbb{C}^m,\mathbb{C}^p)} |u(iw)|^2 dw \\
&\leq \Big(\sup_{\xi\in\mathbb{R}} \|G(i\xi)\|_{\mathscr{L}(\mathbb{C}^m,\mathbb{C}^p)} \Big)^2 \int_{-\infty}^{+\infty} |u(iw)|^2 dw \\
&= \|G\|^2_{H^\infty_{p\times m}} \|u\|^2_{L^2(i\mathbb{R},\mathbb{C}^m)},
\end{aligned}$$

Consequently, $\mathcal{M}_G$ is bounded from $L^2(i\mathbb{R}, \mathbb{C}^m)$ into $L^2(i\mathbb{R}, \mathbb{C}^p)$ with operator norm less than $\|G\|_{H^\infty_{p\times m}}$. In order to conclude that the operator norm is in fact equal to $\|G\|_{H^\infty_{p\times m}}$, we will show that for any $\varepsilon > 0$,

$$\|\mathcal{M}_G\| \geq \|G\|_{H^\infty_{p\times m}} - \varepsilon.$$

Let $\varepsilon > 0$ be fixed. We are going find $u_0 \in L^2(i\mathbb{R}; \mathbb{C}^m) \setminus \{0\}$ satisfying

$$\|\mathcal{M}_G u_0\|_{L^2(i\mathbb{R};\mathbb{C}^p)} \geq \big(\|G\|_{H^\infty_{p\times m}} - \varepsilon\big)\|u_0\|_{L^2(i\mathbb{R};\mathbb{C}^m)}. \tag{23}$$

Set

$$N \stackrel{\text{def}}{=} \|G\|_{H^\infty_{p\times m}} = \inf\big\{\alpha > 0 \colon \operatorname{meas}\big(\{iw \in i\mathbb{R} \colon \overline{\sigma}\big(G(iw)\big) > \alpha\}\big) = 0\big\}$$

and

$$S \stackrel{\text{def}}{=} \big\{iw \in i\mathbb{R} \colon \overline{\sigma}\big(G(iw)\big) > (N - \varepsilon)\big\} \subset i\mathbb{R}.$$

We have seen in Part I that $\overline{\sigma}\big(G(iw)\big)^2$ is the largest eigenvalue of the matrix $G(iw)^*G(iw)$. We denote by $v(iw)$ its associated eigenvector of norm $|v(iw)| = 1$. Notice that $i\mathbb{R}$ may be written as $i\mathbb{R} = \bigcup_{k=1}^\infty B_{i\mathbb{R}}(0, k)$, where

$$B_{i\mathbb{R}}(0, k) \stackrel{\text{def}}{=} \big\{iw \in i\mathbb{R} \colon |w| < k\big\}.$$

By definition of N, the set S is of positive measure. Thus there must exist an integer k_0 such that $S \cap B_{i\mathbb{R}} \neq \emptyset$ and has a positive measure. We define the vector u_0 by

$$u_0(iw) \stackrel{\text{def}}{=} \chi_{S\cap B_{i\mathbb{R}}}(iw) v(iw)$$

We know that u_0 is nonzero on a set of positive measure. Moreover, its L^2-norm is given by

$$\|u_0\|^2_{L^2(i\mathbb{R};\mathbb{C}^m)} = \int_{-\infty}^{+\infty} \chi_{S\cap B_{i\mathbb{R}}}(iw)|v(iw)|^2 dw = \int_{-\infty}^{+\infty} \chi_{S\cap B_{i\mathbb{R}}}(iw) dw$$
$$= \text{meas}(S \cap i\mathbb{R}) < \infty.$$

We may then notice that the eigenvector $v(iw)$ satisfies

$$\begin{aligned}\overline{\sigma}\big(G(iw)\big)^2|v(iw)| &= |G(iw)^*G(iw)v(iw)| \\ &\leq \|G(iw)^*\||G(iw)v(iw)| = \overline{\sigma}\big(G(iw)\big)|G(iw)v(iw)|\end{aligned}$$

Thus, for any $iw \in S \cap i\mathbb{R}$,

$$|G(iw)v(iw)| \geq \overline{\sigma}\big(G(iw)\big)|v(iw)| \geq (N-\varepsilon)|v(iw)|$$

and for any $iw \in i\mathbb{R}$, $|G(iw)u_0(iw)| \geq (N-\varepsilon)|u_0(iw)|$. Consequently u_0 satisfies (23), which concludes the proof of the first equality.

In order to prove the second equality, we may show that $\mathcal{F}\boldsymbol{\gamma}_g = \mathcal{M}_G\mathcal{F}$. Let $\phi \in L^1(\mathbb{R};\mathbb{C}^m) \cap L^2(\mathbb{R};\mathbb{C}^m)$, then $\boldsymbol{\gamma}_g\phi \in L^1(\mathbb{R};\mathbb{C}^m)$ since

$$\|\boldsymbol{\gamma}_g\phi\|_{L^1(\mathbb{R};\mathbb{C}^m)} = \int_{-\infty}^{+\infty} \Big| \int_0^{+\infty} g(\tau)\phi(t-\tau)d\tau \Big| dt$$
$$\leq \|\phi\|_{L^1(\mathbb{R};\mathbb{C}^m)} \int_0^{+\infty} |g(\tau)|d\tau \ < \infty.$$

Moreover, for $w \in \mathbb{R}$, making use of several changes of variables and of the Fubini theorem, we compute

$$\begin{aligned}(\mathcal{F}\boldsymbol{\gamma}_g\phi)(iw) &= \int_{-\infty}^{+\infty} e^{-iwt}\Big(\int_{-\infty}^{t} g(t-\tau)\phi(\tau)d\tau\Big)dt \\ &= \int_{-\infty}^{+\infty} e^{-iwt}\Big(\int_0^{+\infty} g(\tau)\phi(t-\tau)d\tau\Big)dt \\ &= \int_0^{+\infty} g(\tau)\Big(\int_{-\infty}^{+\infty} e^{-iwt}\phi(t-\tau)dt\Big)d\tau \\ &= \int_0^{+\infty} e^{-iw\tau}g(\tau)(\mathcal{F}\phi)(iw)d\tau \\ &= (\mathcal{L}g)(iw)(\mathcal{F}\phi)(iw) \quad = \quad \big(\mathcal{M}_G(\mathcal{F}\phi)\big)(iw).\end{aligned}$$

Thus for any $\phi \in L^1(\mathbb{R};\mathbb{C}^m) \cap L^2(\mathbb{R};\mathbb{C}^m)$, $\mathcal{F}\boldsymbol{\gamma}_g\phi = \mathcal{M}_G\mathcal{F}\phi$. Now let $u \in L^2(\mathbb{R},\mathbb{C}^m)$ and $\varepsilon > 0$, there exists $\phi^\varepsilon \in L^1(\mathbb{R};\mathbb{C}^m) \cap L^2(\mathbb{R};\mathbb{C}^m)$ such that

$\|u-\phi^{\varepsilon}\|_{L^2(\mathbb{R};\mathbb{C}^m)} \leq \varepsilon$ (using the density of $L^1 \cap L^2$ in L^2) and we may rewrite

$$\mathcal{F}\gamma_g u - \mathcal{M}_G \mathcal{F} u = \mathcal{F}\gamma_g(u-\phi^{\varepsilon}) + \mathcal{M}_G\mathcal{F}(\phi^{\varepsilon}-u)$$

in order to deduce that

$$\|\mathcal{F}\gamma_g u - \mathcal{M}_G\mathcal{F}u\|_{L^2(i\mathbb{R};\mathcal{C}^m)} \leq \varepsilon\big(\|\mathcal{F}\gamma_g\| + \|\mathcal{M}_G\mathcal{F}\|\big).$$

This implies that $\mathcal{F}\gamma_g u = \mathcal{M}_G\mathcal{F}u$. Consequently, for any

$$\begin{aligned}
\sup_{u\in L^2(\mathbb{R};\mathbb{C}^m)\setminus\{0\}} \frac{\|\gamma_g u\|_{L^2(\mathbb{R};\mathbb{C}^p)}}{\|u\|_{L^2(\mathbb{R};\mathbb{C}^m)}} &= \sup_{u\in L^2(\mathbb{R};\mathbb{C}^m)\setminus\{0\}} \frac{\|\mathcal{F}\gamma_g u\|_{L^2(i\mathbb{R};\mathbb{C}^p)}}{\|\mathcal{F}u\|_{L^2(i\mathbb{R};\mathbb{C}^m)}} \\
&= \sup_{u\in L^2(\mathbb{R};\mathbb{C}^m)\setminus\{0\}} \frac{\|\mathcal{M}_G\mathcal{F}u\|_{L^2(i\mathbb{R};\mathbb{C}^p)}}{\|\mathcal{F}u\|_{L^2(i\mathbb{R};\mathbb{C}^m)}} \\
&= \sup_{u\in L^2(\mathbb{R};\mathbb{C}^m)\setminus\{0\}} \frac{\|\mathcal{M}_G\mathcal{F}u\|_{L^2(i\mathbb{R};\mathbb{C}^p)}}{\|\mathcal{F}u\|_{L^2(i\mathbb{R};\mathbb{C}^m)}} \\
&= \sup_{v\in L^2(i\mathbb{R};\mathbb{C}^m)\setminus\{0\}} \frac{\|\mathcal{M}_G v\|_{L^2(i\mathbb{R};\mathbb{C}^p)}}{\|v\|_{L^2(i\mathbb{R};\mathbb{C}^m)}},
\end{aligned}$$

since $\mathcal{F}$ is unitary from $L^2(\mathbb{R};\mathbb{C}^m)$ onto $L^2(i\mathbb{R};\mathbb{C}^m)$. □

***Definition 5.17* (Hankel operator).** — The linear map $\Gamma_g\colon L^2(\mathbb{R}_+;\mathbb{C}^m) \to L^2(\mathbb{R}_+;\mathbb{C}^p)$, defined by

$$(\Gamma_g u)(t) \stackrel{\text{def}}{=} \chi_{[0,+\infty)}(t)\int_0^{+\infty} g(t+\tau)u(\tau)d\tau.$$

is called the *Hankel operator*.

According to the definition given by [**16**, p.1], Γ_g is a Hankel integral operator. Moreover, notice that Γ_g is the restriction to $L^2(\mathbb{R}_+;\mathbb{C}^m)$ of $(\gamma_g \circ \mathfrak{R})$, which is also projected on $L^2(\mathbb{R}_+;\mathbb{C}^p)$, i.e.

$$\Gamma_g \stackrel{\text{def}}{=} \chi_{[0,+\infty)}\Big((\gamma_g\circ\mathfrak{R})\Big|_{L^2(\mathbb{R}_+;\mathbb{C}^m)}\Big).$$

Indeed, let $u \in L^2(\mathbb{R}_+;\mathbb{C}^p)$, then

$$\begin{aligned}
[(\gamma_g\circ\mathfrak{R})u](t) &= \int_{-\infty}^{t} g(t-\tau)u(-\tau)d\tau \\
&= \int_{-t}^{+\infty} g(t+\tau)u(\tau)d\tau \\
&= \int_{0}^{+\infty} g(t+\tau)u(\tau)d\tau
\end{aligned}$$

holds for any $t \in \mathbb{R}$. As a consequence, the norm of the Hankel operator is less than or equal to the norm of γ_g, since

$$\begin{aligned}
\|\Gamma_g\| &= \sup_{u\in L^2(\mathbb{R}_+;\mathbb{C}^m)\setminus\{0\}} \frac{\|\Gamma_g u\|_{L^2(\mathbb{R}_+;\mathbb{C}^p)}}{\|u\|_{L^2(\mathbb{R}_+;\mathbb{C}^m)}} \\
&= \sup_{u\in L^2(\mathbb{R}_+;\mathbb{C}^m)\setminus\{0\}} \frac{\|\chi_{[0,\infty)}\gamma_g \boldsymbol{\mathfrak{R}}\, u\|_{L^2(\mathbb{R}_+;\mathbb{C}^p)}}{\|u\|_{L^2(\mathbb{R}_+;\mathbb{C}^m)}} \\
&= \sup_{u\in L^2(\mathbb{R}_+;\mathbb{C}^m)\setminus\{0\}} \frac{\|\chi_{[0,\infty)}\gamma_g \boldsymbol{\mathfrak{R}}\, u\|_{L^2(\mathbb{R};\mathbb{C}^p)}}{\|u\|_{L^2(\mathbb{R};\mathbb{C}^m)}} \\
&\leq \sup_{u\in L^2(\mathbb{R};\mathbb{C}^m)\setminus\{0\}} \frac{\|\gamma_g \boldsymbol{\mathfrak{R}}\, u\|_{L^2(\mathbb{R};\mathbb{C}^p)}}{\|u\|_{L^2(\mathbb{R};\mathbb{C}^m)}} \\
&= \sup_{u\in L^2(\mathbb{R};\mathbb{C}^m)\setminus\{0\}} \frac{\|\gamma_g \boldsymbol{\mathfrak{R}}\, u\|_{L^2(\mathbb{R};\mathbb{C}^p)}}{\|\boldsymbol{\mathfrak{R}}\, u\|_{L^2(\mathbb{R};\mathbb{C}^m)}} \\
&= \sup_{u\in L^2(\mathbb{R};\mathbb{C}^m)\setminus\{0\}} \frac{\|\gamma_g u\|_{L^2(\mathbb{R};\mathbb{C}^p)}}{\|u\|_{L^2(\mathbb{R};\mathbb{C}^m)}} = \|\gamma_g\|,
\end{aligned}$$

Here the inequality is justified by the fact that $L^2(\mathbb{R}_+,\mathbb{C}^m) \subset L^2(\mathbb{R},\mathbb{C}^m)$ and $\chi_{[0,\infty)}$ only removes nonnegative terms. We also used that $\boldsymbol{\mathfrak{R}}\, u$ has the same L^2-norm as u. As a consequence,

$$\|\Gamma_g\| \leq \|\gamma_g\|.$$

Furthermore, the Hankel operator may be rewritten as

$$\Gamma_g = \Psi\Phi, \tag{24}$$

since for any $u \in L^2(\mathbb{R}_+;\mathbb{C}^n)$,

$$\Psi\Phi u(t) = \chi_{[0,+\infty)}(t) C e^{At}\Phi u = \chi_{[0,+\infty)}(t)\int_0^{+\infty} C e^{(t+\tau)A} B u(\tau) d\tau = \Gamma_g(t).$$

***Remark 5.18* (Order of (S) and rank of Γ_g).** — The rank of the Hankel operator is equal to the order of (**S**). In effect, the decomposition (24), first, allows to observe that $\operatorname{rank}(\Gamma_g) \leq n$. The input map Φ has values in $\mathbb{C}^n$, consequently $\operatorname{Ran}(\Gamma_g) = \Psi(\operatorname{Ran}(\Phi))$ is spanned by at most n elements of $L^2(\mathbb{R}_+;\mathbb{C}^p)$. For example, denoting by $\{e_k\}_{k=1}^n$ the canonical basis of $\mathbb{C}^n$, we see that $\operatorname{Ran}(\Gamma_g)$ is spanned by $\{\Psi e_k\}_{k=1}^n$, since any element of $\operatorname{Ran}(\Gamma_g)$ has the form

$$\Psi x = \Psi\Big(\sum_{k=1}^n \langle x\,, e_k\rangle_{\mathbb{C}^n} e_k\Big) = \sum_{k=1}^n \langle x\,, e_k\rangle_{\mathbb{C}^n} \Psi e_k$$

for some $x \in \mathbb{C}^n$. Furthermore, since we assumed that the system is observable, the kernel of Ψ is equal to zero (see [**8**, Section 2.4.1, p.80]). As a consequence, for any complex numbers $c_1, c_2, \ldots, c_n$ such that

$$c_1\Psi e_1 + c_2\Psi e_2 + \ldots + c_n\Psi e_n = 0,$$

the linearity of Ψ implies that $\Psi(c_1e_1 + c_2e_2 + \ldots + c_ne_n) = 0$, and Ψ being injective we obtain $c_1e_1 + c_2e_2 + \ldots + c_ne_n = 0$. The set $\{e_k\}_{k=1}^n$ being linearly independent, necessarily $c_1 = c_2 = \ldots = c_n = 0$. Hence, $\{\Psi e_k\}_{k=1}^n$ is linearly independent and

$$\operatorname{rank}(\Gamma_g) = n.$$

6. Balanced truncation

We may now introduce the balanced truncation method. A key step of the method is to change coordinates in the state space so that states that are the most controllable are also the most observable. We will latter define how controllability and observability are quantified, and how this definition is related to the Gramians and the singular values of Γ_g.

6.1. Change of coordinates. — Let $T \in \mathbb{C}^{n\times n}$ be an invertible matrix, and consider the change of coordinates in the state space

$$z = T^{-1}\widetilde{z}.$$

The system $(\mathbf{S})$ becomes

$$(\widetilde{\mathbf{S}}):\ \begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}(T^{-1}\widetilde{z})(t) = AT^{-1}\widetilde{z}(t) + Bu(t), \quad \widetilde{z}(0) = Tz_0 \\ y(t) = CT^{-1}\widetilde{z}(t) \end{cases}$$

Since $\frac{\mathrm{d}}{\mathrm{d}t}(T^{-1}\widetilde{z}) = T^{-1}\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{z}$, we can multiply (on the left) the first equation by T and obtain

$$(\widetilde{\mathbf{S}}):\ \begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}\widetilde{z}(t) = \widetilde{A}\widetilde{z}(t) + \widetilde{B}u(t), \qquad \widetilde{z}(0) = Tz_0 \\ y(t) = \widetilde{C}\widetilde{z}(t) \end{cases}$$

where $\widetilde{A} = TAT^{-1}$, $\widetilde{B} = TB$ and $\widetilde{C} = CT^{-1}$. The controllability and observability Gramians for the system $(\widetilde{\mathbf{S}})$ are respectively given by

$$\widetilde{\mathcal{R}} = T\mathcal{R}T^*, \quad \text{and} \quad \widetilde{\mathcal{Q}} = (T^{-1})^*\mathcal{Q}(T^{-1}). \tag{25}$$

Indeed, since $(T^*)^{-1} = (T^{-1})^*$ and since the definition of matrix exponential by power series yields $e^{TAT^{-1}t} = Te^{At}T^{-1}$, we can rewrite the Gramians of

$(\widetilde{\mathbf{S}})$ as

$$\begin{aligned}\widetilde{\mathcal{R}} &= \int_0^\infty e^{t\tilde{A}}\widetilde{B}\widetilde{B}^* e^{t\tilde{A}^*}dt \\ &= \int_0^\infty e^{tTAT^{-1}}TBB^*T^* e^{t(T^{-1})^*A^*T^*}dt \\ &= \int_0^\infty Te^{tA}T^{-1}TBB^*T^*(T^{-1})^* e^{tA^*}T^*dt \\ &= T\Big(\int_0^\infty e^{tA}BB^*e^{tA^*}dt\Big)T^*\end{aligned}$$

and

$$\begin{aligned}\widetilde{\mathcal{Q}} &= \int_0^\infty e^{t\tilde{A}^*}\widetilde{C}^*\widetilde{C}e^{t\tilde{A}}dt \\ &= \int_0^\infty e^{t(T^{-1})^*A^*T^*}(T^{-1})^*C^*CT^{-1}e^{tTAT^{-1}}dt \\ &= \int_0^\infty (T^{-1})^*e^{tA^*}T^*(T^{-1})^*C^*CT^{-1}Te^{tA}T^{-1}dt \\ &= (T^{-1})^*\Big(\int_0^\infty e^{tA^*}C^*Ce^{tA}dt\Big)T^{-1}.\end{aligned}$$

Remark 6.1 **(Input-output relation).** — Notice that $(\widetilde{\mathbf{S}})$ has the same impulse response g as the original system $(\mathbf{S})$,

$$\widetilde{C}e^{t\tilde{A}}\widetilde{B} = (CT^{-1})e^{t(TAT^{-1})}(TB) = C(T^{-1}T)e^{At}(T^{-1}T)B = g(t).$$

This implies that the change of coordinates does not have any effect on the transfer function, and it also appear when computing

$$\begin{aligned}\widetilde{C}(sI_n - \widetilde{A})^{-1}\widetilde{B} &= CT^{-1}(sI_n - TAT^{-1})^{-1}TB \\ &= CT^{-1}\big(T(sI_n - A)T^{-1}\big)^{-1}TB \\ &= C(T^{-1}T)(sI_n - A)^{-1}(T^{-1}T)B \;=\; G(s).\end{aligned}$$

A change of coordinates in the state space does not modify the relationship between inputs and outputs.

6.2. Balancing. — The following Proposition gives a sufficient condition for the existence of a transformation of the state space, that will produce a new system whose Gramians are equal and diagonal. It is a fundamental step of the Balanced truncation methods, because it allows to decide how to truncate the state space.

***Definition 6.2* (Minimal).** — The system (**S**) is called *minimal* if (A, B) is controllable and (A, C) is observable.

***Definition 6.3* (Balanced).** — A system is called *balanced* if its controllability and observability Gramians are equal and diagonal.

***Proposition 6.4* (Balancing transformation).** — *Assume that* (**S**) *is minimal. Then the singular value decomposition*[4] *of* $\mathcal{R}^{\frac{1}{2}}\mathcal{Q}\mathcal{R}^{\frac{1}{2}}$ *is given by*

$$\mathcal{R}^{\frac{1}{2}}\mathcal{Q}\mathcal{R}^{\frac{1}{2}} = U\Sigma^2 U^*,$$

where Σ *is the following diagonal matrix whose diagonal entries are positive and ordered decreasingly:*

$$\Sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{pmatrix}.$$

Furthermore, the invertible matrix $T \overset{def}{=} \Sigma^{\frac{1}{2}} U^* \mathcal{R}^{-\frac{1}{2}}$*, of inverse denoted by* S*, yields a balanced system* $(\widetilde{\mathbf{S}})$ *whose controllability and observability Gramians* $\widetilde{\mathcal{R}}, \widetilde{\mathcal{Q}}$ *are given by*

$$\widetilde{\mathcal{R}} = T\mathcal{R}T^*, \qquad \widetilde{\mathcal{Q}} = S^*\mathcal{Q}S,$$

and satisfy

$$\widetilde{\mathcal{R}} = \widetilde{\mathcal{Q}} = \Sigma.$$

Proof. — The system (**S**) being minimal, the hermitian matrices $\mathcal{R}$, $\mathcal{Q}$ are strictly positive (see Proposition 5.14), hence so is $\mathcal{R}^{\frac{1}{2}}\mathcal{Q}\mathcal{R}^{\frac{1}{2}}$. Consequently there exists a unitary matrix U and a diagonal matrix D such that $\mathcal{R}^{\frac{1}{2}}\mathcal{Q}\mathcal{R}^{\frac{1}{2}} = UDU^*$ and

$$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

4. See the Part I.

with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$ (since $\mathcal{R}^{\frac{1}{2}}\mathcal{Q}\mathcal{R}^{\frac{1}{2}}$ is definite positive). For $k \in \{1, \ldots, n\}$, set $\sigma_k \stackrel{\text{def}}{=} \sqrt{\lambda_k}$ and

$$\Sigma \stackrel{\text{def}}{=} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{pmatrix}$$

in order to bring out that

$$\big(\mathcal{R}^{\frac{1}{2}}U\Sigma^{-\frac{1}{2}}\big)^*\mathcal{Q}\big(\mathcal{R}^{\frac{1}{2}}U\Sigma^{-\frac{1}{2}}\big) = \Sigma^{-\frac{1}{2}}U^*\big(\mathcal{R}^{\frac{1}{2}}\mathcal{Q}\mathcal{R}^{\frac{1}{2}}\big)U\Sigma^{-\frac{1}{2}} = \Sigma.$$

In order to keep the same notation as (25), we set

$$T^{-1} \stackrel{\text{def}}{=} \mathcal{R}^{\frac{1}{2}}U\Sigma^{-\frac{1}{2}}.$$

We have obtained $(T^{-1})^*\mathcal{Q}T^{-1} = \Sigma$. Moreover, $T = \Sigma^{\frac{1}{2}}U^*\mathcal{R}^{-\frac{1}{2}}$ and satisfies

$$T\mathcal{R}T^* = \big(\Sigma^{\frac{1}{2}}U^*\mathcal{R}^{-\frac{1}{2}}\big)\mathcal{R}\big(\mathcal{R}^{-\frac{1}{2}}U\Sigma^{\frac{1}{2}}\big) = \Sigma,$$

the second equation to verify in (25). □

***Remark 6.5* (Singular values of Γ_g).** — The eigenvalues of $\Gamma_g^*\Gamma_g$ and of $\widetilde{\mathcal{R}}\widetilde{\mathcal{Q}}$ are nonnegative since both operators are self-adjoint. Moreover, any σ^2 different from zero is an eigenvalue of $\Gamma_g^*\Gamma_g$, if and only if, it also is an eigenvalue of $\mathcal{R}\mathcal{Q}$ (and of $\widetilde{\mathcal{R}}\widetilde{\mathcal{Q}}$, since the identity $\mathcal{Q}\mathcal{R} = T\widetilde{\mathcal{R}}\widetilde{\mathcal{Q}}T^{-1}$ implies that $\mathcal{R}\mathcal{Q}$ and $\widetilde{\mathcal{R}}\widetilde{\mathcal{Q}}$ have the same eigenvalues $\{\sigma_k^2\}_{k=1}^n$). Here we follow the arguments given by [**12**, section 2.3, p.1121-1122].

First, suppose that $\sigma^2 \neq 0$ is an eigenvalue of $\Gamma_g^*\Gamma_g$, meaning that $\Gamma_g^*\Gamma_g u = \sigma^2 u$ with $u \in L^2(\mathbb{R}_+; \mathbb{C}^m) \setminus \{0\}$. Rewriting Γ_g with the input and output maps we get

$$\Phi^*\Psi^*\Psi\Phi u = \sigma^2 u.$$

Then multiplying this equation on the left by Φ we can observe that

$$\mathcal{R}\mathcal{Q}\Phi u = \Phi\Phi^*\Psi^*\Psi\Phi u = \sigma^2\Phi u.$$

Since $u \neq 0$, Φu cannot be equal to zero. Otherwise, we would have by linearity $\sigma^2 u = \Gamma_g^*\Gamma_g u = (\Phi^*\Psi^*\Psi)\Phi u = 0$. Thus, σ^2 is an eigenvalue of $\mathcal{R}\mathcal{Q}$. Suppose now that σ^2 is an eigenvalue of $\mathcal{R}\mathcal{Q}$, that is

$$\Phi\Phi^*\Psi^*\Psi x = \sigma^2 u$$

where $x \in \mathbb{C}^n$. Multiply this time by $\Phi^*\Psi^*\Psi$ on the left, to get

$$\Gamma_g^*\Gamma_g(\Phi^*\Psi^*\Psi u) = (\Phi^*\Psi^*\Psi)\Phi\Phi^*\Psi^*\Psi x = \sigma^2\Phi^*\Psi^*\Psi u.$$

The vector $\Phi^*\Psi^*\Psi u$ cannot be equal to zero since otherwise by linearity $\mathcal{RQ}u = \Phi(\Phi^*\Psi^*\Psi u)$ would also be equal to zero. The map $\mathcal{RQ}$ being bijective, this would mean that $u = 0$. Thus, σ^2 is an eigenvalue of $\Gamma_g^*\Gamma_g$.

The diagonal elements $\{\sigma_k\}_{k=1}^n$ of the balanced Gramians are thus both singular values of Γ_g and the square roots of the eigenvalues of $\mathcal{RQ}$. They are invariant with respect to a change of coordinates in the state space. This observation leads to the following definition.

***Definition 6.6* (Hankel singular values).** — The square roots of the eigenvalues of $\mathcal{RQ}$ (the product of the controllability and observability Gramians) are called the *Hankel singular values* of the system.

***Remark 6.7* (Norm of Γ_g).** — The norm of the Hankel operator is

$$\|\Gamma_g\| = \sigma_1$$

Indeed, using the preceding remark, as well as the fact that the norm of a self-adjoint operator is equal to its spectral radius (see [**17**, Theorem VI.6, p.192]), we deduce that

$$\|\Gamma_g\|^2 = \|\Gamma_g^*\Gamma_g\| = \max\{\lambda\colon \lambda \in \text{Spectrum}(\Gamma_g^*\Gamma_g)\} = \max\{\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2\} = \sigma_1^2.$$

As a consequence,

$$\sigma_1 = \|\Gamma_g\| \leq \|\gamma_g\| = \|G\|_{H^\infty_{p\times m}}.$$

6.3. Quantification of observability and controllability. — Let $\widetilde{\mathcal{R}} = T\mathcal{R}T^*$ and $\widetilde{Q} = S^*\mathcal{Q}S$ be the balanced Gramians of the system

$$(\widetilde{\mathbf{S}})\colon \begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}\widetilde{z}(t) = \widetilde{A}\widetilde{z}(t) + \widetilde{B}u(t), \quad \widetilde{z}(0) = Tz_0 \\ y(t) = \widetilde{C}\widetilde{z}(t) \end{cases}$$

obtained in the preceding paragraph (Section 6.2). Similarly to section 5.2, we may define the linear and bounded maps $\widetilde{\Psi}\colon \mathbb{C}^n \to L^2(\mathbb{R}_+;\mathbb{C}^p)$ and $\widetilde{\Phi}\colon L^2(\mathbb{R}_+;\mathbb{C}^m) \to \mathbb{C}^n$ defined by

$$(\widetilde{\Psi}x)(t) \overset{\text{def}}{=} \begin{cases} \widetilde{C}e^{\widetilde{A}t}x, & t \geq 0 \\ 0, & t < 0 \end{cases}, \qquad \widetilde{\Phi}u \overset{\text{def}}{=} \int_0^{+\infty} e^{\widetilde{A}\tau}\widetilde{B}u(\tau)d\tau.$$

We have seen that these maps satisfy $\widetilde{\mathcal{R}} = \widetilde{\Phi}\widetilde{\Phi}^*$ and $\widetilde{\mathcal{Q}} = \widetilde{\Psi}^*\widetilde{\Psi}$. Moreover, we remarked that for some $x \in \mathbb{C}^n$,$u \in L^2(\mathbb{R}_+;\mathbb{C}^m)$,

$$\widetilde{\Psi}x = z \quad \text{and} \quad \widetilde{\Phi}u = z^\infty(0),$$

where z and z^∞ are the respective solutions to

$$(\widetilde{\mathbf{S}}_{\mathrm{o}}):\ \begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}z(t) = \widetilde{A}z(t), \quad z(0) = x \\ y(t) = \widetilde{B}z(t), \end{cases}$$

and

$$(\widetilde{\mathbf{S}}_{\mathrm{c}})\ :\ \begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}z^\infty(t) = \widetilde{A}z^\infty(t) + \widetilde{B}(\mathfrak{R}\, u)(t), \quad t \in \mathbb{R}_- \\ \lim\limits_{\tau\to-\infty} z^\infty(\tau) = 0. \end{cases}$$

Let $\{e_k\}_{k=1}^n$ be the canonical basis of $\mathbb{C}^n$. Any element x of $\mathbb{C}^n$ may be rewritten as

$$x = \sum_{k=1}^{n} \big\langle x\,, e_k \big\rangle_{\mathbb{C}^n} e_k,$$

and both Gramians map x to

$$\widetilde{\mathcal{Q}}x = \widetilde{\mathcal{R}}x = \sum_{k=1}^{n} \sigma_k \big\langle x\,, e_k \big\rangle_{\mathbb{C}^n} e_k.$$

We want to give definitions of how much observable and how much controllable the state x is.

6.3.1. *Quantification of observability.* — Let $x \in \mathbb{C}^n$, and consider the system $(\widetilde{\mathbf{S}}_o)$ where there is no input and x is the initial condition. In this system, the output function is $y = \widetilde{\Psi}x$. Observe that the $L^2(\mathbb{R}_+;\mathbb{C}^p)$-norm of y may be rewritten as

$$\|y\|^2_{L^2(\mathbb{R}_+;\mathbb{C}^p)} = \big\langle \widetilde{\Psi}x\,, \widetilde{\Psi}x \big\rangle_{L^2(\mathbb{R}_+;\mathbb{C}^p)} = \big\langle x\,, \widetilde{\mathcal{Q}}x \big\rangle_{\mathbb{C}^n} = \sum_{k=1}^{n} \sigma_k |\langle x\,, e_k\rangle_{\mathbb{C}^n}|^2.$$

Definition 6.8 **(More observable).** — Let $x_1, x_2 \in \mathbb{C}^n$ be of norm one. We say that x_1 *is more observable than* x_2 if $y_1 \overset{\text{def}}{=} \widetilde{\Psi}x_1$ and $y_2 \overset{\text{def}}{=} \widetilde{\Psi}x_2$ satisfy

$$\|y_1\|_{L^2(\mathbb{R}_+;\mathbb{C}^p)} > \|y_2\|_{L^2(\mathbb{R}_+;\mathbb{C}^p)}.$$

Taking into account that the singular values are ordered decreasingly, we remark that a vector $x \in \mathbb{C}^n$ of norm one will lead to an output function of larger norm if its components corresponding to the first directions $e_1, e_2, \dots$ are larger. Also, for two vectors x_1, x_2 both of norm one, if the components of x_1 corresponding to the first directions $e_1, e_2, \dots$ are larger than those of x_2, then x_1 is more observable than x_2. We may then say that the directions $e_1, e_2, \dots$ of the state space $\mathbb{C}^n$ allow to "better observe" a state.

6.3.2. *Quantification of controllability.* — Let $u \in L^2(\mathbb{R}_+;\mathbb{C}^m)$ and consider the system $(\widetilde{\mathbf{S}}_c)$ defined above, whose solution z^∞ satisfies $z^\infty(0) = \widetilde{\Phi}u$. The minimal energy used to reach, via the solution z^∞, a state x at time $t = 0$ is given by

$$\min\{\| \mathfrak{R}\, u\|_{L^2(\mathbb{R}_-;\mathbb{C}^m)} \colon u \in L^2(\mathbb{R}_+;\mathbb{C}^m),\ \widetilde{\Phi}u = x\}$$

if this minimum exists. The following Proposition states that this minimum does exists, and gives a formula for the input function of minimal energy, which involves the controllability Gramian.

Proposition 6.9. — *Let $x \in \mathbb{C}^n$, the minimum*

$$\begin{aligned}&\min\{\|\mathfrak{R}u\|_{L^2(\mathbb{R}_-;\mathbb{C}^m)} \colon u \in L^2(\mathbb{R}_+;\mathbb{C}^m),\ \widetilde{\Phi}u = x\}\\ &= \min\{\|u\|_{L^2(\mathbb{R}_+;\mathbb{C}^m)} \colon u \in L^2(\mathbb{R}_+;\mathbb{C}^m),\ \widetilde{\Phi}u = x\}\end{aligned}$$

is achieved at $u_{opt} \overset{def}{=} \widetilde{\Phi}^*\widetilde{\mathcal{R}}^{-1}x$.

That is, $u_{\rm opt}$ is an input function of minimal norm such that the solution of $(\widetilde{\mathbf{S}}_c)$ attains x at time $t = 0$.

Proof. — This may be shown following the arguments of [**21**, Proposition 1.5.7, p.17], for the operators $\widetilde{\Phi}$ and $\widetilde{\mathcal{R}}$. Indeed, $u_{\rm opt}$ belongs to $L^2(\mathbb{R}_+;\mathbb{C}^m)$ by definition of $\widetilde{\Phi}$, moreover $\widetilde{\Phi}u_{\rm opt} = \widetilde{\Phi}\widetilde{\Phi}^*\widetilde{\mathcal{R}}^{-1}x = \widetilde{\mathcal{R}}\widetilde{\mathcal{R}}^{-1}x = x$. Let $u \in L^2(\mathbb{R}_-;\mathbb{C}^m)$ be any other input function satisfying $\widetilde{\Phi}u = x$. Then, $\widetilde{\Phi}(u_{\rm opt} - u) = 0$ and since $\ker(\Phi) = \mathrm{Ran}(\widetilde{\Phi}^*)^\perp$ we deduce that $w \overset{\rm def}{=} u_{\rm opt} - u$ lies in $\mathrm{Ran}(\widetilde{\Phi}^*)^\perp$. Though, we know from its definition that $u_{\rm opt}$ belongs to $\mathrm{Ran}(\widetilde{\Phi}^*)$. Consequently, the Pythagoras' theorem yields

$$\|u\|^2_{L^2(\mathbb{R}_-;\mathbb{C}^m)} = \|u_{\rm opt}\|^2_{L^2(\mathbb{R}_-;\mathbb{C}^m)} + \|w\|^2_{L^2(\mathbb{R}_-;\mathbb{C}^m)} \geq \|u_{\rm opt}\|^2_{L^2(\mathbb{R}_-;\mathbb{C}^m)}. \qquad \square$$

Definition 6.10 **(More controllable).** — Let $x_1, x_2 \in \mathbb{C}^n$ be of norm one. We say that x_1 *is more controllable than* x_2 if less energy is needed to reach x_1 than to reach x_2, meaning that $u^1_{\rm opt} \overset{\rm def}{=} \widetilde{\Phi}^*\widetilde{\mathcal{R}}^{-1}x_1$ and $u^2_{\rm opt} \overset{\rm def}{=} \widetilde{\Phi}^*\widetilde{\mathcal{R}}^{-1}x_2$ satisfy

$$\|u^1_{\rm opt}\|_{L^2(\mathbb{R}_+;\mathbb{C}^m)} > \|u^2_{\rm opt}\|_{L^2(\mathbb{R}_+;\mathbb{C}^m)}.$$

Observe that the minimal energy needed to reach $x \in \mathbb{C}^n$ is

$$\begin{aligned}\big\|u_{\text{opt}}\big\|^2_{L^2(\mathbb{R}_-;\mathbb{C}^m)} &= \big\langle \widetilde{\Phi}^*\widetilde{\mathcal{R}}^{-1}x\,,\widetilde{\Phi}^*\widetilde{\mathcal{R}}^{-1}x\big\rangle_{L^2(\mathbb{R}_-;\mathbb{C}^m)}\\ &= \big\langle \widetilde{\mathcal{R}}^{-1}x\,,\widetilde{\Phi}\widetilde{\Phi}^*\widetilde{\mathcal{R}}^{-1}x\big\rangle_{\mathbb{C}^n}\\ &= \big\langle \widetilde{\mathcal{R}}^{-1}x\,,x\big\rangle_{\mathbb{C}^n}\\ &= \Big\langle \widetilde{\mathcal{R}}^{-1}\Big(\sum_{k=1}^n\langle x\,,e_k\rangle_{\mathbb{C}^n}e_k\Big)\,,x\Big\rangle_{\mathbb{C}^n} = \sum_{k=1}^n\frac{1}{\sigma_k}|\langle x\,,e_k\rangle_{\mathbb{C}^n}|^2,\end{aligned}$$

and recall that the Hankel singular values satisfy

$$\frac{1}{\sigma_1}\leq\frac{1}{\sigma_2}\leq\ldots\leq\frac{1}{\sigma_n}.$$

Consequently, the minimal energy needed to reach a vector $x \in \mathbb{C}^n$ of norm one will be lower, if the components of x corresponding to the first directions $e_1, e_2, \ldots$ are larger. Also, for two vector x_1, x_2 of norm one, if the components of x_1 corresponding to the first directions $e_1, e_2, \ldots$ are larger than those of x_2, then x_1 is more controllable than x_2. As for the observability, we may then say that the directions $e_1, e_2, \ldots$ of the state space $\mathbb{C}^n$ allow to "better" control" a state.

These remarks suggest that if we truncate the state space $\mathbb{C}^n$ by removing, in the balanced system, the directions

$$e_{\ell+1}, e_{\ell+2}, \ldots, e_n$$

then we should obtain truncated states which will be more observable and controllable than if we had truncated other directions of $\mathbb{C}^n$.

6.4. Truncation. — For this section, in order to lighten the notation, we suppose that the system

$$(\mathbf{S}):\ \begin{cases}\dfrac{\mathrm{d}}{\mathrm{d}t}z(t) = Az(t)+Bu(t), \quad z(0)=z_0\\ y(t)=Cz(t)\end{cases}$$

is already balanced, meaning that the system from which $(\mathbf{S})$ originates is minimal, with a Hurwitz matrix (hence A is also Hurwitz). Consequently, the observability and controllability Gramians of $(\mathbf{S})$ have the form

$$\mathcal{Q}=\mathcal{R}=\begin{pmatrix}\sigma_1 & & 0\\ & \ddots & \\ 0 & & \sigma_n\end{pmatrix}$$

where the Hankel singular values are positive and ordered decreasingly

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n > 0.$$

Recall that $(\mathbf{S})$ and the original system have the same transfer function G and the same impulse response g. We introduce the following partitions of the matrices A, B and C:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} \quad \text{and} \quad C = \begin{pmatrix} C_1 & C_2 \end{pmatrix}$$

where $A_{11} \in \mathbb{C}^{\ell \times \ell}$, $B_1 \in \mathbb{C}^{\ell \times m}$ and $C_1 \in \mathbb{C}^{p \times \ell}$, and $\ell < n$ is such that $\sigma_\ell < \sigma_{\ell+1}$.

The truncation of the state space $\mathbb{C}^n$ consists in retaining the first ℓ components of the states, as hinted in Section 6.3. This equate to defining the approximate system $(\mathbf{S}_\ell)$ by

$$(\mathbf{S}_\ell)\colon \begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} z_\ell(t) = A_{11} z_\ell(t) + B_1 u(t), \quad z_\ell(0) = z_0^\ell \\ y_\ell(t) = C_1 z_\ell(t) \end{cases}$$

where z_0^ℓ is defined by keeping the first ℓ component of z_0. The following Proposition brings out a property of the balanced truncation method: the matrix A_{11} of the approximate system remains Hurwitz.

Proposition 6.11. — *Assume that the system* $(\mathbf{S})$ *is balanced with Hankel singular values* $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n > 0$*, and* A *is a Hurwitz matrix. Then,* A_{11} *is a Hurwitz matrix and the system* $(\mathbf{S}_\ell)$ *defined above is balanced with singular values* $\sigma_1 \geq \ldots \geq \sigma_\ell$*.*

Remark 6.12. — Since A_{11} is Hurwitz, the transfer function of $(\mathbf{S}_\ell)$ belongs to $H^\infty_{p \times m}$.

Proof. — Defining the diagonal matrices $\Sigma_1 \in \mathbb{C}^{\ell \times \ell}$ and $\Sigma_2 \in \mathbb{C}^{(n-\ell) \times (n-\ell)}$ by

$$\Sigma_1 = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_\ell \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} \sigma_{\ell+1} & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{pmatrix},$$

which have no common eigenvalues since the Hankel singular values are ordered decreasingly and $\sigma_\ell > \sigma_{\ell+1}$, we rewrite the Gramians of $(\mathbf{S})$ as

$$\mathcal{Q} = \mathcal{R} = \begin{pmatrix} \Sigma_1 & \mathbb{O}_{\ell, n-\ell} \\ \mathbb{O}_{n-\ell, \ell} & \Sigma_2 \end{pmatrix}.$$

A first observation is that, since A is Hurwitz, Σ satisfies the Lyapunov equations (see Proposition 5.13)

$$\begin{aligned}(26)\qquad &\begin{pmatrix} A_{11}^* & A_{21}^* \\ A_{12}^* & A_{22}^* \end{pmatrix} \begin{pmatrix} \Sigma_1 & \mathbb{O}_{\ell,n-\ell} \\ \mathbb{O}_{n-\ell,\ell} & \Sigma_2 \end{pmatrix} \\ &+ \begin{pmatrix} \Sigma_1 & \mathbb{O}_{\ell,n-\ell} \\ \mathbb{O}_{n-\ell,\ell} & \Sigma_2 \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} + \begin{pmatrix} C_1^* \\ C_2^* \end{pmatrix} \begin{pmatrix} C_1 & C_2 \end{pmatrix} = 0\end{aligned}$$

and

$$\begin{aligned}(27)\qquad &\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} \Sigma_1 & \mathbb{O}_{\ell,n-\ell} \\ \mathbb{O}_{n-\ell,\ell} & \Sigma_2 \end{pmatrix} \\ &+ \begin{pmatrix} \Sigma_1 & \mathbb{O}_{\ell,n-\ell} \\ \mathbb{O}_{n-\ell,\ell} & \Sigma_2 \end{pmatrix} \begin{pmatrix} A_{11}^* & A_{21}^* \\ A_{12}^* & A_{22}^* \end{pmatrix} + \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} \begin{pmatrix} B_1^* & B_2^* \end{pmatrix} = 0.\end{aligned}$$

Consequently, Σ_1 is the solution to the Lyapunov equations

$$A_{11}^*\Sigma_1 + \Sigma_1 A_{11} + C_1^* C_1 = 0 \tag{28}$$

$$A_{11}\Sigma_1 + \Sigma_1 A_{11}^* + B_1 B_1^* = 0. \tag{29}$$

Let $\lambda \in \mathbb{C}$ be any eigenvalue of A_{11}, our aim is to show that necessarily $\Re(\lambda) < 0$. Fix $\{v_1, \dots, v_\kappa\}$ a basis for the eigenspace $\ker(A_{11} - \lambda I_\ell)$ associated to λ. Denoting by V the matrix whose columns are the vectors v_k

$$V = \begin{pmatrix} | & | & & | \\ v_1 & v_2 & \dots & v_\kappa \\ | & | & & | \end{pmatrix},$$

we observe that the eigenspace associated to λ is also the Range of V. This implies that A_{11} maps any vector x belonging $\mathrm{Ran}(V)$ to λx, in other words the map $A_{11}V \in \mathscr{L}(\mathbb{C}^\kappa, \mathbb{C}^n)$ only multiplies vectors by λ. We will use this property and the Lyapunov equations satisfied by Σ_1 to uncover the sign of $\Re(\lambda)$. Multiply equation (28) on the left by V^* and on the right by V, this yields

$$\begin{aligned} & (A_{11}V)^*\Sigma_1 V + V^*\Sigma_1(A_{11}V) + V^* C_1^* C_1 V = 0 \\ \Leftrightarrow \quad & (\lambda^* + \lambda) V^*\Sigma_1 V + V^* C_1^* C_1 V = 0 \\ \Leftrightarrow \quad & 2\Re(\lambda) V^*\Sigma_1 V = -(C_1 V)^* C_1 V. \end{aligned}$$

Choosing some nonzero $x \in \mathbb{C}^\kappa$ and using that Σ_1 is positive definite and $(C_1V)^*C_1V$ is nonnegative, we deduce that

$$\Re(\lambda) = -\frac{\left\langle (C_1V)^*C_1Vx\,,x\right\rangle_{\mathbb{C}^n}}{2\left\langle \Sigma Vx\,,Vx\right\rangle_{\mathbb{C}^n}} \leq 0$$

(note that Vx cannot be zero since V is injective). It remains to prove that the real part of λ cannot be zero. Suppose by contradiction that $\lambda = iw$, with $w \in \mathbb{R}$ (we will come to the contradiction that iw is also an eigenvalue of A, while A is Hurwitz). In that case, the preceding computation also yields $(C_1V)^*C_1V = 0$, thus

$$C_1V = 0$$

Multiplying the equation (28), this time only on right by V, yields

$$\begin{aligned} & A_{11}^*\Sigma_1V + \Sigma_1A_{11}V = 0 \\ \Leftrightarrow \quad & A_{11}^*\Sigma_1V = -iw\Sigma_1V. \end{aligned}$$

Now using the equation (29), multiplied on the left by $V^*\Sigma_1$ and on the right by Σ_1V, we obtain

$$\begin{aligned} & (A_{11}^*\Sigma_1V)^*\Sigma_1^2V + V^*\Sigma_1^2(A_{11}^*\Sigma_1V) + V^*\Sigma_1B_1B_1^*\Sigma_1V = 0 \\ \Leftrightarrow \quad & iwV^*\Sigma_1^3V + -iwV^*\Sigma_1^3V + V^*\Sigma_1B_1B_1^*\Sigma_1V = 0 \\ \Leftrightarrow \quad & (B_1^*\Sigma_1V)^*(B_1^*\Sigma_1V) = 0, \end{aligned}$$

which implies that

$$B_1^*\Sigma_1V = 0.$$

With this new information, multiplying the equation (29) on the right by Σ_1V, we obtain the following relation

$$\begin{aligned} & A_{11}\Sigma_1^2V + \Sigma_1(A_{11}^*\Sigma_1V) = 0 \\ \Leftrightarrow \quad & A_{11}(\Sigma_1^2V) = iw(\Sigma_1^2V). \end{aligned}$$

In other words, $\mathrm{Ran}(\Sigma_1^2V) \subset \mathrm{Ran}(V)$ (since $\mathrm{Ran}(V)$ is the eigenspace associated to λ), which also means $\Sigma_1^2(\mathrm{Ran}(V)) \subset \mathrm{Ran}(V)$. Choose an eigenvalue of $\Sigma_1^2\big|_{\mathrm{Ran}(V)}$, it also is an eigenvalue of Σ_1^2 (since $\mathrm{Ran}(V) \subset \mathbb{C}^n$), consequently it has the form σ^2 with $\sigma > 0$. Take $x \in \mathrm{Ran}(V) \setminus \{0\}$ an eigenvector associated to σ^2, this vector satisfies

$$\Sigma_1^2x = \sigma^2x \qquad \text{and} \qquad A_{11}x = iwx.$$

We claim that $\begin{pmatrix} x \\ \mathbb{O}_{n-\ell,1} \end{pmatrix}$ is an eigenvalue of A associated with the eigenvalue iw, that is

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x \\ \mathbb{O}_{n-\ell,1} \end{pmatrix} = iw \begin{pmatrix} x \\ \mathbb{O}_{n-\ell,1} \end{pmatrix}$$

and this contradicts the hypothesis stating that A is Hurwitz. To prove the claim, it only remains to check that $A_{21}x = 0$. Multiply on the right the equations (26) and (27) respectively by

$$\begin{pmatrix} x \\ \mathbb{O}_{n-\ell,1} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \Sigma_1 x \\ \mathbb{O}_{n-\ell,1} \end{pmatrix}.$$

Since $C_1V = 0$ and $B_1^*\Sigma_1 V = 0$, the terms involving B and C are null and we get

$$\begin{pmatrix} A_{11}^* & A_{21}^* \\ A_{12}^* & A_{22}^* \end{pmatrix} \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \begin{pmatrix} x \\ \mathbb{O}_{n-\ell,1} \end{pmatrix} + \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x \\ \mathbb{O}_{n-\ell,1} \end{pmatrix} = 0$$

which is equivalent to

$$\begin{cases} A_{11}^*\Sigma_1 x + \Sigma_1 A_{11} x = 0 \\ A_{12}^*\Sigma_1 x + \Sigma_2 A_{21} x = 0 \end{cases}$$

and

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \begin{pmatrix} \Sigma_1 x \\ \mathbb{O}_{n-\ell,1} \end{pmatrix} + \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \begin{pmatrix} A_{11}^* & A_{21}^* \\ A_{12}^* & A_{22}^* \end{pmatrix} \begin{pmatrix} \Sigma_1 x \\ \mathbb{O}_{n-\ell,1} \end{pmatrix} = 0$$

which is equivalent to

$$\begin{cases} A_{11}\Sigma_1^2 x + \Sigma_1 A_{11}^*\Sigma_1 x = 0 \\ A_{21}\Sigma_1^2 x + \Sigma_2 A_{12}^*\Sigma_1 x = 0 \end{cases}.$$

From the second row of each equation, we deduce that $\Sigma_2^2 A_{21}x = A_{21}\Sigma_1^2 x$, and having chosen x as an eigenvector of Σ_1 associated with σ^2, we obtain

$$\Sigma_2^2 A_{21} x = \sigma^2 A_{21} x.$$

Remember that σ^2 is an eigenvalue of Σ_1^2, thus it cannot also be an eigenvalue of Σ_2^2 (since the diagonal matrices Σ_1 and Σ_2 have different eigenvalues) and $A_{21}x$ has to be equal to zero.

Having proven that A_{11} is Hurwitz and having observed that Σ_1 satisfies the Lyapunov equations (28) and (29), we can now conclude that the Gramians $\mathcal{R}^\ell$ and $\mathcal{Q}^\ell$ of (S_ℓ) are both equal to Σ_1 (again Proposition 5.13). □

7. Error of the balanced truncation

The balanced truncation method has the advantage that a priori error bounds for the error are available.

7.1. Lower bound for the error. — The Proposition that follows stresses that there is a limit to how well one can approximate (**S**) by the truncated system ($\mathbf{S}_\ell$). Before considering the error between transfer functions, we look at the error committed when approximating a square matrix by any other matrix whose rank is strictly less than the dimension (of the matrices).

Lemma 7.1. — *Suppose that $M \in \mathbb{C}^{n\times n}$ is any square matrix whose singular value decomposition is given bellow*

$$M = U \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} V^*$$

(where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ and U, V are unitary matrices). Let ℓ be an integer less than n. Take M_ℓ any square matrix in $\mathbb{C}^{n\times n}$ whose rank is less than or equal to ℓ. Then, the norm of the difference $M - M_\ell$ is minored, with

$$\|M - M_\ell\| \geq \lambda_{\ell+1}.$$

Proof. — Let the columns of U and V be denoted by $u_1, u_2, \ldots, u_n$ and $v_1, v_2, \ldots, v_n$ respectively. Recall that $\{u_1, u_2, \ldots, u_n\}$ and $\{v_1, v_2, \ldots, v_n\}$ are both orthonormal sets. The dimension of $\operatorname{span}\{v_1, \ldots, v_{\ell+1}\}$ is $\ell+1$. Whereas, since $\operatorname{rank}(M_\ell) \leq \ell$, the dimension of $\ker(M_\ell)$ has to be greater than $n - \ell$. Hence, the spaces $\operatorname{span}\{v_1, \ldots, v_{\ell+1}\}$ and $\ker(M_\ell)$ share a nonzero element, which will be denoted by x. From its definition we observe that x satisfies

$$(M - M_\ell)x = Mx = M\Big(\sum_{k=1}^{\ell+1} \langle x\,, v_k\rangle_{\mathbb{C}^n} v_k\Big) = \sum_{k=1}^{\ell+1} \langle x\,, v_k\rangle_{\mathbb{C}^n} \sigma_k u_k.$$

The singular values of M being ordered decreasingly, we may bound from bellow the norm of $(M - M_\ell)x$ the following way

$$|(M - M_\ell)x|^2 = \sum_{k=1}^{\ell+1} |\langle x\,, v_k\rangle_{\mathbb{C}^n}|^2 \lambda_k^2 \geq \lambda_{\ell+1}^2 \sum_{k=1}^{\ell+1} |\langle x\,, v_k\rangle_{\mathbb{C}^n}|^2 = \lambda_{\ell+1}^2 |x|^2.$$

Thus,

$$\|M - M_\ell\| \geq \frac{|(M - M_\ell)x|}{|x|} \geq \lambda_{\ell+1} \qquad \square$$

As we have seen before, there is a connection between the $H^\infty_{p\times m}$-norm of the transfer function and the norm of the input-output map (whose norm is σ_1, the largest Hankel singular value):

$$\|G\|_{H^\infty_{p\times m}} \geq \|\Gamma_g\|.$$

Using this inequality and Lemma 7.1, we will be able to bound from bellow the H^∞-error in term of the Hankel singular values.

Proposition 7.2. — *Consider the transfer function G of the system (S), whose Hankel singular values are $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n > 0$. Let the transfer function of the truncated system be denoted by G_ℓ. Then, the error between the system (S) and the truncated system admits a lower bound, with*

$$\|G - G_\ell\|_{H^\infty_{p\times m}} \geq \sigma_{\ell+1}$$

Proof. — The impulse response of the truncated system will be denoted by g_ℓ, meaning that $g_\ell(t) = C_1 e^{tA_{11}} B_1$. Define the transfer function of the error system

$$E \stackrel{\text{def}}{=} G - G_\ell.$$

As G, G_ℓ belong to $H^\infty(\mathbb{C}_+;\mathbb{C}^{p\times m})$, the transfer function of the error system also belongs to this space (since $H^\infty(\mathbb{C}_+;\mathbb{C}^{p\times m})$ is a vector space). The impulse response g_{err} of the error system is given by $g - g_\ell$, since

$$\mathcal{L}(g - g_\ell) = \mathcal{L}(g) - \mathcal{L}(g_\ell) = G - G_\ell = E \stackrel{\text{def}}{=} \mathcal{L}g_{\text{err}}$$

(see the uniqueness theorem given by [**3**, Theorem 1.7.3, p.40]). Hence, the Hankel operator of E is Γ_{g-g_ℓ} (which is equal to $\Gamma_g - \Gamma_{g_\ell}$, by linearity of the integral) and satisfies

$$\|E\|_{H^\infty_{p\times m}} \geq \|\Gamma_g - \Gamma_{g_\ell}\|.$$

(see section 5.3). In order to obtain the desired result, we have to show that the right-hand side is bounded from bellow by $\sigma_{\ell+1}$. Both Hankel operators $\Gamma_g, \Gamma_{g_\ell}$ are defined from $L^2(\mathbb{R}_+;\mathbb{C}^m)$ into $L^2(\mathbb{R}_+;\mathbb{C}^p)$. We may multiply them

on the left and on the right by operators of norm one (respectively denoted by P_o and P_c), so that

$$\|\Gamma_g - \Gamma_{g_\ell}\| = \|P_o\|\,\|\Gamma_g - \Gamma_{g_\ell}\|\,\|P_c\| \geq \|P_o\Gamma_g P_c - P_o\Gamma_{g_\ell}P_c\|.$$

Hence, in order to conclude, we will have to bound from bellow the right-hand side above. Moreover, if P_c takes values in $\mathbb{C}^n$ and P_o goes into $\mathbb{C}^n$ then $(P_o\Gamma_g P_c), (P_o\Gamma_{g_\ell}P_c)$ are square matrices, and $\operatorname{rank}(P_o\Gamma_{g_\ell}P_c) \leq \operatorname{rank}(\Gamma_{g_\ell}) = \ell$. Denoting by $\lambda_1 \geq \ldots \geq \lambda_n$ the singular values of $P_o\Gamma_g P_c$, we would obtain from Lemma 7.1 that

$$\|P_o\Gamma_g P_c - P_o\Gamma_{g_\ell}P_c\| \geq \lambda_{\ell+1}.$$

Taking into account these observations, the desired result will be obtained if P_o and P_c are additionally chosen so that the singular values of $P_o\Gamma_g P_c$ are the Hankel singular values $\sigma_1, \ldots, \sigma_n$. Define $P_o\colon L^2(\mathbb{R}_+;\mathbb{C}^p) \to \mathbb{C}^n$ and $P_c\colon \mathbb{C}^n \to L^2(\mathbb{R}_+;\mathbb{C}^m)$ by

$$P_o \stackrel{\text{def}}{=} \mathcal{Q}^{-\frac{1}{2}}\Psi^* \qquad \text{and} \qquad P_c \stackrel{\text{def}}{=} \Phi^*\mathcal{R}^{-\frac{1}{2}}.$$

They are of unit norm since for any $y \in L^2(\mathbb{R}_+;\mathbb{C}^p)$,

$$\begin{aligned} |P_o y|^2 &= \left\langle \mathcal{Q}^{-\frac{1}{2}}\Psi y\,,\,\mathcal{Q}^{-\frac{1}{2}}\Psi y\right\rangle_{\mathbb{C}^n} \\ &= \left\langle \Psi\mathcal{Q}^{-1}\Psi^* y\,,\,y\right\rangle_{L^2(\mathbb{R}_+;\mathbb{C}^p)} \\ &= \left\langle \Psi(\Psi^*\Psi)^{-1}\Psi^* y\,,\,y\right\rangle_{L^2(\mathbb{R}_+;\mathbb{C}^p)} \;=\; |y|^2_{L^2(\mathbb{R}_+;\mathbb{C}^p)}, \end{aligned}$$

and for any $x \in \mathbb{C}^n$,

$$\begin{aligned} \|P_c x\|^2_{L^2(\mathbb{R}_+;\mathbb{C}^m)} &= \left\langle \Phi^*\mathcal{R}^{-\frac{1}{2}}x\,,\,\Phi^*\mathcal{R}^{-\frac{1}{2}}x\right\rangle_{L^2(\mathbb{R}_+;\mathbb{C}^m)} \\ &= \left\langle \mathcal{R}^{-\frac{1}{2}}\Phi\Phi^*\mathcal{R}^{-\frac{1}{2}}x\,,\,x\right\rangle_{\mathbb{C}^n} \\ &= \left\langle \mathcal{R}^{-\frac{1}{2}}\mathcal{R}\mathcal{R}^{-\frac{1}{2}}x\,,\,x\right\rangle_{\mathbb{C}^n} \;=\; |x|^2 \end{aligned}$$

By definition, the singular values of the matrix $P_o\Gamma_g P_c$, which rewrites as

$$P_o\Gamma_g P_c = (\mathcal{Q}^{-\frac{1}{2}}\Psi^*)(\Psi\Phi)(\Phi^*\mathcal{R}^{-\frac{1}{2}}) = \mathcal{Q}^{-\frac{1}{2}}\mathcal{Q}\mathcal{R}\mathcal{R}^{-\frac{1}{2}} = \mathcal{Q}^{\frac{1}{2}}\mathcal{R}^{\frac{1}{2}},$$

are the square roots of the eigenvalues of

$$\left(\mathcal{Q}^{\frac{1}{2}}\mathcal{R}^{\frac{1}{2}}\right)^*\left(\mathcal{Q}^{\frac{1}{2}}\mathcal{R}^{\frac{1}{2}}\right) = \mathcal{R}^{\frac{1}{2}}\mathcal{Q}\mathcal{R}^{\frac{1}{2}}.$$

However, we have seen in Proposition 6.4 that the Hankel singular values may be found by diagonalizing the matrix $\mathcal{R}^{\frac{1}{2}}\mathcal{Q}\mathcal{R}^{\frac{1}{2}}$ and

$$\mathcal{R}^{\frac{1}{2}}\mathcal{Q}\mathcal{R}^{\frac{1}{2}} = U \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix} U^*.$$

This implies that the singular values of $P_o\Gamma_g P_c$ are the Hankel singular values. □

7.2. Upper bound for the error. — As in section 6.4, we assume that (**S**) is already balanced, originates from a minimal system with a Hurwitz matrix, and we denote the Hankel singular values by

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n > 0.$$

In order to prove the upper bound for the H^∞-error, we will have to take into account that there might be repeated values among the $n-\ell$ Hankel singular values

$$\sigma_{\ell+1} \geq \sigma_{\ell+2} \geq \ldots \geq \sigma_n > 0. \tag{30}$$

Hence, we rewrite (30) as

$$\underbrace{\sigma_{i_m} = \ldots = \sigma_{i_m}}_{n_m \text{ times}} > \underbrace{\sigma_{i_{m-1}} = \ldots = \sigma_{i_{m-1}}}_{n_{m-1} \text{ times}} > \ldots > \underbrace{\sigma_{i_1} = \ldots = \sigma_{i_1}}_{n_1 \text{ times}} > 0$$

where m is the number of distinct elements in (30) and n_k denotes how many times the Hankel singular value σ_{i_k} appears in (30) (hence, $n-\ell = n_1 + n_2 + \ldots + n_m$). Consequently, the Gramians $\mathcal{Q} = \mathcal{R}$ of (**S**) also have the shape

$$\begin{pmatrix} \sigma & & & & & & 0 \\ & \ddots & & & & & \\ & & \sigma_\ell & & & & \\ & & & \sigma_{\ell+1} & & & \\ & & & & \ddots & & \\ 0 & & & & & & \sigma_n \end{pmatrix} = \left(\begin{array}{c|ccc} \Sigma & & & 0 \\ \hline & \sigma_{i_m} I_{n_m} & & \\ & & \ddots & \\ & & & \sigma_{i_2} I_{n_2} \\ 0 & & & \sigma_{i_1} I_{n_1} \end{array}\right)$$

with

$$\Sigma \stackrel{\text{def}}{=} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_\ell \end{pmatrix}.$$

Remark 7.3. — Notice that, contrary to the previous notation chosen for the Hankel singular values, σ_{i_m} denotes the largest of the remaining Hankel singular values in (30), while σ_{i_1} is the smallest. This choice of notation is made to the advantage of the proof of Theorem 7.6.

The Proposition that follows is the principal step to deduce the upper bound for the H^∞-error. During this step we will assume that all the truncated directions $e_{\ell+1}, e_{\ell+2}, \ldots, e_n$ correspond to the same Hankel singular value, i.e.

$$\sigma_{\ell+1} = \sigma_{\ell+2} = \ldots = \sigma_n.$$

After this step, we will truncate iteratively the system (**S**), removing at each iteration directions e_k corresponding to the same Hankel singular value. Then, applying this Proposition 7.5 at each iteration, will allow to deduce the estimate of the H^∞-error.

Lemma 7.4. — *Let $\widetilde{A} \in \mathbb{C}^{\widetilde{n}}$ be a Hurwitz matrix, and define the transfer function*

$$E(s) \stackrel{def}{=} \left[\begin{array}{c|c} \widetilde{A} & \widetilde{B} \\ \hline \widetilde{C} & \widetilde{D} \end{array}\right],$$

where $\widetilde{B} \in \mathbb{C}^{\widetilde{n}\times\widetilde{m}}$, $\widetilde{C} \in \mathbb{C}^{\widetilde{p}\times\widetilde{n}}$ and $\widetilde{D} \in \mathbb{C}^{\widetilde{p}\times\widetilde{m}}$. Denote by $\widetilde{\mathcal{Q}}$ the observability Gramian of the system associated with $E(s)$. If the following equality holds

$$\widetilde{C}^*\widetilde{D} + \widetilde{\mathcal{Q}}\widetilde{B} = 0, \tag{31}$$

then the norm of E is given by

$$\big\|E\big\|^2_{H^\infty_{\widetilde{p}\times\widetilde{m}}} = \|\widetilde{D}^*\widetilde{D}\|.$$

Proof. — We define the transfer function $E\,\tilde{}\,(s)$ (called the *para-Hermitian conjugate of* $E(s)$) by

$$E\,\tilde{}\,(s) \stackrel{\text{def}}{=} \left[\begin{array}{c|c} -\widetilde{A}^* & -\widetilde{C}^* \\ \hline \widetilde{B}^* & \widetilde{D}^* \end{array}\right] = \widetilde{B}^*(-sI - \widetilde{A}^*)^{-1}\widetilde{C}^* + \widetilde{D}^*.$$

This para-Hermitian conjugate is useful because on the imaginary axis (for $w \in \mathbb{R}$), $E^{\sim}(iw)$ agrees with the adjoint of $E(iw)$, which is given by

$$E(s)^* = \widetilde{B}^*(\overline{s}I - \widetilde{A}^*)^{-1}\widetilde{C}^* + \widetilde{D}^*.$$

Notice that the matrix $E(iw)^*$ is well defined since $\widetilde{A}^*$ is Hurwitz (hence, iw is not an eigenvalue of A^*). In order to prove this Lemma, our aim will be to show that if $E^{\sim}(s)E(s)$ satisfies (31), then it is necessarily constant, with

$$E^{\sim}(s)E(s) = \widetilde{D}^*\widetilde{D}, \tag{32}$$

for any $s \in \mathbb{C}$ such that $E^{\sim}(s)$ and $E(s)$ are well defined. This property would permit to rewrite the $H^{\infty}_{\widetilde{p}\times\widetilde{m}}$-norm of $E(s)$ as

$$\begin{aligned}\big\|E\big\|^2_{H^{\infty}_{\widetilde{p}\times\widetilde{m}}} &= \sup_{w\in\mathbb{R}} \|E(iw)\|^2 \\ &= \sup_{w\in\mathbb{R}} \|E(iw)^*E(iw)\| \\ &= \sup_{w\in\mathbb{R}} \|E^{\sim}(iw)E(iw)\| \;=\; \|\widetilde{D}^*\widetilde{D}\|.\end{aligned}$$

The product $E^{\sim}(s)E(s)$ is given by

$$E^{\sim}(s)E(s) = \left[\begin{array}{cc|c} -\widetilde{A}^* & -\widetilde{C}^*\widetilde{C} & -\widetilde{C}^*\widetilde{D} \\ \mathbb{O}_{\widetilde{n}} & A & \widetilde{B} \\ \hline \widetilde{B}^* & \widetilde{D}^*\widetilde{C} & \widetilde{D}^*\widetilde{D} \end{array}\right]$$

(see the Appendix, Proposition 10.3). A change of coordinate in the state space does not affect the transfer function, we define the change of coordinate T (and give its inverse T^{-1}) by

$$T \stackrel{\text{def}}{=} \begin{pmatrix} I_{\widetilde{n}} & -\widetilde{\mathcal{Q}} \\ \mathbb{O}_{\widetilde{n}} & I_{\widetilde{n}} \end{pmatrix}, \qquad T^{-1} = \begin{pmatrix} I_{\widetilde{n}} & \widetilde{\mathcal{Q}} \\ \mathbb{O}_{\widetilde{n}} & I_{\widetilde{n}} \end{pmatrix}.$$

Then, as we have seen in the section 6.1, the product may be rewritten as

$$\begin{aligned} E^{\sim}(s)E(s) &= \left[\begin{array}{c|c} T\begin{pmatrix} -\widetilde{A}^* & -\widetilde{C}^*\widetilde{C} \\ \mathbb{O}_{\widetilde{n}} & A \end{pmatrix}T^{-1} & T\begin{pmatrix} -\widetilde{C}^*\widetilde{D} \\ \widetilde{B} \end{pmatrix} \\ \hline \begin{pmatrix} \widetilde{B}^* & \widetilde{D}^*\widetilde{C} \end{pmatrix}T^{-1} & \widetilde{D}^*\widetilde{D} \end{array}\right] \\ &= \left[\begin{array}{cc|c} -\widetilde{A}^* & -\big(\widetilde{A}^*\widetilde{\mathcal{Q}} + \widetilde{\mathcal{Q}}\widetilde{A} + \widetilde{C}^*\widetilde{C}\big) & -\big(\widetilde{C}^*\widetilde{D} + \widetilde{\mathcal{Q}}\widetilde{B}\big) \\ \mathbb{O}_{\widetilde{n}} & A & \widetilde{B} \\ \hline \widetilde{B}^* & \big(\widetilde{C}^*\widetilde{D} + \widetilde{\mathcal{Q}}\widetilde{B}\big)^* & \widetilde{D}^*\widetilde{D} \end{array}\right]. \end{aligned}$$

Since $\widetilde{A}$ is Hurwitz, the observability Gramian $\mathcal{Q}$ is solution to the Lyapunov equation $\widetilde{A}^*\widetilde{\mathcal{Q}} + \widetilde{\mathcal{Q}}\widetilde{A} + \widetilde{C}^*\widetilde{C} = 0$. This, in addition to (31), implies that

$$E\,\tilde{}\,(s)E(s) = \left[\begin{array}{cc|c} -\widetilde{A}^* & \mathbb{O}_{\widetilde{n}} & \mathbb{O}_{\widetilde{n},\widetilde{m}} \\ \mathbb{O}_{\widetilde{n}} & A & \widetilde{B} \\ \hline \widetilde{B}^* & \mathbb{O}_{\widetilde{n}} & \widetilde{D}^*\widetilde{D} \end{array}\right]$$

$$= \begin{pmatrix} \widetilde{B}^* & \mathbb{O}_{\widetilde{n}} \end{pmatrix} \left(sI_{2\widetilde{n}} - \begin{pmatrix} -\widetilde{A}^* & \mathbb{O}_{\widetilde{n}} \\ \mathbb{O}_{\widetilde{n}} & A \end{pmatrix} \right)^{-1} \begin{pmatrix} \mathbb{O}_{\widetilde{n},\widetilde{m}} \\ \widetilde{B} \end{pmatrix} + \widetilde{D}^*\widetilde{D}.$$

The first term in the right-hand side being the matrix $\mathbb{O}_{\widetilde{m}}$, we deduce that (32) holds. □

Proposition 7.5. — *Suppose that* (**S**) *is already balanced, with Hankel singular values*

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_\ell \geq \sigma_{\ell+1} = \sigma_{\ell+2} = \ldots = \sigma_n.$$

Then the H^∞-error of approximation admits the following upper bound

$$\|G - G_\ell\|_{H^\infty_{p\times m}} \leq 2\sigma_{\ell+1}.$$

Proof. — Let $E_{11}(s) \in \mathbb{C}^{p\times m}$ be the following transfer function

$$E_{11}(s) \stackrel{\text{def}}{=} G(s) - G_\ell(s) = \left[\begin{array}{ccc|c} A_{11} & \mathbb{O}_\ell & \mathbb{O}_{\ell,n-\ell} & B_1 \\ \mathbb{O}_\ell & A_{11} & A_{12} & B_1 \\ \mathbb{O}_{n-\ell,\ell} & A_{21} & A_{22} & B_2 \\ \hline -C_1 & C_1 & C_2 & \mathbb{O}_{p,m} \end{array}\right] \stackrel{\text{def}}{=} \left[\begin{array}{c|c} A_\ell & B_\ell \\ \hline C_\ell & D_\ell \end{array}\right].$$

(see the Appendix, Proposition 10.3), where $A_\ell \in \mathbb{C}^{(n+\ell)\times(n+\ell)}$, $B_\ell \in \mathbb{C}^{(n+\ell)\times m}$, $C_\ell \in \mathbb{C}^{p\times(n+\ell)}$ and $D_\ell \in \mathbb{C}^{p\times m}$. Our aim is to show that the $H^\infty_{p\times m}$-norm of this transfer function is less than $2\sigma_{\ell+1}$. We are going to make use of Lemma 7.4 by building another transfer function $E(s) \in \mathbb{C}^{(p+m)\times(m+p)}$

$$E(s) \stackrel{\text{def}}{=} \left[\begin{array}{c|c} \widetilde{A} & \widetilde{B} \\ \hline \widetilde{C} & \widetilde{D} \end{array}\right]$$

such that $E_{11}(s)$ is a submatrix of $E(s)$, i.e.

$$E(s) = \begin{pmatrix} E_{11}(s) & E_{12}(s) \\ E_{21}(s) & E_{22}(s) \end{pmatrix},$$

and whose $H^\infty_{(p+m)\times(m+p)}$-norm is equal to the bound $2\sigma_{\ell+1}$. This way, we will be allowed to conclude that

$$\big\|E_{11}\big\|_{H^\infty_{p\times m}} \leq \big\|E\big\|_{H^\infty_{(p+m)\times(m+p)}} = 2\sigma_{\ell+1},$$

where the first inequality comes from noticing that for any $s \in \mathbb{C}_+$, we have $\|E_{11}(s)\| \leq \|E(s)\|$. Indeed, denoting by $c_k(s)$ and $\widetilde{c}_k(s)$ the columns of respectively $E_{11}(s)$ and $E(s)$, we may observe that

$$\begin{aligned}
\|E(s)\|^2 &= \sup\Big\{\big|E(s)\widetilde{u}\big|^2 \colon \widetilde{u} \in \mathbb{C}^{m+p},\ |\widetilde{u}| = 1\Big\} \\
&= \sup\Big\{\Big|\sum_{k=1}^{m+p} \widetilde{c}_k(s)\widetilde{u}_k\Big|^2 \colon \widetilde{u} \in \mathbb{C}^{m+p},\ |\widetilde{u}| = 1\Big\} \\
&\geq \sup\Big\{\Big|\sum_{k=1}^{m+p} \widetilde{c}_k(s)\widetilde{u}_k\Big|^2 \colon \widetilde{u} \in \mathbb{C}^{m+p},\ \widetilde{u}_i = 0\ \forall i > m,\ |\widetilde{u}| = 1\Big\} \qquad (33) \\
&= \Big\{\Big|\sum_{k=1}^{m} \widetilde{c}_k(s)u_k\Big|^2 \colon u \in \mathbb{C}^{m},\ |\widetilde{u}| = 1\Big\} \\
&\geq \Big\{\Big|\sum_{k=1}^{m} c_k(s)u_k\Big|^2 \colon u \in \mathbb{C}^{m},\ |\widetilde{u}| = 1\Big\} \qquad (34) \\
&= \Big\{\Big|\sum_{k=1}^{m} \widetilde{c}_k(s)u_k\Big|^2 \colon u \in \mathbb{C}^{m},\ |\widetilde{u}| = 1\Big\} \\
&\geq \Big\{\big|E_{11}(s)u\big|^2 \colon u \in \mathbb{C}^{m},\ |\widetilde{u}| = 1\Big\} \ =\ \|E_{11}(s)\|^2.
\end{aligned}$$

Here, the components of vectors $u \overset{\text{def}}{=} (u_1, \ldots, u_m)^T$ and $\widetilde{u} \overset{\text{def}}{=} (\widetilde{u}_1, \ldots, \widetilde{u}_{m+p})^T$, and both inequalities hold since, for (33) we took the supremum over a subset of $\{\widetilde{u} \in \mathbb{C}^{m+p} \colon |\widetilde{u}| = 1\}$, and for (34) we removed nonnegative terms.

Let $\widetilde{\mathcal{Q}}$ denote the observability Gramian of the system represented by $E(s)$. We now have to find $E(s)$ such that,

$$\widetilde{C}^*\widetilde{D} + \widetilde{\mathcal{Q}}\widetilde{B} = 0, \qquad (35)$$

and choose $\widetilde{D}$ satisfying $\big\|\widetilde{D}^*\widetilde{D}\big\|^{\frac{1}{2}} = 2\sigma_{\ell+1}$, so that Lemma 7.4 may yield

$$\big\|E\big\|_{H^\infty_{(p+m)\times(m+p)}} = \big\|\widetilde{D}^*\widetilde{D}\big\|^{\frac{1}{2}} = 2\sigma_{\ell+1}.$$

First, in the system associated with $E_{11}(s)$, we apply the following change of coordinate

$$T = \begin{pmatrix} \frac{1}{2}I_\ell & \frac{1}{2}I_\ell & \mathbb{O}_{\ell,n-\ell} \\ \frac{1}{2}I_\ell & -\frac{1}{2}I_\ell & \mathbb{O}_{\ell,n-\ell} \\ \mathbb{O}_{n-\ell,\ell} & \mathbb{O}_{n-\ell,\ell} & I_{n-\ell} \end{pmatrix}, \quad T^{-1} = \begin{pmatrix} I_\ell & I_\ell & \mathbb{O}_{\ell,n-\ell} \\ I_\ell & -I_\ell & \mathbb{O}_{\ell,n-\ell} \\ \mathbb{O}_{n-\ell,\ell} & \mathbb{O}_{n-\ell,\ell} & I_{n-\ell} \end{pmatrix}.$$

Then the transfer function E_{11} also has the form

$$E_{11}(s) = \left[\begin{array}{c|c} TA_\ell T^{-1} & TB_\ell \\ \hline C_\ell T^{-1} & D_\ell \end{array}\right] = \left[\begin{array}{ccc|c} A_{11} & \mathbb{O}_\ell & \frac{1}{2}A_{12} & B_1 \\ \mathbb{O}_\ell & A_{11} & -\frac{1}{2}A_{12} & \mathbb{O}_{\ell,m} \\ A_{21} & -A_{21} & A_{22} & B_2 \\ \hline \mathbb{O}_{p,\ell} & -2C_1 & C_2 & \mathbb{O}_{p,m} \end{array}\right].$$

The construction of $E(s)$ is based on $E_{11}(s)$. We keep the Hurwitz [5] matrix $TA_\ell T^{-1}$ by setting $\widetilde{A} \stackrel{\text{def}}{=} TA_\ell T^{-1}$ and enlarge the three other matrices defining them as

$$\widetilde{B} \stackrel{\text{def}}{=} \begin{pmatrix} TB_\ell & B^+ \end{pmatrix} \stackrel{\text{def}}{=} \begin{pmatrix} B_1 & B_1^+ \\ \mathbb{O}_{\ell,m} & B_2^+ \\ B_2 & B_3^+ \end{pmatrix} \in \mathbb{C}^{(n+\ell)\times(m+p)},$$

and

$$\widetilde{C} \stackrel{\text{def}}{=} \begin{pmatrix} C_\ell T^{-1} \\ C^+ \end{pmatrix} \stackrel{\text{def}}{=} \begin{pmatrix} \mathbb{O}_{p,\ell} & -2C_1 & C_2 \\ C_1^+ & C_2^+ & C_3^+ \end{pmatrix} \in \mathbb{C}^{(p+m)\times(n+\ell)}.$$

While $\widetilde{D}$ is fixed as

$$\widetilde{D} \stackrel{\text{def}}{=} \begin{pmatrix} D_\ell & 2\sigma_{\ell+1}I_p \\ 2\sigma_{\ell+1}I_m & \mathbb{O}_{m,p} \end{pmatrix} = 2\sigma_{\ell+1}\begin{pmatrix} \mathbb{O}_{p,m} & I_p \\ I_m & \mathbb{O}_{m,p} \end{pmatrix},$$

so that $\left\|\widetilde{D}^*\widetilde{D}\right\|^{\frac{1}{2}} = \overline{\sigma}(\widetilde{D}^*\widetilde{D})^{\frac{1}{2}} = 2\sigma_{\ell+1}$. The input and output spaces of the system associated with $E(s)$ are both of dimension $p+m$, and we may observe

5. Indeed, the matrix $A_\ell = \begin{pmatrix} A_{11} & \mathbb{O}_{\ell,n} \\ \mathbb{O}_{n,\ell} & A \end{pmatrix}$ is Hurwitz, since its eigenvalues are necessarily also eigenvalues of A_{11} and/or of A. Moreover, the eigenvalues of $\widetilde{A} = TA_\ell T^{-1}$ are also eigenvalues of A_ℓ, thus $\widetilde{A}$ is also a Hurwitz matrix.

that $E_{11}(s)$ is indeed a submatrix of $E(s)$ since

$$\begin{aligned}
E(s) &= \widetilde{C}\big(sI_{n+\ell} - \widetilde{A}\big)^{-1}\widetilde{B} + \widetilde{D} \\
&= \begin{pmatrix} C_\ell T^{-1} \\ C^+ \end{pmatrix} \big(sI_{n+\ell} - TA_\ell T^{-1}\big)^{-1} \begin{pmatrix} TB_\ell & B^+ \end{pmatrix} + \begin{pmatrix} D_\ell & 2\sigma_{\ell+1}I_p \\ 2\sigma_{\ell+1}I_m & \mathbb{O}_{m,p} \end{pmatrix} \\
&= \begin{pmatrix} C_\ell T^{-1}\big(sI_{n+\ell} - TA_\ell T^{-1}\big)^{-1}TB_\ell + D_\ell & C_\ell T^{-1}\big(sI_{n+\ell} - TA_\ell T^{-1}\big)^{-1}B^+ + 2\sigma_{\ell+1}I_p \\ C^+\big(sI_{n+\ell} - TA_\ell T^{-1}\big)^{-1}TB_\ell + 2\sigma_{\ell+1}I_m & C^+\big(sI_{n+\ell} - TA_\ell T^{-1}\big)^{-1}B^+ \end{pmatrix},
\end{aligned}$$

where we recognize the first block as $E_{11}(s)$. Thereafter, it remains to choose B^+ and C^+ such that (35) is fulfilled. Recall that the system (S) has controllability and observability Gramians given by

$$\mathcal{R} = \mathcal{Q} = \begin{pmatrix} \Sigma & \mathbb{O}_{\ell,n-\ell} \\ \mathbb{O}_{n-\ell} & \sigma_{\ell+1}I_{n-\ell} \end{pmatrix}, \quad \text{where } \Sigma \overset{\text{def}}{=} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_\ell \end{pmatrix}$$

and that satisfy the Lyapunov equations

$$\begin{aligned}
&\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}\begin{pmatrix} \Sigma & \mathbb{O}_{\ell,n-\ell} \\ \mathbb{O}_{n-\ell} & \sigma_{\ell+1}I_{n-\ell} \end{pmatrix} + \begin{pmatrix} \Sigma & \mathbb{O}_{\ell,n-\ell} \\ \mathbb{O}_{n-\ell} & \sigma_{\ell+1}I_{n-\ell} \end{pmatrix}\begin{pmatrix} A_{11}^* & A_{21}^* \\ A_{12}^* & A_{22}^* \end{pmatrix} \\
&\qquad\qquad = -\begin{pmatrix} B_1B_1^* & B_1B_2^* \\ B_2B_1^* & B_2B_2^* \end{pmatrix}
\end{aligned}$$

and

$$\begin{aligned}
&\begin{pmatrix} \Sigma & \mathbb{O}_{\ell,n-\ell} \\ \mathbb{O}_{n-\ell} & \sigma_{\ell+1}I_{n-\ell} \end{pmatrix}\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} + \begin{pmatrix} A_{11}^* & A_{21}^* \\ A_{12}^* & A_{22}^* \end{pmatrix}\begin{pmatrix} \Sigma & \mathbb{O}_{\ell,n-\ell} \\ \mathbb{O}_{n-\ell} & \sigma_{\ell+1}I_{n-\ell} \end{pmatrix} \\
&\qquad\qquad = -\begin{pmatrix} C_1^*C_1 & C_1^*C_2 \\ C_2^*C_1 & C_2^*C_2 \end{pmatrix}.
\end{aligned}$$

These equations may be rewritten as [6]

$$A_{11}\Sigma + \Sigma A_{11}^* = -B_1B_1^* \tag{36}$$

$$\Sigma A_{11} + A_{11}^*\Sigma = -C_1^*C_1 \tag{37}$$

$$\sigma_{\ell+1}A_{22} + \sigma_{\ell+1}A_{22}^* = -B_2B_2^* = -C_2^*C_2 \tag{38}$$

6. Two more equations are recovered by computing the adjoint on both sides of equations (39) and (40).

$$\sigma_{\ell+1}A_{12} + \Sigma A_{21}^* = -B_1 B_2^* \tag{39}$$

$$\Sigma A_{12} + \sigma_{\ell+1}A_{21}^* = -C_1^* C_2 \tag{40}$$

We begin by choosing C^+ and simultaneously finding the observability Gramian $\mathcal{Q}$ which is the solution of the Lyapunov equation

$$\widetilde{\mathcal{Q}}\widetilde{A} + \widetilde{A}^*\widetilde{\mathcal{Q}} = -\widetilde{C}^*\widetilde{C}, \tag{41}$$

since $\widetilde{A}$ is Hurwitz. We look for $\widetilde{\mathcal{Q}}$ in the form

$$\widetilde{\mathcal{Q}} = \begin{pmatrix} \Sigma_1 & \mathbb{O}_\ell & \mathbb{O}_{\ell,n-\ell} \\ \mathbb{O}_\ell & \Sigma_2 & \mathbb{O}_{\ell,n-\ell} \\ \mathbb{O}_{n-\ell,\ell} & \mathbb{O}_{n-\ell,\ell} & \Sigma_3 \end{pmatrix},$$

with diagonal matrices $\Sigma_1 \in \mathbb{C}^{\ell\times\ell}$, $\Sigma_2 \in \mathbb{C}^{\ell\times\ell}$ and $\Sigma_3 \in \mathbb{C}^{(n-\ell)\times(n-\ell)}$. Injecting this matrix in (41) yields

$$\begin{pmatrix} \Sigma_1 A_{11} & \mathbb{O}_\ell & \frac{1}{2}\Sigma_1 A_{12} \\ \mathbb{O}_\ell & \Sigma_2 A_{11} & -\frac{1}{2}\Sigma_2 A_{12} \\ \Sigma_3 A_{21} & -\Sigma_3 A_{21} & \Sigma_3 A_{22} \end{pmatrix} + \begin{pmatrix} A_{11}^*\Sigma_1 & \mathbb{O}_\ell & A_{21}^*\Sigma_3 \\ \mathbb{O}_\ell & A_{11}^*\Sigma_2 & -A_{21}^*\Sigma_3 \\ \frac{1}{2}A_{12}^*\Sigma_1 & -\frac{1}{2}\Sigma_2 A_{12}^* & A_{22}^*\Sigma_3 \end{pmatrix}$$

$$= -\begin{pmatrix} (C_1^+)^* C_1^+ & (C_1^+)^* C_2^+ & (C_1^+)^* C_3^+ \\ (C_2^+)^* C_1^+ & (-2C_1)^*(-2C_1) + (C_2^+)^* C_2^+ & (-2C_1)^* C_2 + (C_2^+)^* C_3^+ \\ (C_3^+)^* C_1^+ & C_2^*(-2C_1) + (C_3^+)^* C_2^+ & C_2^* C_2 + (C_3^+)^* C_3^+ \end{pmatrix}.$$

Moreover, if we choose $C_2^+ \stackrel{\text{def}}{=} \mathbb{O}_{m,\ell}$, the right-hand side of the above equation becomes

$$-\begin{pmatrix} (C_1^+)^* C_1^+ & \mathbb{O}_\ell & (C_1^+)^* C_3^+ \\ \mathbb{O}_\ell & (-2C_1)^*(-2C_1) & (-2C_1)^* C_2 \\ (C_3^+)^* C_1^+ & C_2^*(-2C_1) & C_2^* C_2 + (C_3^+)^* C_3^+ \end{pmatrix}.$$

Setting $\Sigma_3 \stackrel{\text{def}}{=} 2\sigma_{\ell+1} I_{n-\ell}$ and $C_3^+ \stackrel{\text{def}}{=} B_2^*$, we get from (38) that the equation

$$\Sigma_3 A_{22} + A_{22}^*\Sigma_3 = -(C_2^* C_2 + (C_3^+)^* C_3^+)$$

holds, and the Lyapunov equation becomes

$$\begin{pmatrix} \Sigma_1 A_{11} & \mathbb{O}_\ell & \frac{1}{2}\Sigma_1 A_{12} \\ \mathbb{O}_\ell & \Sigma_2 A_{11} & -\frac{1}{2}\Sigma_2 A_{12} \\ 2\sigma_{\ell+1} A_{21} & -2\sigma_{\ell+1} A_{21} & 2\sigma_{\ell+1} A_{22} \end{pmatrix} + \begin{pmatrix} A_{11}^*\Sigma_1 & \mathbb{O}_\ell & 2\sigma_{\ell+1} A_{21}^* \\ \mathbb{O}_\ell & A_{11}^*\Sigma_2 & -2\sigma_{\ell+1} A_{21}^* \\ \frac{1}{2}A_{12}^*\Sigma_1 & -\frac{1}{2}\Sigma_2 A_{12}^* & 2\sigma_{\ell+1} A_{22}^* \end{pmatrix}$$
$$= -\begin{pmatrix} (C_1^+)^* C_1^+ & \mathbb{O}_\ell & (C_1^+)^* B_2^* \\ \mathbb{O}_\ell & (-2C_1)^*(-2C_1) & (-2C_1)^* C_2 \\ B_2 C_1^+ & C_2^*(-2C_1) & C_2^* C_2 + B_2 B_2^* \end{pmatrix}.$$

Then, defining $\Sigma_2 \overset{\text{def}}{=} 4\Sigma$, we obtain from (37) and (40) that both equations

$$\begin{cases} \Sigma_2 A_{11} + A_{11}^*\Sigma_2 = -(-2C_1)^*(-2C_1) \\ -\dfrac{1}{2}\Sigma_2 A_{12} - 2\sigma_{\ell+1} A_{21}^* = -(-2C_1)^* C_2 \end{cases}$$

hold, as well as $-2\sigma_{\ell+1} A_{21} - \frac{1}{2}A_{12}^*\Sigma_2 = -C_2^*(-2C_1)$ since we may compute the adjoint on both sides of the last equation. The Lyapunov equation becomes

$$\begin{pmatrix} \Sigma_1 A_{11} & \mathbb{O}_\ell & \frac{1}{2}\Sigma_1 A_{12} \\ \mathbb{O}_\ell & 4\Sigma A_{11} & -2\Sigma A_{12} \\ 2\sigma_{\ell+1} A_{21} & -2\sigma_{\ell+1} A_{21} & 2\sigma_{\ell+1} A_{22} \end{pmatrix} + \begin{pmatrix} A_{11}^*\Sigma_1 & \mathbb{O}_\ell & 2\sigma_{\ell+1} A_{21}^* \\ \mathbb{O}_\ell & 4A_{11}^*\Sigma & -2\sigma_{\ell+1} A_{21}^* \\ \frac{1}{2}A_{12}^*\Sigma_1 & -2\Sigma A_{12}^* & 2\sigma_{\ell+1} A_{22}^* \end{pmatrix}$$
$$= -\begin{pmatrix} (C_1^+)^* C_1^+ & \mathbb{O}_\ell & (C_1^+)^* B_2^* \\ \mathbb{O}_\ell & (-2C_1)^*(-2C_1) & (-2C_1)^* C_2 \\ B_2 C_1^+ & C_2^*(-2C_1) & C_2^* C_2 + B_2 B_2^* \end{pmatrix}.$$

Finally, multiplying (39) on the left by $2\sigma_{\ell+1}\Sigma^{-1}$ and comparing it with the equation (coming from the matrix system above)

$$\frac{1}{2}\Sigma_1 A_{12} + 2\sigma_{\ell+1} A_{21}^* = -(C_1^+)^* B_2^*,$$

we deduce that setting $\Sigma_1 \overset{\text{def}}{=} 4\sigma_{\ell+1}^2\Sigma^{-1}$ and $C_1^+ \overset{\text{def}}{=} 2\sigma_{\ell+1} B_1^*\Sigma^{-1}$ implies that both the above equation (hence also $2\sigma_{\ell+1} A_{21} + \frac{1}{2}A_{12}^*\Sigma_1 = -B_2 C_1^+$ by computing the adjoints) and

$$\Sigma_1 A_{11} + A_{11}^*\Sigma_1 = -(C_1^+)^* C_1^+,$$

hold. To summarize, we have chosen

$$C^+ \overset{\text{def}}{=} \begin{pmatrix} 2\sigma_{\ell+1} B_1^*\Sigma^{-1} & \mathbb{O}_{m,\ell} & B_2^* \end{pmatrix}$$

and the observability Gramian of the system represented by $E(s)$ is

$$\widetilde{\mathcal{Q}} = \begin{pmatrix} 4\sigma_{\ell+1}\Sigma^{-1} & \mathbb{O}_{\ell} & \mathbb{O}_{\ell,n-\ell} \\ \mathbb{O}_{\ell} & 4\Sigma & \mathbb{O}_{\ell,n-\ell} \\ \mathbb{O}_{n-\ell,\ell} & \mathbb{O}_{n-\ell,\ell} & 2\sigma_{\ell+1}I_{n-\ell} \end{pmatrix}.$$

It remains to choose B^+ so that the equation (35) is also satisfied, i.e.

$$2\sigma_{\ell+1}\begin{pmatrix} 2\sigma_{\ell+1}\Sigma^{-1}B_1 & \mathbb{O}_{\ell,p} \\ \mathbb{O}_{\ell,m} & -2C_1^* \\ B_2 & C_2^* \end{pmatrix} + \begin{pmatrix} 4\sigma_{\ell+1}^2\Sigma^{-1}B_1 & 4\sigma_{\ell+1}^2\Sigma^{-1}B_1^+ \\ \mathbb{O}_{\ell,m} & 4\Sigma B_2^+ \\ 2\sigma_{\ell+1}B_2 & 2\sigma_{\ell+1}B_3^+ \end{pmatrix} = 0.$$

Thus, to conclude the proof, we may choose $B_1^+ \stackrel{\text{def}}{=} \mathbb{O}_{p,\ell}$, $B_2^+ \stackrel{\text{def}}{=} \sigma_{\ell+1}\Sigma^{-1}C_1^*$ and $B_3^+ \stackrel{\text{def}}{=} -C_2^*$. □

Theorem 7.6. — *The approximate system* $(\mathbf{S}_\ell)$ *defined above satisfies the following bound for the error*

$$\left\|G - G_\ell\right\|_{H_{p\times m}^\infty} \leq 2\sum_{k=1}^{m}\sigma_{i_k}.$$

Proof. — Proving this Theorem consists in applying m times the Proposition 7.5. At the beginning of this section, we have rewritten the Gramians of $(\mathbf{S})$ as

$$\mathcal{R} = \mathcal{Q} = \left(\begin{array}{c|ccc} \Sigma & & & 0 \\ \hline & \sigma_{i_m}I_{n_m} & & \\ & & \ddots & \\ & & & \sigma_{i_1}I_{n_1} \\ 0 & & & \end{array}\right)$$

Set $A^{i_0} \stackrel{\text{def}}{=} A$, $B^{i_0} \stackrel{\text{def}}{=} B$, $C^{i_0} \stackrel{\text{def}}{=} C$ and $G^{i_0}(s) \stackrel{\text{def}}{=} G(s)$. For any $k \in \{1,\ldots,m\}$, we define the system

$$(\mathbf{S}^{i_k})\colon \begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}z(t) = A_{11}^{i_k}z(t) + B_1^{i_k}u(t) \\ y(t) = C_1^{i_k}z(t) \end{cases}$$

to be the system $(\mathbf{S}^{i_{k-1}})$ from which the states corresponding to the singular value σ_{i_k} have been truncated. Set $N_k = \sum_{j=1}^{k} n_j$. This means that the

matrices $A_{11}^{i_k} \in \mathbb{C}^{(n-N_k)\times(n-N_k)}$, $B_1^{i_k} \in \mathbb{C}^{(n-N_k)\times m}$ and $C_1^{i_k} \in \mathbb{C}^{p\times(n-N_k)}$ satisfy

$$A_{11}^{i_{k-1}} = \begin{pmatrix} A_{11}^{i_k} & A_{12}^{i_k} \\ A_{21}^{i_k} & A_{22}^{i_k} \end{pmatrix}, \quad B_1^{i_{k-1}} = \begin{pmatrix} B_1^{i_k} \\ B_2^{i_k} \end{pmatrix}, \quad C_1^{i_{k-1}} = \begin{pmatrix} C_1^{i_k} & C_2^{i_k} \end{pmatrix}.$$

We know from Proposition 6.11 that the truncated matrices A^{i_k} are Hurwitz and that the truncated system $(\mathbf{S}^{i_k})$ is balanced with singular values

$$\sigma_1 \geq \ldots \geq \sigma_\ell > \underbrace{\sigma_{i_m} = \ldots = \sigma_{i_m}}_{n_m \text{ times}} > \ldots > \underbrace{\sigma_{i_{k-1}} = \ldots = \sigma_{i_{k-1}}}_{n_{k-1} \text{ times}}.$$

Proposition 7.5 states that $\left\|G^{i_{k-1}} - G^{i_k}\right\|_{H^\infty_{p\times m}} \leq 2\sigma_{i_k}$. For $k = m$, we obtain $G^{i_m} = G_\ell$ and conclude from the triangular inequality that

$$\left\|G - G_\ell\right\|_{H^\infty_{p\times m}} = \left\|G^{i_0} - G^{i_m}\right\|_{H^\infty_{p\times m}} \leq \sum_{k=1}^{m} \left\|G^{i_{k-1}} - G^{i_k}\right\|_{H^\infty_{p\times m}} \leq 2\sum_{k=1}^{m} \sigma_{i_k}. \ \square$$

8. Balanced POD

The balanced proper orthogonal decomposition method is an approximation of balanced truncation. This method relies on approximating the Gramians by using data from simulations, and then, computing the balancing transformation directly from these approximations (using POD/SVD). In this section, we mainly follow [**18**].

8.1. Empirical Gramians. — To begin, notice that the matrices B, C and their adjoints may be rewritten by using the columns of B and C^*. For $u \in \mathbb{C}^m$, we may rewrite Bu as

$$Bu = \begin{pmatrix} | & & | \\ b_1 & \ldots & b_m \\ | & & | \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix} = \sum_{k=1}^{m} b_k u_k$$

The adjoint of B is given, for any $x \in \mathbb{C}^n$, by

$$B^* x = \begin{pmatrix} \langle x\,, b_1\rangle_{\mathbb{C}^n} \\ \vdots \\ \langle x\,, b_m\rangle_{\mathbb{C}^n} \end{pmatrix} = \begin{pmatrix} — & b_1^* & — \\ & \vdots & \\ — & b_m^* & — \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

As for the matrix C, having this time c_k represent the columns of the adjoint of C, we may rewrite for $y \in \mathbb{C}^p$

$$C^* y = \begin{pmatrix} | & & | \\ c_1 & \dots & c_m \\ | & & | \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} = \sum_{k=1}^{p} c_k y_k$$

The matrix C is given, for any $x \in \mathbb{C}^n$, by

$$Cx = \begin{pmatrix} \langle x \, , c_1 \rangle_{\mathbb{C}^n} \\ \vdots \\ \langle x \, , c_p \rangle_{\mathbb{C}^n} \end{pmatrix} = \begin{pmatrix} — & c_1^* & — \\ & \vdots & \\ — & c_p^* & — \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

We make use of these observations to rewrite the Gramians. Let $w \in \mathbb{C}^n$ be some state element, then by definition of B^* and C, we have

$$B^* e^{A^* t} x = \begin{pmatrix} \langle x \, , e^{At} b_1 \rangle_{\mathbb{C}^n} \\ \vdots \\ \langle x \, , e^{At} b_m \rangle_{\mathbb{C}^n} \end{pmatrix}, \quad C e^{At} x = \begin{pmatrix} \langle x \, , e^{A^* t} c_1 \rangle_{\mathbb{C}^n} \\ \vdots \\ \langle x \, , e^{A^* t} c_p \rangle_{\mathbb{C}^n} \end{pmatrix}.$$

Injecting this in the expressions of the Gramians, as well as the definitions of B and C^*, we deduce that

$$\mathcal{R} x = \int_0^{+\infty} e^{At} B (B^* e^{A^* t} x) dt = \int_0^{+\infty} e^{At} \sum_{k=1}^{m} b_k \big\langle x \, , e^{At} b_k \big\rangle_{\mathbb{C}^n} dt$$

and

$$\mathcal{Q} x = \int_0^{+\infty} e^{A^* t} C^* (C e^{At} x) dt = \int_0^{+\infty} e^{A^* t} \sum_{k=1}^{p} c_k \big\langle x \, , e^{A^* t} c_k \big\rangle_{\mathbb{C}^n} dt$$

For $k \in \{1, \dots, m\}$, the function $q_k(t) \overset{\text{def}}{=} e^{At} b_k$ is the solution of

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} z(t) = A z(t), \quad \forall t \in \mathbb{R}_+ \\ z(0) = b_k. \end{cases}$$

For $k \in \{1, \dots, p\}$, the function $w_k(t) \overset{\text{def}}{=} e^{A^* t} c_k$ is the solution of

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} z(t) = A^* z(t) \quad \forall t \in \mathbb{R}_+ \\ z(0) = c_k. \end{cases}$$

As a consequence, we may rewrite the Gramians as

$$\mathcal{R}x = \int_0^{+\infty} \sum_{k=1}^{m} q_k(t)\big\langle x\,, q_k(t)\big\rangle_{\mathbb{C}^n} dt, \quad \mathcal{Q}x = \int_0^{+\infty} \sum_{k=1}^{p} w_k(t)\big\langle x\,, w_k(t)\big\rangle_{\mathbb{C}^n} dt.$$

From now onward, instead of using exact Gramians $\mathcal{R}, \mathcal{Q}$, we will consider the operators $\mathcal{R}_N$ and $\mathcal{Q}_N$, respectively called the *approximate ,or empirical, controllability and observability Gramians*,

$$\mathcal{R}x \approx \mathcal{R}_N x = \sum_{j=1}^{N} \alpha_j \sum_{k=1}^{m} q_k(t_j)\big\langle x\,, q_k(t_j)\big\rangle_{\mathbb{C}^n},$$

and

$$\mathcal{Q}x \approx \mathcal{Q}_N x = \sum_{j=1}^{N} \alpha_j \sum_{k=1}^{p} w_k(t_j)\big\langle x\,, w_k(t_j)\big\rangle_{\mathbb{C}^n}.$$

These operators are constructed by applying a quadrature rule to the integrals, and $\{\alpha_j\}_{j=1}^N$ are positive weights. Note that the weights α_j depend on N, even if this does not appear in the notation. We may observe that the empirical Gramians can be factored as the product of two matrices. In effect, for $\mathcal{R}_N \in \mathbb{C}^{n\times n}$ we have

$$\mathcal{R}_N x = \sum_{j=1}^{N} \sum_{k=1}^{m} \sqrt{\alpha_j} q_k(t_j)\big\langle x\,, \sqrt{\alpha_j} q_k(t_j)\big\rangle_{\mathbb{C}^n} = XX^* x,$$

where the matrix $X \in \mathbb{C}^{n\times(mN)}$ is defined by

$$X \stackrel{\text{def}}{=} \begin{pmatrix} | & & | & & | & & | \\ \sqrt{\alpha_1} q_1(t_1) & \dots & \sqrt{\alpha_N} q_1(t_N) & \dots\dots & \sqrt{\alpha_1} q_m(t_1) & \dots & \sqrt{\alpha_N} q_m(t_N) \\ | & & | & & | & & | \end{pmatrix}.$$

The adjoint $X^* \in \mathbb{C}^{(mN)\times n}$ of X is given by

$$X^*x = \begin{pmatrix} \langle x\,,\sqrt{\alpha_1}q_1(t_1)\rangle_{\mathbb{C}^n} \\ \vdots \\ \langle x\,,\sqrt{\alpha_N}q_1(t_N)\rangle_{\mathbb{C}^n} \\ \vdots \\ \langle x\,,\sqrt{\alpha_1}q_m(t_1)\rangle_{\mathbb{C}^n} \\ \vdots \\ \langle x\,,\sqrt{\alpha_N}q_m(t_N)\rangle_{\mathbb{C}^n} \end{pmatrix},$$

for any $x \in \mathbb{C}^n$. As for the empirical observability Gramian $\mathcal{Q}_N \in \mathbb{C}^{n\times n}$, it rewrites as

$$\mathcal{Q}_N x = \sum_{j=1}^{N}\sum_{k=1}^{p} \sqrt{\alpha_j}w_k(t_j)\langle x\,,\sqrt{\alpha_j}w_k(t_j)\rangle_{\mathbb{C}^n} = YY^*x,$$

with $Y \in \mathbb{C}^{n\times(pN)}$ defined by

$$Y \overset{\text{def}}{=} \begin{pmatrix} | & & | & & | & & | \\ \sqrt{\alpha_1}w_1(t_1) & \dots & \sqrt{\alpha_N}w_1(t_N) & \dots\dots & \sqrt{\alpha_1}w_p(t_1) & \dots & \sqrt{\alpha_N}w_p(t_N) \\ | & & | & & | & & | \end{pmatrix}.$$

The empirical Gramians may not be of rank n.

8.2. Balancing using POD. — In order to know which state to truncate, we need to balance the system. Thus, we have to find a change of coordinate $T \in \mathbb{C}^{n\times n}$ (whose inverse will be denoted by S) in the state space, such that in the new system the Gramians $T\mathcal{R}T^*$ and $S^*\mathcal{Q}S$ are equal and diagonal. We also have seen that, assuming controllability and observability, such a matrix T exists (see Proposition 6.4).

From now on, instead of assuming that we have access to the exact Gramians, we want to use the empirical Gramians $\mathcal{R}_N, \mathcal{Q}_N$ and find a change of coordinate T (of inverse S) such that $T\mathcal{R}_N T^*$ and $S^*\mathcal{Q}_N S$ are equal and diagonal. Analogously to the case of the exact Gramians, we give the following definition.

Definition 8.1 **(Empirical Hankel singular values)**
The square roots of the eigenvalues of $\mathcal{R}_N\mathcal{Q}_N$ (the product of the empirical controllability and observability Gramians) are called the *empirical Hankel singular values*.

Since the empirical Gramians depend on N, in what follows, many quantities also depends on N, even if this does not appear in the notation. Consider the matrix $Y^*X \in \mathbb{C}^{(pN)\times(mN)}$. Let its singular value decomposition be given by

$$Y^*X = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} \Sigma_1 & \mathbb{O}_{r,(mN)} \\ \mathbb{O}_{(pN),r} & \mathbb{O}_{(pN),(mN)} \end{pmatrix} \begin{pmatrix} V_1^* \\ V_2^* \end{pmatrix} = U_1\Sigma_1 V_1^*, \tag{42}$$

where

$$\Sigma_1 \overset{\text{def}}{=} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{pmatrix} \qquad \text{and} \qquad r \overset{\text{def}}{=} \operatorname{rank}(Y^*X).$$

Recall that the diagonal components of Σ_1 are positive and ordered decreasingly. The size of Y^*X does not depend on the size of the state space $\mathbb{C}^n$. Compared to the SVD needed in Proposition 6.4, this is a computational advantage if the dimension of the input and output space are small compared to n, and mN, pN are consequently small compared to n (for N not too large).

Proposition 8.2. — *The rectangular matrices T_1, S_1 defined by*

$$T_1 \overset{def}{=} \Sigma_1^{-\frac{1}{2}} U_1^* Y^* \in \mathbb{C}^{r\times n}, \quad S_1 \overset{def}{=} XV_1\Sigma_1^{-\frac{1}{2}} \in \mathbb{C}^{n\times r}$$

satisfy $T_1S_1 = I_r$. Moreover, the empirical Gramians satisfy

$$T_1\mathcal{R}_N T_1^* = S_1^*\mathcal{Q}_N S_1 = \Sigma_1.$$

Remark 8.3 **(If Y^*X has rank n).** — If mN and pN were equal to or larger than n, then it would be possible to have $\operatorname{rank}(Y^*X) = n$. In practice, we rather choose mN and pN to be much smaller than the dimension of the state space. However, we may observe that if we assume that Y^*X has rank n, then Proposition 8.2 affirms that T_1 is square and invertible, with inverse S_1. Moreover,

$$T_1\mathcal{R}_N T_1^* = S_1^*\mathcal{Q}_N S_1 = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{pmatrix}$$

and the empirical Hankel singular values $\{\sigma\}_{k=1}^{n}$ are positive and ordered decreasingly. In this case, T_1 is a balancing transformation (with respect to the empirical Gramians).

Proof of Proposition 8.2. — Similarly to Proposition 6.4, we will find expressions for T_1 and S_1 by considering the singular value decomposition of a matrix. Indeed, multiplying (42) on the left by $(Y^*X)^*$ yields $X^*\mathcal{Q}_N X = X^*YY^*X = V_1\Sigma_1^2V_1^*$, consequently

$$\Sigma_1^{-\frac{1}{2}}V_1^*X^*\mathcal{Q}_N X V_1\Sigma_1^{-\frac{1}{2}} = \Sigma_1 \;\Leftrightarrow\; S_1^*\mathcal{Q}_N S_1 = \Sigma_1,$$

where S_1 is defined by $S_1 \overset{\text{def}}{=} XV_1\Sigma_1^{-\frac{1}{2}}$. The singular value decomposition of Y^*X also allows to find a matrix T_1 such that $T_1S_1 = I_r$. Indeed, multiplying (42) on the left by $\Sigma_1^{-\frac{1}{2}}U_1^*$ and on the right by $V_1\Sigma_1^{-\frac{1}{2}}$ yields

$$(\Sigma_1^{-\frac{1}{2}}U_1^*Y^*)(XV_1\Sigma_1^{-\frac{1}{2}}) = I_r,$$

and we define $T_1 \overset{\text{def}}{=} \Sigma_1^{-\frac{1}{2}}U_1^*Y^*$. It remains to check that

$$\begin{aligned} T_1\mathcal{R}_N T_1^* &= \Sigma_1^{-\frac{1}{2}}U_1^*(Y^*X)(X^*Y)U_1\Sigma_1^{-\frac{1}{2}} \\ &= \Sigma_1^{-\frac{1}{2}}U_1^*(U_1\Sigma_1V_1^*)(V_1\Sigma_1U_1^*)U_1\Sigma_1^{-\frac{1}{2}} = \Sigma_1. \end{aligned}$$ □

The following Proposition states that we may transform (**S**) ny using the change of coordinates T defined bellow, into an "almost balanced" system (with respect to the empirical Gramians). The first r columns of T are T_1 and the first r rows of its inverse are S_1. The new system has empirical Gramians of the form (43), whose product is given by (44), hence Σ_1 contains all the positive empirical Hankel singular values.

Proposition 8.4. — *Assume that Y^*X has rank $r < n$. Let, T_1, S_1 be the matrices defined in the Proposition 8.2. Then there exists $T_2 \in \mathbb{C}^{(n-r)\times n}$ and $S_2 \in \mathbb{C}^{n\times(n-r)}$ such that for*

$$T \overset{def}{=} \begin{pmatrix} T_1 \\ T_2 \end{pmatrix}, \quad S \overset{def}{=} \begin{pmatrix} S_1 & S_2 \end{pmatrix},$$

T is invertible with inverse S, and

$$T\mathcal{R}_N T^* = \begin{pmatrix} \Sigma_1 & \mathbb{O}_{r,(n-r)} \\ \mathbb{O}_{(n-r),r} & L \end{pmatrix}, \quad S^*\mathcal{Q}_N S = \begin{pmatrix} \Sigma_1 & \mathbb{O}_{r,(n-r)} \\ \mathbb{O}_{(n-r),r} & M \end{pmatrix}, \tag{43}$$

and furthermore

$$T\mathcal{R}_N\mathcal{Q}_N S = \begin{pmatrix} \Sigma_1^2 & \mathbb{O}_{r,(n-r)} \\ \mathbb{O}_{(n-r),r} & \mathbb{O}_{n-r} \end{pmatrix}. \tag{44}$$

Proof. — Our aim is to find T_2, S_2 satisfying $TS = I_n$, i.e.

$$\begin{pmatrix} T_1S_1 & T_1S_2 \\ T_2S_1 & T_2S_2 \end{pmatrix} = \begin{pmatrix} I_r & \mathbb{O}_{r,(n-r)} \\ \mathbb{O}_{(n-r),r} & I_{n-r} \end{pmatrix}. \tag{45}$$

The Proposition 8.2 already gives $T_1S_1 = I_r$, which also implies that $\text{rank}(T_1) = r$. Consequently, the dimension of $\ker(T_1)$ is $n - r \neq 0$. Let $\{s_{r+1}, s_{r+2}, \ldots, s_n\}$ be a basis for $\ker(T_1)$ and define the matrix $S_2 \in \mathbb{C}^{n\times(n-r)}$ whose columns are the vectors s_k for $r + 1 \leq k \leq n$, i.e.

$$S_2 \overset{\text{def}}{=} \begin{pmatrix} | & & | \\ s_{r+1} & \ldots & s_n \\ | & & | \end{pmatrix}.$$

Any element of $\text{Ran}(S_2)$ will belong to $\ker(T_1)$, thus necessarily $T_1S_2 = \mathbb{O}_{r,(n-r)}$. Moreover, the columns of $S = \begin{pmatrix} S_1 & S_2 \end{pmatrix}$ must be linearly independent. Indeed, let $\alpha = (\alpha_1, \ldots, \alpha_n) \neq 0_{\mathbb{C}^n}$, denote by $s_1, \ldots, s_r$ the columns of S_1. If $\sum_{k=1}^n \alpha_k s_k$ were equal to zero, using that S_1 maps this linear combination to $\begin{pmatrix} \alpha_1, \ldots, \alpha_r, 0, \ldots, 0 \end{pmatrix}^T$ we would deduce that the first r coefficients $\alpha_1, \ldots, \alpha_r$ are equal to zero and, as a consequence, $\sum_{k=r+1}^n \alpha_k s_k = 0_{\mathbb{C}^n}$. The columns of S_2 being independent, the remaining coefficients would also be necessarily equal to zero. Hence, S is invertible and we may build T_2 from the last $n - r$ lines of the inverse of S, meaning that rewriting the inverse of S as

$$S^{-1} \overset{\text{def}}{=} \begin{pmatrix} P_1 \\ P_2 \end{pmatrix},$$

with $P_1 \in \mathbb{C}^{r,n}$ and $P_2 \in \mathbb{C}^{(n-r),n}$, we set $T_2 \overset{\text{def}}{=} P_2$. Since

$$S^{-1}S = \begin{pmatrix} P_1S_1 & P_1S_2 \\ P_2S_1 & P_2S_2 \end{pmatrix} = \begin{pmatrix} I_r & \mathbb{O}_{r,n-r} \\ \mathbb{O}_{n-r,r} & I_{n-r} \end{pmatrix},$$

the matrix T_2 automatically satisfies $T_2S_1 = \mathbb{O}_{(n-r),r}$ and $T_2S_2 = I_{n-r}$. Thus (45) holds.

The remaining relations (43), (44) follow from the definitions of T and S. After the change of coordinates, the empirical Gramians have the form

$$T\mathcal{R}_N T^* = \begin{pmatrix} T_1\mathcal{R}_N T_1^* & T_1\mathcal{R}_N T_2^* \\ T_2\mathcal{R}_N T_1^* & T_2\mathcal{R}_N T_2^* \end{pmatrix}, \quad S^*\mathcal{Q}_N S = \begin{pmatrix} S_1^*\mathcal{Q}_N S_1 & S_1^*\mathcal{Q}_N S_2 \\ S_2^*\mathcal{Q}_N S_1 & S_2^*\mathcal{Q}_N S_2 \end{pmatrix}.$$

From Proposition 8.2, we already know that $T_1\mathcal{R}_N T_1^* = S_1^*\mathcal{Q}_N S_1 = \Sigma_1$. We consider the following equations, which originate from (45),

$$\begin{cases} S_1^* T_2^* = \mathbb{O}_{(n-r),r} \\ T_1 S_2 = \mathbb{O}_{r,(n-r)} \end{cases},$$

and rewrite them as

$$\begin{cases} \big(\Sigma_1^{-\frac{1}{2}} V_1^* X^*\big) T_2^* = \mathbb{O}_{r,(n-r)} \\ \big(\Sigma_1^{-\frac{1}{2}} U_1^* Y^*\big) S_2 = \mathbb{O}_{r,(n-r)} \end{cases} \Leftrightarrow \begin{cases} \big(\Sigma_1^{\frac{1}{2}} V_1^*\big) X^* T_2^* = \mathbb{O}_{r,(n-r)} \\ \big(\Sigma_1^{\frac{1}{2}} U_1^*\big) Y^* S_2 = \mathbb{O}_{r,(n-r)} \end{cases}$$

The singular value decomposition of Y^*X (42) yields

$$\begin{cases} \Sigma_1^{\frac{1}{2}} V_1^* = \Sigma_1^{-\frac{1}{2}} U_1^* Y^* X \\ \Sigma_1^{\frac{1}{2}} U_1^* = \Sigma_1^{-\frac{1}{2}} V_1^* X^* Y \end{cases} \Leftrightarrow \begin{cases} \Sigma_1^{\frac{1}{2}} V_1^* = T_1 X \\ \Sigma_1^{\frac{1}{2}} U_1^* = S_1^* Y \end{cases}.$$

Combining the four obtained equations yields

$$T_1\mathcal{R}_N T_2^* = \mathbb{O}_{r,(n-r)} = T_2\mathcal{R}_N T_1^* \quad \text{and} \quad S_1^*\mathcal{Q}_N S_2 = \mathbb{O}_{r,(n-r)} = S_2^*\mathcal{Q}_N S_1.$$

We define $L \overset{\text{def}}{=} T_2\mathcal{R}_N T_2^*$ and $M \overset{\text{def}}{=} S_2^*\mathcal{Q}_N S_2$, so that (43) holds. In order to verify (44), it remains to check that $LM = \mathbb{O}_{n-r}$. This follows again from equations originating from (45): $T_2 S_2 = I_{n-r} = S_2^* T_2^*$ and $T_1 S_2 = \mathbb{O}_{n-r,r}$ (or $T_2 S_1 = \mathbb{O}_{r,n-r}$). Indeed, LM rewrites as

$$\begin{aligned} LM &= T_2\mathcal{R}_N\big(T_2^* S_2^*\big)\mathcal{Q}_N S_2 \\ &= T_2\mathcal{R}_N\mathcal{Q}_N S_2 \\ &= T_2 X\big(X^*Y\big)Y^* S_2 \\ &= T_2 X\big(U_1\Sigma_1 V_1^*\big)^* Y^* S_2 \\ &= T_2\big(X V_1 \Sigma_1^{-\frac{1}{2}}\big)\Sigma_1^2\big(\Sigma_1^{-\frac{1}{2}} U_1^* Y^*\big) S_2 \\ &= T_2 S_1 \Sigma_1^2 T_1 S_2 \;=\; \mathbb{O}_{n-r}. \end{aligned}$$

□

Suppose we apply the change of coordinates $z = S\widetilde{z}$ to the system $(\mathbf{S})$. We have seen that the obtained system is

$$(\widetilde{\mathbf{S}}):\ \begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}\widetilde{z}(t) = \widetilde{A}z(t) + \widetilde{B}u(t) \\ y(t) = \widetilde{C}z(t) \end{cases}$$

where

$$\begin{cases} \widetilde{A} = TAS = \begin{pmatrix} T_1AS_1 & T_1AS_2 \\ T_2AS_1 & T_2AS_2 \end{pmatrix} \\ \widetilde{B} = TB = \begin{pmatrix} T_1B \\ T_2B \end{pmatrix} \\ \widetilde{C} = CS = \begin{pmatrix} CS_1 & CS_2 \end{pmatrix}, \end{cases} \tag{46}$$

and

$$\widetilde{z}(0) = Tz_0 = \begin{pmatrix} T_1z_0 \\ T_2z_0 \end{pmatrix}.$$

Choose $\ell < r = \operatorname{rank}(Y^*X)$, then the approximate system may be found by truncating the system

$$(\mathbf{S}_r):\ \begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}z_r(t) = \big(T_1AS_1\big)z_r(t) + \big(T_1B\big)u(t), \quad z_r(0) = T_1z_0 \\ y(t) = \big(CS_1\big)z_r(t). \end{cases}$$

Hence, the approximation is possible just by computing T_1 and S_1.

8.3. Application to the true Gramians. — The results given by Propositions 8.2 and 8.4 can be used to find a balancing transformation for a minimal system, or even for a system which may not minimal. If instead of defining X and Y as before, we had diagonalized the exacts Gramians $\mathcal{R}, \mathcal{Q}$ the following way

$$\mathcal{R} = P_RD_RP_R^*, \quad \mathcal{Q} = P_QD_QP_Q^*,$$

and defined $X, Y \in \mathbb{C}^{n\times n}$ by $X \overset{\text{def}}{=} P_RD_R^{\frac{1}{2}}P_R^*$ and $Y \overset{\text{def}}{=} P_QD_Q^{\frac{1}{2}}P_Q^*$. Then, $\mathcal{R} = XX^*$ and $\mathcal{Q} = YY^*$.

If $(\mathbf{S})$ is minimal, the Proposition 8.2 affirms that T_1 (defined above) is a balancing transformation. Moreover, in the general case where the system may not be minimal, one may show, by using the Proposition 8.4 and a modification of Proposition 6.11, that $(\mathbf{S}_r)$ defined above is balanced, minimal, and that its Gramians are given by Σ_1. Furthermore, [**20**, Theorems 2.1 and 2.2,

p.1321-1323] affirms that (**S**) and ($\mathbf{S}_r$) have the same input-output behavior, by showing that all Hankel singular values of the error system corresponding to $G - G_r$ are equal to zero (where G_r denotes the transfer function of ($\mathbf{S}_r$)).

9. Algorithms

Algorithm 1 is a strict derivation of the results from Section 6. Consequently it is, in theory, applicable to minimal systems. However, it may be preferable

Algorithm 1 Balanced truncation: first method

1. **Compute** the Gramians $\mathcal{R}$ and $\mathcal{Q}$
2. **Diagonalization** $\mathcal{R}^{\frac{1}{2}}\mathcal{Q}\mathcal{R}^{\frac{1}{2}} = U\Sigma^2U^*$
3. **Set** $T = \Sigma^{\frac{1}{2}}U^*\mathcal{R}^{-\frac{1}{2}}$ and $S = \mathcal{R}^{\frac{1}{2}}U\Sigma^{-\frac{1}{2}}$
4. **Choose** $\ell \leq n$
5. **Set** T_f = first ℓ rows of T. Set S_f = first ℓ columns of S
6. **Set** $A_\ell = T_\mathrm{f}AS_\mathrm{f}$, $B_\ell = T_\mathrm{f}B$ and $C_\ell = CS_\mathrm{f}$

to use the Algorithm 2 even for a system which is both controllable and observable, as inverting the computed Gramians may lead to numerical errors. In particular, Algorithm 1 asks to invert Σ, while the diagonal entries of this matrix can be very small. This second method may be used both for minimal and non-minimal systems, with a supplementary constrain for non-minimal systems: ℓ must be chosen less than $\mathrm{rank}(Y^*X)$. Algorithm 3 differs from Al-

Algorithm 2 Balanced truncation: second method

1. **Compute** the Gramians $\mathcal{R}$ and $\mathcal{Q}$
2. **Diagonalization**: $\mathcal{R} = P_R D_R P_R^*$ and $\mathcal{Q} = P_Q D_Q P_Q^*$
3. **Set** $X = P_R D_R^{\frac{1}{2}} P_R^*$ and $Y = P_Q D_Q^{\frac{1}{2}} P_Q^*$
4. **SVD** : $Y^*X = U\Sigma V^*$
5. **Set** U_1 = first r columns of U. **Set** V_1 = first r columns of V. **Set** Σ_1 = the $r \times r$ left upper block of Σ
6. **Set** $r = \mathrm{rank}(Y^*X)$
7. **Choose** $\ell \leq r$
8. **Set** $T_1 = \Sigma_1^{-\frac{1}{2}}U_1^*Y^*$ and $S_1 = XV_1\Sigma_1^{-\frac{1}{2}}$
9. **Set** T_f = first ℓ rows of T_1. **Set** S_f = first ℓ columns of S_1
10. **Set** $A_\ell = T_\mathrm{f}AS_\mathrm{f}$, $B_\ell = T_\mathrm{f}B$ and $C_\ell = CS_\mathrm{f}$

gorithm 2 by its 3 first steps (the construction of the empirical Gramians), and

because the fourth step involves the singular value decomposition of a matrix whose size is independent of n.

Algorithm 3 Balanced proper orthogonal decomposition

1. **For** $k \in \{1, \dots, m\}$, **solve**

$$\frac{\mathrm{d}}{\mathrm{d}t} z(t) = Az(t), \quad z(0) = b_k.$$

for the time discretization $[t_1, \dots, t_N]$ prescribed by the quadrature rule, the solution is $q_k \in \mathbb{C}^N$

2. **For** $k \in \{1, \dots, p\}$, **solve**

$$\frac{\mathrm{d}}{\mathrm{d}t} z(t) = A^* z(t), \quad z(0) = c_k$$

for the time discretization $[t_1, \dots, t_N]$ prescribed by the quadrature rule, the solution is $w_k \in \mathbb{C}^N$

3. **Set**

$$X = \begin{pmatrix} | & & | & & | & & | \\ \sqrt{\alpha_1} q_1(t_1) & \dots & \sqrt{\alpha_N} q_1(t_N) & \dots\dots & \sqrt{\alpha_1} q_m(t_1) & \dots & \sqrt{\alpha_N} q_m(t_N) \\ | & & | & & | & & | \end{pmatrix}$$

and

$$Y = \begin{pmatrix} | & & | & & | & & | \\ \sqrt{\alpha_1} w_1(t_1) & \dots & \sqrt{\alpha_N} w_1(t_N) & \dots\dots & \sqrt{\alpha_1} w_p(t_1) & \dots & \sqrt{\alpha_N} w_p(t_N) \\ | & & | & & | & & | \end{pmatrix}$$

4. **SVD** : $Y^*X = U\Sigma V^*$
5. **Set** $U_1 =$ first r columns of U. **Set** $V_1 =$ first r columns of V. **Set** $\Sigma_1 =$ the $r \times r$ left upper block of Σ
6. **Set** $r = \mathrm{rank}(Y^*X)$
7. **Choose** $\ell \leq r$
8. **Set** $T_1 = \Sigma_1^{-\frac{1}{2}} U_1^* Y^*$ and $S_1 = X V_1 \Sigma_1^{-\frac{1}{2}}$
9. **Set** $T_{\mathrm{f}} =$ first ℓ rows of T_1. **Set** $S_{\mathrm{f}} =$ first ℓ columns of S_1
10. **Set** $A_\ell = T_{\mathrm{f}} A S_{\mathrm{f}}$, $B_\ell = T_{\mathrm{f}} B$ and $C_\ell = C S_{\mathrm{f}}$

10. Numerical experiments

We conduct a numerical experiment using `Matlab` and the code is available at `https://github.com/chrdz/Balanced_truncation_2018`. We consider the

2-d heat equation with boundary control and Neumann observation, where $\Omega \stackrel{\text{def}}{=} (0,\pi)\times(0,\pi)$ and the space variable is denoted by $x=(x_1,x_2)$. The control and observation are done on $\Gamma \stackrel{\text{def}}{=} \{(x_1,0)\colon 0<x_1<\pi\}$.

$$(\mathbf{P})\begin{cases} \partial_t z = \Delta z & \text{in } (0,T)\times\Omega \\ z=0 & \text{on } (0,T)\times[\partial\Omega\setminus\Gamma] \\ z=u & \text{on } (0,T)\times\Gamma \\ z(0,\cdot,\cdot)=z^0 & \text{on } \Omega \\ y=\partial_n z & \text{on } (0,T)\times\Gamma \end{cases}$$

Here, by ∂_n we denote the normal derivative, which in this case corresponds to $-\partial_{x_2} z$ over Γ. We realize a finite difference space semi-discretization in order to obtain a finite dimensional input-output system and then use the balanced truncation method to reduce the order of this system. Let $N\in\mathbb{N}$ and let the uniform space step be given by $h=\frac{\pi}{N+1}$. The mesh points are denoted by $x_{i,j} \stackrel{\text{def}}{=} (ih,jh)$ for $i,j\in\{0,\dots,N+1\}$.

The unknown at each time t are $z_{i,j}(t) \stackrel{\text{def}}{=} z(t,ih,jh)$, for $i,j\in\{1,\dots,N\}$. We set $z^0_{i,j} \stackrel{\text{def}}{=} z^0(ih,jh)$, for $i,j\in\{0,\dots,N+1\}$. The space semi-discretization is done by using second order centered finite difference for the second derivative and first order forward finite difference for the normal derivative.

$$(\mathbf{P}_h)\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} z_{i,j} = \dfrac{z_{i-1,j}+z_{i,j-1}-4z_{i,j}+z_{i,j+1}+z_{i+1,j}}{h^2} & 0<t<T,\ 1\le i,j\le N \\ z_{i,j}=0 & 0<t<T,\ (i,j)\in\mathcal{I} \\ z_{0,j}=u_j & 0<t<T,\ 0\le j\le N+1 \\ z_{i,j}(0)=z^0_{i,j} & 0\le i,j\le N+1 \\ y_j=-\dfrac{z_{1,j}-z_{0,j}}{h} & 0<t<T,\ 1\le j\le N. \end{cases}$$

where $\mathcal{I}$ corresponds to the indexes of the mesh points belonging to $\partial\Omega\setminus\Gamma$:

$$\begin{aligned} \mathcal{I} \stackrel{\text{def}}{=} & \big\{(i,0)\colon 0\le i\le N+1\big\} \\ & \cup\big\{(N+1,j)\colon 0\le j\le N+1\big\} \\ & \cup\big\{(i,N+1)\colon 0\le i\le N+1\big\}. \end{aligned}$$

The unknown of the partial differential equation are located at the interior points of the mesh, meaning $\{x_{i,j}\}_{i,j=1}^N$. For these interior points, we introduce another indexing: we define $x_k \stackrel{\text{def}}{=} x_{i,j}$ for $k=(i-1)N+j$, for $i,j\in\{1,\dots,N\}$.

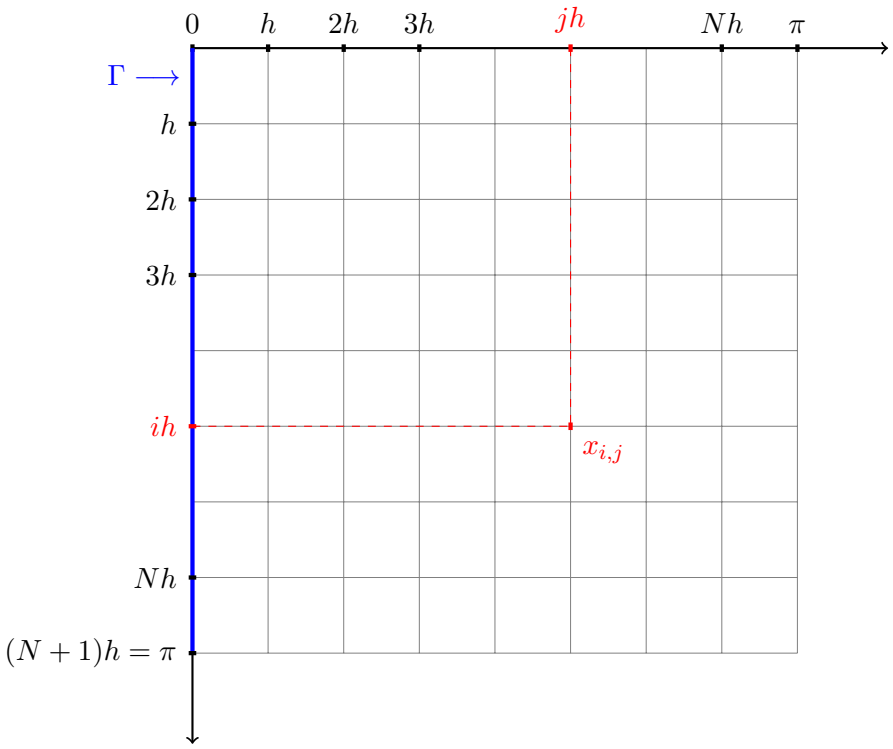

FIGURE 1. Mesh with i, j-indexing

This notation will be necessary to rewrite the semi-discretized PDE as a matrix system.

Similarly to the previous notation $z_{i,j}$, we set $z_k(t) \overset{\text{def}}{=} z_{i,j}(t)$ and $z_k^0 \overset{\text{def}}{=} z_{i,j}^0$. Hence, defining the state space function $Z\colon t \mapsto \mathbb{C}^{N^2}$ and the initial state $Z^0 \in \mathbb{C}^{N^2}$ by

$$Z \overset{\text{def}}{=} \begin{pmatrix} z_{1,1} \\ | \\ z_{1,N} \\ \vdots \\ z_{N,1} \\ | \\ z_{N,N} \end{pmatrix}, \qquad Z^0 \overset{\text{def}}{=} \begin{pmatrix} z_{1,1}^0 \\ | \\ z_{1,N}^0 \\ \vdots \\ z_{N,1}^0 \\ | \\ z_{N,N}^0 \end{pmatrix},$$

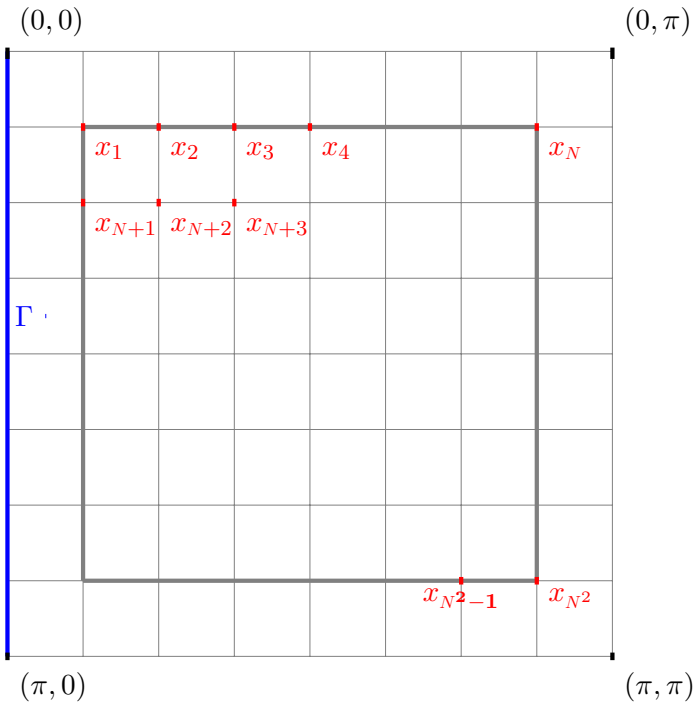

FIGURE 2. Mesh with k-indexing

as well as the input function $U\colon t \mapsto \mathbb{C}^N$ and output function $Y\colon t \mapsto \mathbb{C}^N$ by $U \stackrel{\text{def}}{=} (u_1, \ldots, u_N)^T$ and $Y \stackrel{\text{def}}{=} (y_1, \ldots, y_N)^T$, the semi-discretized problem $(\mathbf{P}_h)$ rewrites as

$$(\mathbf{P}_h) \begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} Z = AZ + BU & 0 < t < T \\ Z(0) = Z^0 & \\ Y = CZ + DU & 0 < t < T. \end{cases}$$

The input and output spaces have the same dimension N, while the dimension of the state space is N^2. The matrix $A \in \mathbb{C}^{N^2 \times N^2}$ is the the pentadiagonal matrix that is defined bellow by using the blocks $M \in \mathbb{C}^{N \times N}$ and I_N:

$$A \stackrel{\text{def}}{=} \frac{1}{h^2} \begin{pmatrix} M & I_N & & 0 \\ I_N & \ddots & \ddots & \\ & \ddots & \ddots & I_N \\ 0 & & I_N & M \end{pmatrix}, \quad M \stackrel{\text{def}}{=} \begin{pmatrix} -4 & 1 & & 0 \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & -4 \end{pmatrix}.$$

The matrices $B \in \mathbb{C}^{N^2 \times N}$, $C \in \mathbb{C}^{N \times N^2}$ and $D \in \mathbb{C}^{N \times N}$ are defined as

$$B \overset{\text{def}}{=} \frac{1}{h^2} \begin{pmatrix} I_N \\ \mathbb{O}_{N^2-N,N} \end{pmatrix}, \quad C \overset{\text{def}}{=} -\frac{1}{h} \begin{pmatrix} I_N & \mathbb{O}_{N,N^2-N} \end{pmatrix}, \quad D = \frac{1}{h} I_N.$$

Without any input, the approximate solution given by implicit Euler time discretization of $(\mathbf{P}_h)$ is given in Figure 3. For the null-controllability of the heat

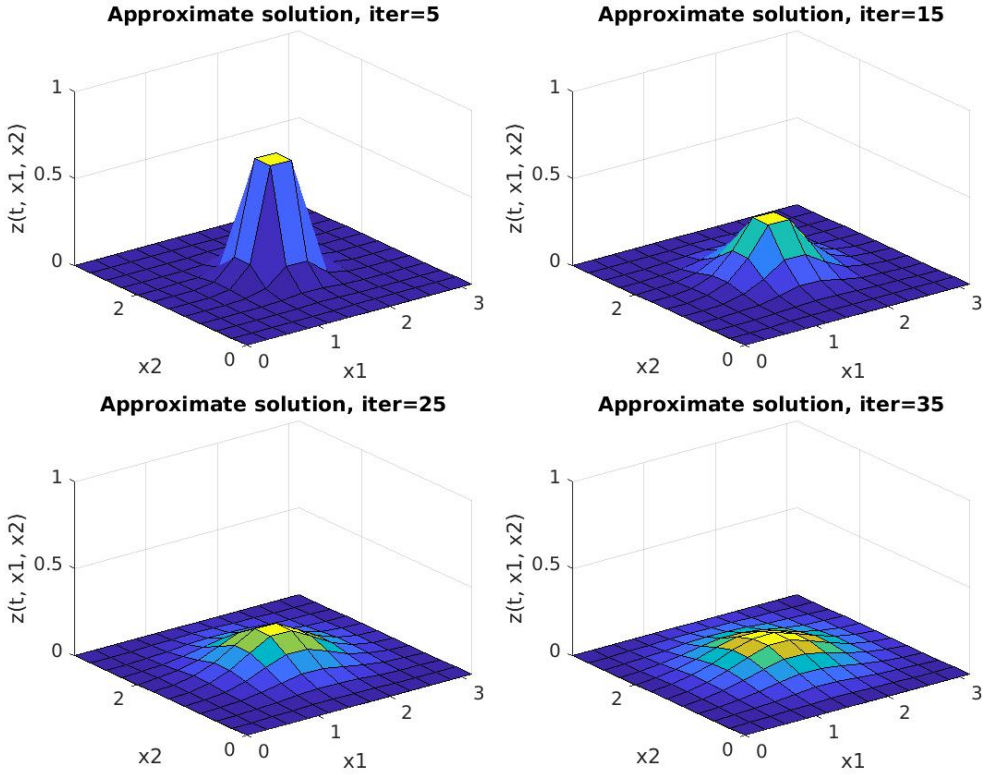

FIGURE 3. Solution without input

equation with boundary control and its finite difference semi-discretization, as well as for the study of the limit of the controls of the semi-discrete systems as mesh-size goes to zero, we refer to [**25**, Section 3, p.18-27]. Controllability of the heat equation and its semi-discrete approximation is also considered in [**24**], notably in the Section 6.3, p.505-506 for the 1-d semi-discretized problem.

The balanced truncation (Algorithm 2) may now be applied to the system $(\mathbf{P}_h)$, in order to reduce the order of the this system from N^2 to $\ell \leq N^2$. The simulations are done for the number of interior points $N = 10$, hence the

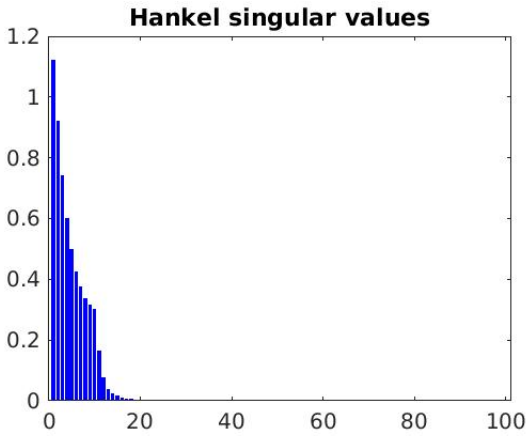

FIGURE 4

system is of order 100. We may choose ℓ by taking into account the decay of the Hankel singular values, which appear in the Figure 4. We observe in Figure 5 the decrease of $\|G(iw) - G_\ell(iw)\|$ across frequencies for ℓ going from 18 to 22, moreover we observe that the theoretical error bounds for the H^∞-error are respected in this experiment (for ℓ going from 15 to 35).

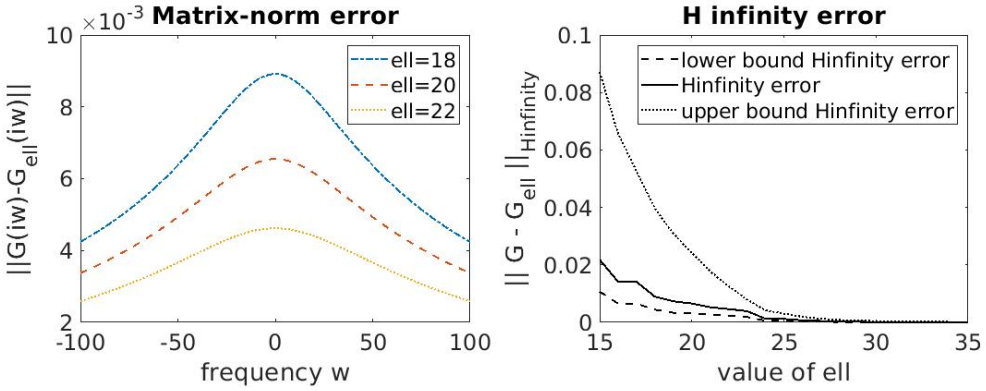

FIGURE 5. Errors and error bounds

Remark 10.1 **(Testing minimality).** — Set $n \stackrel{\text{def}}{=} N^2$. `Matlab` provides a control system toolbox which allows to compute the Gramians $\mathcal{R}$, $\mathcal{Q}$ and the

matrices

$$M_{\mathrm{c}} \overset{\text{def}}{=} \begin{pmatrix} B & AB & \dots & A^{n-1}B \end{pmatrix}, \quad M_{\mathrm{o}} \overset{\text{def}}{=} \begin{pmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{pmatrix}.$$

In order to test controllability, it may be natural to compute the rank of $\mathcal{R}$ or the rank of M_{c} for example. However, as it is indicated in the documentation of `Matlab`, the rank of the M_{c} is very sensitive to rounding errors and errors in the data. Another test for the controllability and observability, which is numerically effective (and provided by `Matlab` as well), is known as the *staircase algorithm*, see [**6**, Section 6, p.173].

APPENDIX

Laplace and Fourier transforms

$\mathcal{L}$ denotes the (unilateral) Laplace transform (see for example [**3**]). For E, F two Banach spaces, we both consider the Laplace transform of E-valued functions and $\mathscr{L}(E,F)$-valued functions. Let $f\colon \mathbb{R}_+ \longrightarrow E$, its Laplace transform $\mathcal{L}f$ is defined by the following limit in E

$$(\mathcal{L}f)(\lambda) \stackrel{\text{def}}{=} \lim_{t\to\infty} \int_0^t e^{-\lambda\tau} f(\tau) d\tau.$$

Here, we are interested in functions $f \in L^2(\mathbb{R}_+; E)$, in which case $\mathcal{L}f$ is well defined on $\mathbb{C}_+$. For operator valued functions $T\colon \mathbb{R}_+ \longrightarrow \mathscr{L}(E,F)$, $\mathcal{L}T$ is defined by the following limit in $\mathscr{L}(E,F)$

$$(\mathcal{L}T)(\lambda) \stackrel{\text{def}}{=} \lim_{t\to\infty} \int_0^t e^{-\lambda\tau} T(\tau) d\tau,$$

where $\int_0^t e^{-\lambda\tau} T(\tau) d\tau$ denotes the operator

$$\Big(x \mapsto \int_0^t e^{-\lambda\tau} T(\tau) x d\tau\Big) \in \mathscr{L}(E,F).$$

In particular, we consider here the Laplace transform of $T(\tau) \stackrel{\text{def}}{=} e^{\tau A} \in \mathscr{L}(\mathbb{C}^n)$ for an Hurwitz matrix A [7], in which case $\mathcal{L}T$ is well defined on the closure of $\mathbb{C}_+$.

$\mathscr{F}$ denotes the Fourier transform. Let X be a Hilbert space and $f\colon \mathbb{R} \longrightarrow X$ an element of $L^1(\mathbb{R}; X)$. The Fourier transform of f, denoted by $\mathscr{F}f$, is defined for any $w \in \mathbb{R}$ by

$$(\mathscr{F}f)(w) \stackrel{\text{def}}{=} \int_{-\infty}^{+\infty} e^{-iw\tau} f(\tau) d\tau.$$

The restriction of $\mathscr{F}$ to $L^1(\mathbb{R}; X) \cap L^2(\mathbb{R}; X)$ extends as a bounded linear operator from $L^2(\mathbb{R}; X)$ into $L^2(\mathbb{R}; X)$ (which is also denoted by $\mathscr{F}$). This follows from the fact that $L^1(\mathbb{R}; X) \cap L^2(\mathbb{R}; X)$ is dense in $L^2(\mathbb{R}; X)$, and from Plancherel's theorem which states that

$$\|\mathscr{F}f\|_{L^2(\mathbb{R};X)} = \sqrt{2\pi}\|f\|_{L^2(\mathbb{R};X)}, \quad \forall f \in L^1(\mathbb{R}; X) \cap L^2(\mathbb{R}; X).$$

7. Meaning that all eigenvalues of A have negative real parts.

Then, for $f \in L^2(\mathbb{R};X)$,

$$\mathscr{F}f \overset{\text{def}}{=} \lim_{k\to\infty} \mathcal{F}f_k \quad \text{and} \quad \|\mathscr{F}f\|_{L^2(\mathbb{R};X)} = \sqrt{2\pi}\|f\|_{L^2(\mathbb{R};X)},$$

where the above limit is taken in $L^2(\mathbb{R};X)$. Here, (f_k) denotes some sequence included in $L^1(\mathbb{R};X)\cap L^2(\mathbb{R};X)$ satisfying $\lim_{k\to\infty} f_k = f$ in $L^2(\mathbb{R};X)$. Moreover, $\frac{1}{\sqrt{2\pi}}\mathscr{F}$ is a unitary operator from $L^2(\mathbb{R};X)$ onto $L^2(\mathbb{R};X)$. For a proof, see [**3**, Theorem 1.8.2, p.45].
However, in order to stay consistent with the studied literature, instead of using the preceding definition $\mathscr{F}$ for the Fourier transform, we use the following definition:

$$(\mathcal{F}f)(iw) \overset{\text{def}}{=} \int_{-\infty}^{+\infty} e^{-iw\tau} f(\tau)d\tau, \quad \forall iw \in i\mathbb{R},\ \forall f \in L^1(\mathbb{R};X).$$

and see $\frac{1}{\sqrt{2\pi}}\mathcal{F}$ as a bijective isometric operator from $L^2(\mathbb{R},X)$ onto $L^2(i\mathbb{R},X)$.

Complex analysis

Proposition 10.2. — *If f : clos$(\mathbb{C}_+) \to \mathbb{C}$ is analytic on $\mathbb{C}_+$, and is also continuous and bounded on the closure of $\mathbb{C}_+$, then*

$$\sup_{\zeta\in\mathbb{C}_+} |f(\zeta)| = \sup_{w\in\mathbb{R}} |f(iw)|.$$

Proof. — For a fixed $\varepsilon > 0$, define $f_\varepsilon\colon \text{clos}(\mathbb{C}_+) \to \mathbb{C}$ by

$$f_\varepsilon(\zeta) = \frac{f(\zeta)}{1+\varepsilon\zeta},$$

which is analytic on $\mathbb{C}_+$ and continuous on clos$(\mathbb{C}_+)$ [8]. Since f is bounded, $f_\varepsilon(\zeta)$ tends to zero as $|\zeta|$ goes to $+\infty$. As a consequence of this, and of the continuity of f_ε on clos$(\mathbb{C}_+)$, we deduce that $|f_\varepsilon|$ is bounded and attains its maximum in clos$(\mathbb{C}_+)$, meaning that there exists $\zeta_0 \in \text{clos}(\mathbb{C}_+)$ satisfying

$$|f_\varepsilon(\zeta_0)| = \max_{\zeta\in\text{clos}(\mathbb{C}_+)} |f_\varepsilon(\zeta)|.$$

Assume by contradiction that $\zeta_0 \in \mathbb{C}_+$, and fix $r > 0$ so that the ball

$$\overline{B}(\zeta_0, r) \overset{\text{def}}{=} \{\zeta \in \mathbb{C}\colon |\zeta - \zeta_0| \le r\}$$

8. Indeed, $1+\varepsilon\zeta$ cannot be equal to zero on $\mathbb{C}_+$, having chosen positive ε.

is included in $\mathbb{C}_+$. The function f_ε is analytic on $\mathbb{C}_+$ and not constant [9], thus the maximum modulus theorem (see [**19**, 10.24, Chapter 10, p.212]) affirms that

$$|f_\varepsilon(\zeta_0)| < \max_{\theta\in[0,2\pi]} |f_\varepsilon(\zeta_0 + re^{i\theta})|.$$

This leads to a contradiction since we already know that

$$\max_{\theta\in[0,2\pi]} |f_\varepsilon(\zeta_0 + re^{i\theta})| \leq |f_\varepsilon(\zeta_0)|.$$

Hence, $\zeta_0 \in i\mathbb{R}$, and for any $\zeta \in \mathrm{clos}(\mathbb{C}_+)$, we have

$$|f_\varepsilon(\zeta)| \leq \max_{w\in\mathbb{R}} |f_\varepsilon(iw)|.$$

The right-hand side of the above equation is bounded by $\sup_{w\in\mathbb{R}} |f(iw)|$ for any $\varepsilon > 0$, since

$$|f_\varepsilon(iw)| = \frac{|f(iw)|}{\sqrt{1+\varepsilon^2 w^2}} \leq |f(iw)|$$

holds for any $w \in \mathbb{R}$. We have obtained that for any $\zeta \in \mathrm{clos}(\mathbb{C}_+)$ and any $\varepsilon > 0$,

$$|f_\varepsilon(\zeta)| \leq \sup_{w\in\mathbb{R}} |f(iw)|.$$

Taking the limit as ε goes to zero yields $\sup_{\zeta\in\mathrm{clos}(\mathbb{C}_+)} |f(\zeta)| \leq \sup_{w\in\mathbb{R}} |f(iw)|$, hence [10]

$$\sup_{\zeta\in\mathbb{C}_+} |f(\zeta)| = \sup_{\zeta\in\mathrm{clos}(\mathbb{C}_+)} |f(\zeta)| = \sup_{w\in\mathbb{R}} |f(iw)|. \qquad \square$$

Sketch of proof of Proposition 5.6. — Denote by $q \overset{\text{def}}{=} mp$ the dimension of $\mathbb{C}^{p\times m}$, and by $\{\mu_k\}_{k=1}^q$ an orthonormal basis of $\mathbb{C}^{p\times m}$. Let $\zeta \in \mathrm{clos}(\mathbb{C}_+)$, then $F(\zeta)$ rewrites as

$$F(\zeta) = \sum_{k=1}^{q} \big\langle F(\zeta), \mu_k \big\rangle_{\mathbb{C}^{p\times m}} \mu_k.$$

9. Indeed, $f(\zeta)$ cannot both be bounded on $\mathrm{clos}(\mathbb{C}_+)$ and have the form $\alpha(1+\varepsilon\zeta)$. Besides, f does not depend on ε.

10. In order to show the first inequality, we may firstly observe that we have the inequality $\sup_{\zeta\in\mathbb{C}_+} |f(\zeta)| \leq \sup_{\zeta\in\mathrm{clos}(\mathbb{C}_+)} |f(\zeta)|$. Secondly, for any $\zeta^\infty \in \mathrm{clos}(\mathbb{C}_+)$, there exists a sequence $(\zeta_k)_k \subset \mathbb{C}_+$ converging toward ζ^∞, and f being continuous, $|f(\zeta_k)| \to |f(\zeta^\infty)|$ as k goes to $+\infty$. However, since for any $k \in \mathbb{N}$ we have the uniform bound $|f(\zeta_k)| \leq \sup_{\zeta\in\mathbb{C}_+} |f(\zeta)|$, the limit must also satisfy $|f(\zeta^\infty)| \leq \sup_{\zeta\in\mathbb{C}_+} |f(\zeta)|$.

Moreover, Pythagoras' theorem yields

$$\|F(\zeta)\|^2 = \sum_{k=1}^{q} \left|\left\langle F(\zeta), \mu_k \right\rangle_{\mathbb{C}^{p\times m}}\right|^2.$$

Hence, the H^∞-norm of F also has the form

$$\|F\|^2_{H^\infty_{p\times m}} = \sup_{\zeta\in\mathbb{C}_+} \|F(\zeta)\|^2 = \sup_{\zeta\in\mathbb{C}_+} \Big(\sum_{k=1}^{q} \left|\left\langle F(\zeta), \mu_k \right\rangle_{\mathbb{C}^{p\times m}}\right|^2 \Big).$$

We may use arguments similar to Proposition 10.2, using the maximum modulus principle for Banach-valued functions (see [**9**, Section III.14, p.230]). Indeed, the function

$$F_\varepsilon(\zeta) \stackrel{\text{def}}{=} \frac{F(\zeta)}{1+\varepsilon\zeta}$$

cannot satisfy, for some $c > 0$, $\|F_\varepsilon(\zeta)\| = c$ for any $\zeta \in \mathbb{C}_+$. Otherwise, $\zeta \mapsto \left\langle F_\varepsilon(\zeta), \mu_k \right\rangle_{\mathbb{C}^{p\times m}}$ would also be constant, for any $k \in \{1, \ldots, q\}$, hence so would be F_ε.

□

Operations on transfer functions

Proposition 10.3 **(Product and sum of tranfer functions)**

Consider the matrices $A_1 \in \mathbb{C}^{n_1\times n_1}, A_2 \in \mathbb{C}^{n_2\times n_2}$, $B_1 \in \mathbb{C}^{n_1\times m_1}, B_2 \in \mathbb{C}^{n_2\times m_2}$, $C_1 \in \mathbb{C}^{p_1\times n_1}, C_2 \in \mathbb{C}^{p_2\times n_2}$ and $D_1 \in \mathbb{C}^{p_1\times m_1}, D_2 \in \mathbb{C}^{p_2\times m_2}$, which define two transfer functions $G_1(s) \in \mathbb{C}^{p_1\times m_1}, G_2(s) \in \mathbb{C}^{p_2\times m_2}$ by

$$G_1(s) \stackrel{def}{=} \left[\begin{array}{c|c} A_1 & B_1 \\ \hline C_1 & D_1 \end{array}\right], \quad G_2(s) \stackrel{def}{=} \left[\begin{array}{c|c} A_2 & B_2 \\ \hline C_2 & D_2 \end{array}\right].$$

When m_1, p_1 respectively agree with m_2, p_2, we set $m \stackrel{def}{=} m_1 = m_2$, $p \stackrel{def}{=} p_1 = p_2$ and the sum $G_1(s) + G_2(s) \in \mathbb{C}^{p\times m}$ is given by

$$G_1(s) + G_2(s) = \left[\begin{array}{cc|c} A_1 & \mathbb{O}_{n_1,n_2} & B_1 \\ \mathbb{O}_{n_2,n_1} & A_2 & B_2 \\ \hline C_1 & C_2 & D_1 + D_2 \end{array}\right].$$

When m_1 agrees with p_2, we set $k \overset{def}{=} m_1 = p_2$ and the product $G_1(s)G_2(s) \in \mathbb{C}^{p_1 \times m_2}$ is given by

$$G_1(s)G_2(s) = \left[\begin{array}{cc|c} A_1 & B_1C_2 & B_1D_2 \\ \mathbb{O}_{n_2,n_1} & A_2 & B_2 \\ \hline C_1 & D_1C_2 & D_1+D_2 \end{array}\right].$$

Proof. — We consider $s \in \mathbb{C}$ for which $G_1(s)$ and $G_2(s)$ are well defined. For the sum, we may directly compute

$$\begin{aligned} &G_1(s)+G_2(s) \\ &= C_1(sI_{n_1}-A_1)^{-1}B_1 + D_1 + C_2(sI_{n_2}-A_2)^{-1}B_2 + D_2 \\ &= \begin{pmatrix} C_1 & C_2 \end{pmatrix} \begin{pmatrix} (sI_{n_1}-A_1)^{-1} & \mathbb{O}_{n_1,n_2} \\ \mathbb{O}_{n_2,n_1} & (sI_{n_2}-A_2)^{-1} \end{pmatrix} \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} + D_1 + D_2 \\ &= \begin{pmatrix} C_1 & C_2 \end{pmatrix} \begin{pmatrix} sI_{n_1}-A_1 & \mathbb{O}_{n_1,n_2} \\ \mathbb{O}_{n_2,n_1} & sI_{n_2}-A_2 \end{pmatrix}^{-1} \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} + D_1 + D_2 \\ &= \begin{pmatrix} C_1 & C_2 \end{pmatrix} \left(sI_{n_1+n_2} - \begin{pmatrix} A_1 & \mathbb{O}_{n_1,n_2} \\ \mathbb{O}_{n_2,n_1} & A_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} + D_1 + D_2 \\ &= \left[\begin{array}{cc|c} A_1 & \mathbb{O}_{n_1,n_2} & B_1 \\ \mathbb{O}_{n_2,n_1} & A_2 & B_2 \\ \hline C_1 & C_2 & D_1+D_2 \end{array}\right]. \end{aligned}$$

Whereas for the product we begin by considering the systems associated with $G_1(s)$ and $G_2(s)$

$$(\mathbf{S}_{G_1}): \left\{ \begin{array}{l} \dfrac{\mathrm{d}}{\mathrm{d}t} z_1 = A_1 z_1 + B_1 f, \quad z_1(0) = z_1^0 \\ y_1 = C_1 z_1 + D_1 f \end{array} \right.$$

and

$$(\mathbf{S}_{G_2}): \left\{ \begin{array}{l} \dfrac{\mathrm{d}}{\mathrm{d}t} z_2 = A_2 z_2 + B_2 u_2, \quad z_2(0) = z_2^0 \\ f = C_2 z_2 + D_2 u_2 \end{array} \right.$$

where the input function of $(\mathbf{S}_{G_1})$ is imposed to be the output function of $(\mathbf{S}_{G_2})$. This is equivalent to considering the system

$$(\mathbf{S}_{G_1G_2}):\ \begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}z_1 = A_1z_1 + B_1(C_2z_2 + D_2u_2), & z_1(0) = z_1^0 \\ \dfrac{\mathrm{d}}{\mathrm{d}t}z_2 = A_2z_2 + B_2u_2, & z_2(0) = z_2^0 \\ y_1 = C_1z_1 + D_1(C_2z_2 + D_2u_2) \end{cases}$$

which, setting $z \overset{\text{def}}{=} \begin{pmatrix} z_1 & z_2 \end{pmatrix}^T$, may be rewritten as

$$(\mathbf{S}_{G_1G_2}):\ \begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}z = \begin{pmatrix} A_1 & B_1C_2 \\ \mathbb{O}_{n_2,n_1} & A_2 \end{pmatrix} z + \begin{pmatrix} B_1D_2 \\ B_2 \end{pmatrix} u_2, \quad z(0) = \begin{pmatrix} z_1^0 & z_2^0 \end{pmatrix}^T \\ y_1 = \begin{pmatrix} C_1 & D_1C_2 \end{pmatrix} z + D_1D_2u_2 \end{cases}$$

whose transfer function is

$$\begin{aligned} G(s) &\overset{\text{def}}{=} \left[\begin{array}{cc|c} A_1 & B_1C_2 & B_1D_2 \\ \mathbb{O}_{n_2,n_1} & A_2 & B_2 \\ \hline C_1 & D_1C_2 & D_1 + D_2 \end{array}\right]. \\ &= \begin{pmatrix} C_1 & D_1C_2 \end{pmatrix} \left(sI_{n_1+n_2} - \begin{pmatrix} A_1 & B_1C_2 \\ \mathbb{O}_{n_2,n_1} & A_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} B_1D_2 \\ B_2 \end{pmatrix} + D_1D_2 \\ &= \begin{pmatrix} C_1 & D_1C_2 \end{pmatrix} \begin{pmatrix} sI_{n_1} - A_1 & -B_1C_2 \\ \mathbb{O}_{n_2,n_1} & sI_{n_2} - A_2 \end{pmatrix}^{-1} \begin{pmatrix} B_1D_2 \\ B_2 \end{pmatrix} + D_1D_2 \\ &= \begin{pmatrix} C_1 & D_1C_2 \end{pmatrix} M^{-1} \begin{pmatrix} B_1D_2 \\ B_2 \end{pmatrix} + D_1D_2. \end{aligned}$$

Here we have set

$$M \overset{\text{def}}{=} \begin{pmatrix} sI_{n_1} - A_1 & -B_1C_2 \\ \mathbb{O}_{n_2,n_1} & sI_{n_2} - A_2 \end{pmatrix}.$$

Remark that M is indeed invertible since multiplying it on the right by the invertible matrix

$$\begin{pmatrix} I_{n_1} & (sI_{n_1} - A_1)^{-1}B_1C_2 \\ \mathbb{O}_{n_2,n_1} & I_{n_2} \end{pmatrix}$$

yields

$$M \begin{pmatrix} I_{n_1} & (sI_{n_1} - A_1)^{-1} B_1 C_2 \\ \mathbb{O}_{n_2,n_1} & I_{n_2} \end{pmatrix} = \begin{pmatrix} sI_{n_1} - A_1 & \mathbb{O}_{n_1,n_2} \\ \mathbb{O}_{n_2,n_1} & sI_{n_2} - A_2 \end{pmatrix}$$

and the right-hand side is invertible.

We may verify that $G(s)$ is indeed equal to $G_1(s)G_2(s)$. In order to do this, we will need to rewrite M^{-1} in term of its block components and their inverses. We just noticed that

$$M = \begin{pmatrix} sI_{n_1} - A_1 & \mathbb{O}_{n_1,n_2} \\ \mathbb{O}_{n_2,n_1} & sI_{n_2} - A_2 \end{pmatrix} \begin{pmatrix} I_{n_1} & -(sI_{n_1} - A_1)^{-1} B_1 C_2 \\ \mathbb{O}_{n_2,n_1} & I_{n_2} \end{pmatrix}.$$

Hence,

$$\begin{aligned} M^{-1} &= \begin{pmatrix} I_{n_1} & (sI_{n_1} - A_1)^{-1} B_1 C_2 \\ \mathbb{O}_{n_2,n_1} & I_{n_2} \end{pmatrix} \begin{pmatrix} (sI_{n_1} - A_1)^{-1} & \mathbb{O}_{n_1,n_2} \\ \mathbb{O}_{n_2,n_1} & (sI_{n_2} - A_2)^{-1} \end{pmatrix} \\ &= \begin{pmatrix} (sI_{n_1} - A_1)^{-1} & (sI_{n_1} - A_1)^{-1} B_1 C_2 (sI_{n_2} - A_2)^{-1} \\ \mathbb{O}_{n_2,n_1} & (sI_{n_2} - A_2)^{-1} \end{pmatrix}. \end{aligned}$$

Using this new form for M^{-1}, we observe that the transfer function $G(s)$ may then be rewritten as

$$\begin{aligned} &\begin{pmatrix} C_1 & D_1 C_2 \end{pmatrix} \begin{pmatrix} (sI_{n_1} - A_1)^{-1} & (sI_{n_1} - A_1)^{-1} B_1 C_2 (sI_{n_2} - A_2)^{-1} \\ \mathbb{O}_{n_2,n_1} & (sI_{n_2} - A_2)^{-1} \end{pmatrix} \begin{pmatrix} B_1 D_2 \\ B_2 \end{pmatrix} \\ &\quad + D_1 D_2, \end{aligned}$$

which, in turn, rewrites as

$$\begin{aligned} G(s) &= C_1 (sI_{n_1} - A_1)^{-1} B_1 C_2 + C_1 (sI_{n_1} - A_1)^{-1} B_1 C_2 (sI_{n_2} - A_2)^{-1} B_2 \\ &\quad + D_1 C_2 (sI_{n_2} - A_2)^{-1} B_2 + D_1 D_2 \\ &= \big(C_1 (sI_{n_1} - A_1)^{-1} + D_1\big)\big(C_2 (sI_{n_2} - A_2)^{-1} + D_2\big) \\ &= G_1(s) G_2(s). \end{aligned}$$

□

References

[1] Aliprantis, C. D., and Burkinshaw, O. *Principles of real analysis*. Academic press, 1998.

[2] Antoulas, A. C. *Approximation of Large-Scale Dynamical Systems*. Society for Industrial and Applied Mathematics, 2005.

[3] Arendt, W., Batty, C. J., Hieber, M., and Neubrander, F. *Vector-valued Laplace Transforms and Cauchy Problems*, 2 ed., vol. 96 of *Monographs in Mathematics*. Birkhaüser Basel, 2011.

[4] Baur, U., Benner, P., and Feng, L. Model order reduction for linear and nonlinear systems: A system-theoretic perspective. *Arch Computat Methods Eng 21*, 2 (2014), 331âĂŞ358.

[5] Benner, P., Mehrmann, V., and Sorensen, D. C., Eds. *Dimension Reduction of Large-Scale Systems* (Oberwolfach, Germany, 2003), vol. 45 of *Lecture Notes in Computational Science and Engineering*, Springer.

[6] Datta, B. N. *Numerical Methods for Linear Control Systems*. Elsevier Science and Technology Books, 2003.

[7] Diestel, J., and Uhl, J. J. *Vector Measures*. American Mathematical Society, 1977.

[8] Dullerud, G. E., and Paganini, F. *A Course in Robust Control Theory, A Convex Approach*. Springer, 2000.

[9] Dunford, N., and T., S. J. *Linear operators, Part 1*. Interscience publishers, Inc., New York, 1958.

[10] Eftang, J. L., Patera, A. T., and Ronquist, E. M. An hp certified reduced basis method for parametrized elliptic partial differential equations. *SIAM Journal on Scientific Computing 32*, 6 (2010), 3170–3200.

[11] Evans, L. C. *Partial Differential Equations*, 2 ed. American Mathematical Society, 2010.

[12] Glover, K. All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds. *International Journal of Control 39*, 6 (1984), 1115–1193.

[13] Gubisch, M., and Volkwein, S. Proper orthogonal decomposition for linear-quadratic optimal control. In *Model Reduction and Approximation: Theory and Algorithms*, P. Benner, M. Ohlberger, A. Cohen, and K. Willcox, Eds. Society for industrial and applied mathematics, 2017, ch. 1, pp. 3–63.

[14] Kunisch, K., and Volkwein, S. Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics. *SIAM Journal on Numerical Analysis 40*, 2 (2002), 492–515.

[15] Partington, J. R. *Linear operators and linear systems, An analytical approach to control theory*. London Mathematical Society Student texts, 2004.

[16] Partington, Jonathan, R. *An introduction to Hankel operators*, vol. 13 of *London Mathematical Society Student Texts*. Cambridge University Press, 1989.

[17] Reed, M., and Simon, B. *Methods of Modern Mathematical Physics, I: Functional Analysis*. Academic Press, 1972.

[18] Rowley, C. W. Model reduction for fluids, using balanced proper orthogonal decomposition. *International Journal of Bifurcation and Chaos 15*, 3 (2005), 997–1013.

[19] Rudin, W. *Real and complex analysis.* International student edition, 1970.

[20] Tombs, M. S., and Postlethwaite, I. Truncated balanced realization of a stable non-minimal state-space system. *International Journal of Control 46*, 4 (1987), 1319–1330.

[21] Tucsnak, M., and Weiss, G. *Observation and Control for Operator Semigroups.* Birkhaüser Verlag AG, 2009.

[22] Volkwein, S. Optimal control of a phase-field model using proper orthogonal decomposition. *ZAMM Journal of applied mathematics and mechanics: Zeitschrift fuür angewandte Mathematik und Mechanik 81*, 2 (2001), 83–97.

[23] Yosida, K. *Functional analysis, Sixth edition.* Springer-Verlag, Berlin Heidelberg New York 1980, 1995.

[24] Zuazua, E. Controllability of partial differential equations and their semi-discrete approximations. *Discrete and continuous dynamical systems 8*, 2 (2002), 469–513.

[25] Zuazua, E. Control and numerical approximation of the w ave and heat equations. *Eur. Math. Soc. III* (2006), 1389–1417.

Charlotte Rodriguez, Master Analyse, Équations aux Dérivées Partielles, Probabilités
E-mail : charlotte.rodriguez@u-bordeaux.fr