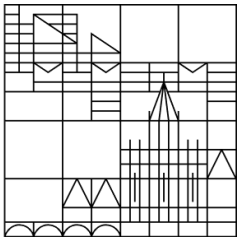
Master Thesis

Data-Driven Model-Order Reduction for Model Predictive Control

by Jan Rohleff



Department of Mathematics and Statistics

Supervisor and 1st Reviewer:	Prof. Dr. Stefan Volkwein
2nd Reviewer:	Jun. Prof. Dr. Behzad Azmi

March 21, 2023

Declaration

Hereby I declare that I have written the following thesis, entitled

Data-Driven Model-Order Reduction for Model Predictive Control

independently and have only used the sources indicated. All passages in the thesis in which words or ideas are taken from other sources have been marked as borrowings in each individual case by indicating the source and in the case of verbatim quotations by means of quotation marks. Furthermore, I assert that this thesis has not been submitted in fulfillment of course requirements for another course.

Konstanz, March 21, 2023 Jan Rohleff

Abstract

In this thesis, quadratic optimal control problems for linear parabolic partial differential equations (PDEs) with time-dependent coefficient functions are considered. After showing the existence and uniqueness of the solution, necessary and sufficient first order optimality conditions are derived. By applying a finite element (FE) discretization, the first-order optimality system can be represented as a linear time-variant (LTV) coupled dynamical system, which encompasses both the state equation and the dual equation. This leads us into the area of dynamical systems. Model predictive control (MPC) is applied to solve the problem over the long-time horizon. To speedup the computational time three data-driven model-order reduction (MOR) techniques are applied: Proper orthogonal decomposition (POD), empirical gramians and extended dynamic mode decomposition (EDMD). Furthermore, an a-posteriori error analysis is conducted to guarantee the accuracy of the reduced model during the MPC. Numerical simulations illustrate the advantages and disadvantages of the various MOR techniques.

Zusammenfassung

In dieser Arbeit werden quadratische optimale Steuerungsprobleme für lineare parabolische partielle Differentialgleichungen (PDEs) mit zeitabhängigen Koeffizientenfunktionen betrachtet. Nachdem die Existenz einer eindeutigen Lösung gezeigt ist, werden notwendige und hinreichende Optimalitätsbedingungen erster Ordnung hergeleitet. Durch die Anwendung einer finiten Elementen (FE) Diskretisierung kann das Optimalitätssystem erster Ordnung als lineares zeitvariantes (LTV) gekoppeltes dynamisches System dargestellt werden, das sowohl die Zustandsgleichung als auch die Dualgleichung umfasst. Dies führt uns in den Bereich der dynamischen Systeme. Model Predictive Control (MPC) wird angewendet, um das Problem über einen langen Zeithorizont zu lösen. Um die Rechenzeit zu beschleunigen, werden drei datenbasierende Model-Order-Reduction (MOR) Techniken angewendet: Proper Orthogonal Decomposition (POD), empirische Gramians und extende dynamic mode Decomposition (EDMD). Darüber hinaus wird eine a-posteriori Fehleranalyse durchgeführt, um eine Approximationsgüte des reduzierten Modells während der MPC zu garantieren. Numerische Simulationen zeigen die Vor- und Nachteile der verschiedenen MOR-Techniken.

Contents

List of Acronyms

DMD	dynamic mode decomposition
EDMD	extended dynamic mode decomposition
FE	finite element
LTI	linear time-invariant
LTV	linear time-variant
MOR	model-order reduction
MPC	model predictive control
ODE	ordinary differential equation
PDE	partial differential equation
POD	proper orthogonal decomposition
ROM	reduced-order modeling
SVD	singular value decomposition
tSVD	truncated singular value decomposition

1 | Introduction

Nowadays optimization problems constrained by time-dependent partial differential equations (PDEs) are fundamental to many areas of engineering, science, and economics. These problems aim to find the best design or control strategy to steer a system towards a desired target while satisfying various constraints. The ability to control these systems optimally has numerous benefits, ranging from improved performance and efficiency to reduced costs and environmental impact.
One example is the administration of chemotherapy drugs to treat cancer. The drug concentration in the patient's body over time can be described by a transport partial differential equation (PDE) and the control input is the drug dosage. The goal is to minimize the toxicity of the drugs while maximizing their effectiveness in treating the cancer. The performance metric could be a combination of the drug concentration in the patient's body and the tumor size. By solving the optimal control problem, we can determine the optimal drug dosage that balances the trade-off between toxicity and effectiveness. This can lead to improved treatment outcomes and reduced side effects for patients undergoing chemotherapy.
Therefore, the study of optimal control for time-variant systems is not only of theoretical importance but also has significant practical applications. This makes it a challenging and rewarding area of research that holds the potential to address real-world problems and enhance our daily lives.
For solving optimal control problem over a large time horizon, Model Predictive Control (MPC) is used (cf. [13, 28]). MPC is a powerful method for solving optimal control problems. It predicts future system behavior and optimizes control actions to achieve desired objectives, taking into account constraints. MPC has proven to be effective in many applications due to its ability to handle complex and dynamic systems, adapt to changing conditions. Thus it can be easily adapted to changing objectives or system conditions, making it a versatile tool for real-time control.
It is generally not possible to derive explicit solution formulas for infinite-dimensional optimal control problems. A classical approach is to use discretization methods, such as finite elements, to approximate the problem as a finite, high-dimensional one. The complexity of the optimization problem is directly tied to the number of degrees of freedom in the discretization. Therefore, the computational demands of MPC can be a challenge, especially for systems with large state dimensions, as it requires solving an optimization problem at each time step.

In this thesis, we propose a data-driven approach to address this challenge by reducing the model complexity of MPC through model-order reduction (MOR) (cf. [1, 15]).

Therefore, we perform several numerical experiments on a specific optimization problem subject to a linear parabolic advection-diffusion equation. Our focus is on linear systems and quadratic cost functionals, which lead to convex optimization problems. The problem can also be extended for nonlinear systems if an efficient realization of the nonlinear term is available. The following data-driven MOR techniques are the evaluated:

- Proper orthogonal decomposition (cf., e.g., [23, 14]),
- Empirical gramains (cf., e.g., [25, 31]),
- Extended dynamic mode decomposition (cf., e.g., [24]).

The proposed data-driven MOR methods utilize snapshots of the system to effectively reduce the dimensionality of the model while preserving its key characteristics. Our goal is to emphasize the differences between the various MOR techniques. To apply these techniques, we first obtain the sufficient first order optimality conditions introducing the adjoint equation. Unlike conventional methods, the MOR techniques are not applied independently on the constraint state and adjoint equation. Instead, they are performed directly on the optimality system of the optimal control problem. This is similar to the previous work in [29] and this thesis serves as a complement and extension. As MPC shifts its time-horizon at each time step, the accuracy of the initial MOR approximation may not hold throughout the entire MPC calculation. One aim is therefore to design an efficient reduced basis update strategy that ensures a satisfactory approximation of the reduced model. This is done with a a-posteriori error estimator between the full and reduced model. Additionally, we evaluate the accuracy of the error measure through tests.

The objective of this research is to demonstrate the effectiveness of data-driven MOR techniques in improving the computational efficiency of MPC while maintaining its control performance.

1.1 Outline

This thesis is organized as follows.

- In Chapter 2, we present the relevant fundamental definitions, notations and theorems throughout this thesis. The focus is on the areas of dynamical systems and control theory, partial differential equations (PDE), and optimization.
- In Chapter 3, we introduce a general linear time-variant optimal control problem. We examine two cases: When the constrained dynamic of the optimal control problem is described by a linear time-variant ordinary differential equation (ODE) or partial differential equation (PDE). We demonstrate the existence and uniqueness of a solution and derive the optimality system of the problem by introducing the adjoint equation. The goal is to transform the optimality system so that it can be viewed as a dynamical system, which will be used for model-order reduction (MOR) in Chapter 4.

- In Chapter 4, we introduce the concept of MOR for optimal control problems considered in the previous chapter. We briefly explain how to apply MOR for our problem and derive a-posteriori error estimates for the reduced model. Then, we provide an overview of three different MOR techniques, proper orthogonal decomposition, empirical gramians and extended dynamic mode decomposition. All of this techniques are data-driven, which means that we require some starting snapshots of the full model in advance before we can execute the model-order reduction.

- In Chapter 5, we examine a particular example of an optimal control problem that was introduced in Chapter 3. We use numerical methods, specifically MPC, to solve this problem. Therefore, we begin the chapter by providing a brief introduction to MPC. The primary goal of this chapter is to demonstrate the effectiveness of the model-order reduction techniques for an introduced reduced MPC with error estimator, and to highlight the pros and cons of the different techniques.

- In chapter 6, we summarize the key takeaways of this thesis and provide insights into potential future research directions.

- In chapter 7, we provide an outlook of how to solve optimal control problems with additional control constraints, which leads to a nonsmooth optimality system. We begin demonstrating how one can solve the optimality system by using the semismooth Newton method. First, we provide a brief overview of the semismooth Newton method.

2 | Preliminaries

This chapter presents the relevant fundamental definitions, notations and theorems used in this thesis. The focus is on the areas of dynamical systems and control theory, partial differential equations (PDE), and optimization.

2.1 Dynamical Systems and Control Theory

We start recalling the most relevant definitions for dynamical systems from [49, Definitions 3.1 and 3.4]

Definition 2.1 (Ordinary differential equation and dynamical system)
Without loss of generality we consider a first order system: Given a function $\mathcal{F} : [t_0, T] \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_y}$ *with* $t_0, T \in \mathbb{R}$, $0 \leq t_0 < T$ *and* $n_y \in \mathbb{N}$. *The equation*

$$\dot{y}(t) = \mathcal{F}(t, y(t)), \quad y(t_0) = y_\circ \tag{2.1}$$

is called explicit ordinary differential equation *(ODE). If the first variable* t *of the map* $\mathcal{F}$ *represents the time* (2.1) *it is called dynamical system.*

Definition 2.2 (Input-output system)
Nonlinear input-output systems *are dynamical system of the form*

$$\dot{y}(t) = \mathcal{F}(t, y(t), u(t)) \qquad \text{for } t \in (0,T), \quad y(0) = y_\circ, \tag{2.2a}$$
$$z(t) = \mathcal{G}(y(t)) \qquad \text{for } t \in (0,T), \tag{2.2b}$$

where $y_\circ \in \mathbb{R}^{n_y}$ *is the initial condition and* $\mathcal{F} : [0,T] \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_y}$, $\mathcal{G} : \mathbb{R}^{n_y} \to \mathbb{R}^{n_z}$ *are given functions. We call* $y \in H^1(0,T;\mathbb{R}^{n_y})$ *the* state, $u \in L^2(0,T;\mathbb{R}^{n_u})$ *the* control *and* $z \in L^2(0,T;\mathbb{R}^{n_z})$ *the* output *of the dynamical system* (2.2).

Definition 2.3 (LTI system)
An input-output system is called linear time-invariant (LTI) dynamical system *if it can be described by the following linear constant coefficient differential equation*

$$\begin{aligned} \dot{y}(t) &= Ay(t) + Bu(t) \quad \text{for } t \in (0,T), \quad y(0) = y_\circ, \\ z(t) &= Cy(t) \qquad\qquad\quad \text{for } t \in (0,T), \end{aligned} \tag{2.3}$$

where $A \in \mathbb{R}^{n_y \times n_y}$, $B \in \mathbb{R}^{n_y \times n_u}$ *and* $C \in \mathbb{R}^{n_z \times n_y}$.

Definition 2.4 (Controllability)
The LTI system is called controllable *on* $[t_0, t_1]$ *if, for any state* $y(t_0) \in \mathbb{R}^{n_y}$*, final time* $t_1 > t_0 \geq 0$ *and final state* $y_{t_1} \in \mathbb{R}^{n_y}$ *there exists a piecewise continuous input* $u : [t_0, t_1] \to \mathbb{R}^{n_u}$ *such that the unique solution of* (2.3) *satisfies* $y(t_1) = y_{t_1}$*. Otherwise, the LTI system is said to be* uncontrollable *on* $[t_0, t_1]$*. The LTI system is called* controllable*, if it is controllable on* $[0, T]$*.*

Definition 2.5 (Observability)
The LTI system (2.3) *is said to be* observable *on* $[t_0, t_1]$*, if for any* $t_1 > t_0 \geq 0$ *the initial state* $y(t_0)$ *can be determined from the knowledge of the input* $u(t)$ *and the output* $y(t)$ *for* $t \in [t_0, t_1]$*. Otherwise* (2.3) *is called* unobservable *on* $[t_0, t_1]$*. The LTI system is called* observable*, if it is observable on* $[0, T]$

Theorem 2.6 (Gronwall's lemma)
Let $I := [a, b]$ *be an intervall,* $u, \alpha \in C(I, \mathbb{R})$ *and* $\beta \in C(I, [0, \infty))$*. Moreover, it holds*

$$u(t) \leq \alpha(t) + \int_a^t \beta(s)u(s)\,\mathrm{d}s$$

for all $t \in I$*. Then the Gronwall inequality*

$$u(t) \leq \alpha(t) + \int_a^t \alpha(s)\beta(s)e^{\int_s^t \beta(\sigma)\mathrm{d}\sigma}\,\mathrm{d}s$$

holds for all $t \in I$*.*

Proof. A proof is given in [9, Theorem 16.6]. □

2.2 Partial Differential Equations (PDE)

Let $(V, \langle\cdot,\cdot\rangle_V)$ and $(H, \langle\cdot,\cdot\rangle_H)$ be two separable Hilbert spaces with $V \subset H$ dense. We assume that $V \hookrightarrow H \simeq H' \hookrightarrow V'$ is a Gelfand triple (as defined in [48, Theorem 17.4]). For this thesis, we focus on the function spaces $V := H^1(\Omega)$ and $H := L^2(\Omega)$ for a bounded domain $\Omega \subset \mathbb{R}^n$, which lead to a Gelfand triple (a precise proof can be found in [48, Theorem 17.4]). In the following, let $T > 0$ and without loss of generality, we assume that all PDEs start at $t_0 = 0$.

We begin this section by defining the natural function space in which the types of partial differential equations that are studied in this thesis are examined.

Definition 2.7
We define

$$W(0,T) := W(0,T;V) := L^2(0,T;V) \cap H^1(0,T;V')$$

the normed linear space of all $\varphi \in L^2(0,T;V)$ *having a (distributional) derivative in* $\partial_t\varphi \in L^2(0,T;V')$*. Then* $W(0,T)$ *endowed with the inner product*

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

is even a Hilbert space (a proof is given in [48, Theorem 25.5]).

Since we are dealing with L^2-functions we need the following theorem to define later on meaningful initial conditions. Especially, we are interested in well-defined point evaluations, for example $y(0)$ in H for a given $y \in W(0,T)$.

Theorem 2.8 (Continuity of $W(0,T)$ functions)
If $\varphi \in W(0,T)$, then $\varphi \in C([0,T];H)$ holds and the embedding

$$W(0,T) \hookrightarrow C([0,T];H)$$

is continuous, i.e. there exists a constant $C > 0$ such that

$$\|\varphi\|_{C([0,T];H)} \leq C\|\varphi\|_{W(0,T)}$$

for all $\varphi \in W(0,T)$.

Proof. A proof is given in [48, Theorem 25.5] □

The following result shows how to evaluate Sobolev functions on the boundary of a smooth domain. This theorem is used to include boundary conditions in the weak formulation of a PDE.

Theorem 2.9 (Trace theorem)
Let $1 \leq p < \infty$ and assume that $\Omega \subset \mathbb{R}^n$ is bounded with Lipschitz-continuous boundary $\partial\Omega$. Then there exists a bounded linear operator

$$T : W^{1,p}(\Omega) \to L^p(\partial\Omega)$$

such that

(i) $Tu = u|_{\partial\Omega}$, if $u \in W^{1,p}(\Omega) \cap C(\bar{\Omega})$.

(ii) $\|Tu\|_{L^p(\partial\Omega)} \leq C\|u\|_{W^{1,p}(\Omega)}$ for each $u \in W^{1,p}(\Omega)$, with the constant C depending only on p and Ω.

Proof. A proof is given in [11, Theorem 1 in Section 5.5] □

Now, we consider a general evolution equation of the form

$$\begin{aligned} y_t(t) + a(t; y(t), \cdot) &= f(t,\cdot) \quad \text{in } V',\ t \in (0,T), \\ y(0) &= y_\circ, \end{aligned} \tag{2.4}$$

where $a : [0,T] \times V \times V \to \mathbb{R}$, $f : [0,T] \times V \to \mathbb{R}$ and $y_\circ \in H$ stands for the initial condition. The aim is to develop a solution theory for the evolution equation (2.4). We aim to prove the existence and uniqueness of a solution, and thus, we require the following definition.

Definition 2.10 (Continuity and coercivity of time-dependent bilinear form)
A time-dependent bilinear form $a(t,\cdot,\cdot) : V \times V \to \mathbb{R}$ is called continuous *if*

(i) $|a(t,\varphi,\psi)| \leq \gamma\|\varphi\|_V\|\psi\|_V$ for all $\varphi, \psi \in V$ a.e. in $[0,T]$,

and coercive *if*

(ii) $a(t,\varphi,\varphi) \geq \eta_1 \|\varphi\|_V^2 - \eta_2 \|\varphi\|_H^2$ *for all* $\varphi \in V$ *a.e. in* $[0,T]$,

for constants $\gamma, \eta_1 > 0$ *and* $\eta_2 \geq 0$ *which do not depend on* t.

Now, we can state the final theorem, which guarantees the well-posedness, existence and uniqueness of a solution of the general evolution equation.

Theorem 2.11 (Existence and uniqueness of the evolution equation)
Let V *and* H *be two separable Hilbert spaces such that* $V \hookrightarrow H \simeq H' \hookrightarrow V'$ *is Gelfand triple. If the bilinear form* $a(t,\cdot,\cdot)$ *is continuous and coercive,* $y_\circ \in H$ *and* $f \in L^2(0,T;V')$ *then there exists a unique solution* $y \in W(0,T)$ *of* (2.4). *Furthermore, it holds*

$$\|y\|_{W(0,T)} \leq C\Big(\|y_\circ\|_H + \|f\|_{L^2(0,T;V')}\Big)$$

for a constant $C > 0$ *which is independent of* $y_\circ$ *and* f.

Proof. A proof is given in [8, pp. 512-520] □

2.3 Optimization

For this section, let $(Y, \|\cdot\|_Y)$, $(U, \|\cdot\|_U)$ and $(H, \|\cdot\|_H)$ be real Hilbert spaces with $Y \subset V$. We consider a a controlled evolution equation

$$\begin{aligned} y_t(t) + a(t; y(t), \cdot) &= \mathcal{B}(t)u(t) + f(t,\cdot) \quad \text{in } V',\ t \in (0,T), \\ y(0) &= y_\circ, \end{aligned} \tag{2.5}$$

where $\mathcal{B} \in L^2(0,T;\mathcal{L}(U,V'))$. The goal is to determine the control u as a solution of an optimization problem. Therefore, we define the following convex objective function $\mathcal{J} : Y \times U :\to \mathbb{R}$ by

$$\mathcal{J}(y,u) = \frac{1}{2}\|\mathcal{C}y - z_{\mathsf{d}}\|_H^2 + \frac{\sigma}{2}\|u\|_U^2, \tag{2.6}$$

where $z_{\mathsf{d}} \in H$ is a desired state, $y = y(u) \in Y$ is the solution of the controlled evolution equation (2.5), $\mathcal{C} : Y \to H$ is linear and bounded operator, $\sigma > 0$ is a regularization parameter. The optimal control problem reads as follow:

$$\min \mathcal{J}(y,u) \quad \text{subject to (s.t.)} \quad (y,u) \text{ satisfies (2.5)}. \tag{2.7}$$

This naturally leads us to the fundamental definitions of optimal control problems, and in this section, we present the main theorems.

Definition 2.12 (Optimal control)
Let $\mathcal{J} : Y \times U \to \mathbb{R}$ *be the* cost functional *and* $\mathcal{S} : U \to Y$ *be the solution operator of the dynamical system constraints. We define the reduced cost functional by*

$$\hat{\mathcal{J}}(u) := \mathcal{J}(\mathcal{S}u, u) : U \to \mathbb{R}.$$

An element $\bar{u} \in U$ *is called* optimal control *with corresponding* optimal state $\bar{y} = \mathcal{S}\bar{u}$, *if it holds*

$$\hat{J}(\bar{u}) \leq \hat{J}(u) \quad \textit{for all } u \in U.$$

An important concept of differentiability in general spaces, needed for optimal control problems, is the Gâteaux differentiability.

Definition 2.13 (Gâteaux differentiability)
Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be two normed vector spaces, $U \subset X$ an open set and $f : U \to Y$. The function f is called Gâteaux differentiable *in $u \in U$, if there is a linear and continuous operator $\mathcal{A} = \mathcal{A}_u : X \to Y$ such that*

$$\lim_{t \to 0} \frac{f(u + tu^\delta) - f(u) - Au^\delta}{t} = 0$$

holds for all $u^\delta \in X \backslash \{0\}$. The operator $\mathcal{A}$ is called the Gâteaux derivative *of f in u and we write $f'(u) := \mathcal{A} \in L(X, Y)$.*
If f is Gâteaux differentiable for every $u \in U$, we call the function f Gâteaux differentiable *and the function $f' : U \to \mathcal{L}(X, Y)$ the Gâteaux derivative of f.*

Since we primarily focus on convex optimal control problems, we often ustilize the favorable properties of convex optimization. As a result, we present the following theorem.

Theorem 2.14 (Existence and uniqueness for convex optimization problems)
Let $(X, \|\cdot\|_X)$ be a normed vector space and let $C \subseteq X$ be a nonempty and convex subset of X. For a convex cost functional $\mathcal{J} : C \to \mathbb{R}$ it holds:

(i) Every local minimum of $\mathcal{J}$ is a global minimum.

(ii) The set of minima is convex.

(iii) If $\mathcal{J}$ is strictly convex and $\bar{x} \in C$ is a local minimum of $\mathcal{J}$, then it is a unique strict local minimum of f.

(iv) Is C open and $\bar{x} \in C$ a stationary point (requires that $\mathcal{J}$ is continuous differentiable), then $\bar{x}$ is the global minimum of $\mathcal{J}$ in C.

Proof. The proof is analog to the proof given in [47, Theorem 3.7]. □

Theorem 2.15
Let $U_{ad} \subset U$ be a nonempty, bounded, closed and convex subset of U. Let $z_d \in H$, $\sigma \geq 0$ and the mapping $\mathcal{G} : U \to H$ be a linear and continuous operator. Then there exists an optimal control $\bar{u}$ solving

$$\min_{u \in U_{ad}} \mathcal{J}(u) := \frac{1}{2}\|\mathcal{G}u - z_d\|_H^2 + \frac{\sigma}{2}\|u\|_U^2. \tag{2.8}$$

If $\sigma > 0$ or $\mathcal{G}$ is injective, then the solution is unique.

Proof. A proof is given in [42, Theorem 2.14]. □

Theorem 2.16
If $\sigma > 0$ and U_{ad} is nonempty, closed and convex, problem (2.8) *has a unique optimal solution.*

Proof. A proof is given in [42, Theorem 2.16]. □

3 | Optimality Systems for linear time-variant Input-Output Systems

This chapter introduces a general linear time-variant optimal control problem. In Section 3.1, we begin by considering the optimal control problem when the dynamics are described by a linear time-variant ordinary differential equation (ODE). We show the existence and uniqueness of a solution and derive the optimality system of the problem by introducing the adjoint equation. The goal is to transform the optimality system so that it can be viewed as a dynamical system, which will be used for model-order reduction in Chapter 4. In Section 3.2, we extend the linear time-variant optimal control problem to a constrained dynamic governed by partial differential equations (PDE). We provide an overview of the finite element discretization of the PDE in space and show how it can be related back to the ODE case. Additionally, we discuss the difference in the order of optimization and discretization of the problem. Finally, we will also take into account the scenario where there are additional control constraints.

3.1 The ODE Case

In this section, we introduce a continuous optimal control problem governed by ordinary differential equations (ODEs). Therefore, let the state space be given by $Y := H^1(0,T;\mathbb{R}^{n_y})$ and the control space by $U := L^2(0,T;\mathbb{R}^{n_u})$. We will consider the following particular case of (2.2)

$$\dot{y}(t) = A(t)y(t) + B(t)u(t) + f(t) \qquad \text{for } t \in (0,T], \quad y(0) = y_\circ, \tag{3.1a}$$
$$z(t) = C(t)y(t) \qquad \text{for } t \in (0,T) \tag{3.1b}$$

with $f \in L^2(0,T;\mathbb{R}^{n_y})$, $A(t) \in \mathbb{R}^{n_y \times n_y}$, $B(t) \in \mathbb{R}^{n_y \times n_u}$ and $C(t) \in \mathbb{R}^{n_z \times n_y}$ for $t \in (0,T)$. System (3.1) is a linear time-variant system and moreover we assume that $A \in C([0,T],\mathbb{R}^{n_y \times n_y})$ is coercive, $B \in L^2(0,T;\mathbb{R}^{n_y \times n_u})$ and $C \in L^2(0,T;\mathbb{R}^{n_z \times n_y})$.

We are interested in controlling (3.1) in an optimal way. Thus we define the following objective function $\mathcal{J} : Y \times U :\to \mathbb{R}$ by

$$\begin{aligned}\mathcal{J}(y,u) &= \frac{1}{2}\int_0^T \|C(t)y(t) - z_{\mathsf{d}}(t)\|_Q^2 \,\mathrm{d}t + \frac{\sigma}{2}\int_0^T \|u(t)\|_R^2 \,\mathrm{d}t \\ &= \frac{1}{2}\int_0^T \|z(t) - z_{\mathsf{d}}(t)\|_Q^2 \,\mathrm{d}t + \frac{\sigma}{2}\int_0^T \|u(t)\|_R^2 \,\mathrm{d}t,\end{aligned} \tag{3.2}$$

where $z_{\mathsf{d}} \in L^2(0,T;\mathbb{R}^{n_z})$ is a desired state, $\sigma > 0$ is the regularization parameter, $Q \in \mathbb{R}^{n_z \times n_z}$ is a symmetric positive semi-definite matrix, $R \in \mathbb{R}^{n_u \times n_u}$ is a symmetric positive definite matrix, $\|\cdot\|_Q = \langle \cdot, \cdot \rangle_Q^{1/2}$ and $\|\cdot\|_R = \langle \cdot, \cdot \rangle_R^{1/2}$ hold. Moreover, we want to define the standard euclidean inner product by $\langle y, z \rangle_{\mathbb{R}^{n_y}} := \sum_{i=1}^{n_y} y_i z_i = y^\top z$ for $y, z \in \mathbb{R}^{n_y}$. The goal is to find a control $u \in L^2(0,T;\mathbb{R}^{n_u})$ that minimizes $\mathcal{J}$ subject to the constraint (3.1):

$$\min \mathcal{J}(y,u) \quad \text{subject to (s.t.)} \quad (y,u) \text{ satisfies (3.1).} \tag{$\mathbf{P}_{\mathrm{ODE}}$}$$

Definition 3.1 (Weighting matrices)
We define the following symmetric and positiv definite weighting matrices

(i) control space: $R \in \mathbb{R}^{n_u \times n_u}$,

(ii) state space: $W \in \mathbb{R}^{n_y \times n_y}$.

The inner products are defined analog, for example $\langle u, \tilde{u} \rangle_R = u^\top R \tilde{u}$ *for* $u, \tilde{u} \in \mathbb{R}^{n_u}$. *Note that, if we choose* $R = I \in \mathbb{R}^{n_u \times n_u}$, *where* I *is the identity, then we obtain the standard Euclidean inner product in* $\mathbb{R}^{n_u}$. *For the output space we define the following symmetric positive semidefinite weighting matrix*

(iii) output space: $Q \in \mathbb{R}^{n_y \times n_y}$

Including the weighting matrices gives us numerical flexibility and later, considering PDEs as constraints, the weighting matrices will be replaced by finite element matrices.

3.1.1 Theoretical Results

We start this subsection by showing the existence and uniqueness of an optimal solution to ($\mathbf{P}_{\mathrm{ODE}}$).

Theorem 3.2 (Existence and uniqueness of ($\mathbf{P}_{\mathrm{ODE}}$))
The optimal control problem ($\mathbf{P}_{\mathrm{ODE}}$) *has a unique solution* $(\bar{y}, \bar{u}) \in Y \times U$.

Proof. Firstly, we show the existence and uniqueness of a solution of the state equation (3.1a). Therefore, we consider equation (3.1a), as an equation in $(\mathbb{R}^{n_y})' = \mathbb{R}^{n_y}$, i.e.

$$\langle y_t(t), \varphi \rangle_W - \langle A(t)y(t), \varphi \rangle_W = \langle B(t)u(t) + f(t), \varphi \rangle_W \quad \text{for all } \varphi \in \mathbb{R}^{n_y},$$

Since A is coercive all requirements of Theorem 2.11 are fulfilled and there exists a unique solution of (3.1a) with

$$\|y\|_Y \leq \tilde{c}\Big(\|y_\circ\|_{\mathbb{R}^{n_y}} + \|B(\cdot)u + f\|_{L^2(0,T;\mathbb{R}^{n_y})} \Big), \tag{3.3}$$

where $\tilde{c} > 0$. Now, let $\mathcal{S} : U \to Y$ be the solution operator of the state equation (3.1a). Furthermore, we define the operator $\mathcal{G} : U \to L^2(0,T;\mathbb{R}^{n_z})$ by $(\mathcal{G}u)(t) := C(t)\mathcal{S}u(t)$. Then $\mathcal{G}$ is obviously linear in u and bounded, since it holds

$$\begin{aligned} \|\mathcal{G}u\|_{L^2(0,T;\mathbb{R}^{n_z})} &= \|C(\cdot)\mathcal{S}u\|_{L^2(0,T;\mathbb{R}^{n_z})} \\ &\leq \|C\|_{L^2(0,T;\mathbb{R}^{n_z \times n_y})} \|\mathcal{S}u\|_Y \\ &\leq \tilde{c}\|C\|_{L^2(0,T;\mathbb{R}^{n_z \times n_y})} \Big(\|y_\circ\|_{\mathbb{R}^{n_y}} + \|B(\cdot)u + f\|_{L^2(0,T;\mathbb{R}^{n_y})} \Big) < \infty. \end{aligned}$$

Consequently, Theorem 2.16 guarantees a unique solution $\bar{u} \in U$ of

$$\min_{u \in L^2(0,T;\mathbb{R}^{n_u})} \hat{\mathcal{J}}(u) := \frac{1}{2} \int_0^T \|(\mathcal{G}u)(t) - z_{\mathsf{d}}(t)\|_Q^2 \, \mathrm{d}t + \frac{\sigma}{2} \int_0^T \|u(t)\|_R^2 \, \mathrm{d}t$$

and therefore of $(\mathbf{P}_{\mathrm{ODE}})$. The optimal state $\bar{y}$ is given by $\bar{y} = \mathcal{S}\bar{u} \in Y$. □

The Lagrange functional $\mathcal{L} : Y \times U \times U \to \mathbb{R}$ associated with $(\mathbf{P}_{\mathrm{ODE}})$ is given by

$$\mathcal{L}(y, u, p) = \mathcal{J}(y, u) + \int_0^T \langle \dot{y}(t) - A(t)y(t) - B(t)u(t) - f(t), p(t) \rangle_W \, \mathrm{d}t. \tag{3.4}$$

A first order sufficient optimality condition of $(\mathbf{P}_{\mathrm{ODE}})$ can be derived either from the stationary condition of the reduced problem $\nabla \hat{\mathcal{J}}(\bar{u}) = 0$ (see i.e. Theorem 2.14) or from stationarity conditions of the Lagrangian (see, e.g. [4]), i.e. $\nabla \mathcal{L}(\bar{y}, \bar{u}, \bar{p}) = 0$. In the following, we focus on the first approach.

3.1.2 Reduced Problem

We start this subsection by defining the inhomogeneous part $\hat{y} \in Y$ of the state equation as the solution of

$$\dot{\hat{y}}(t) = A(t)\hat{y}(t) + f(t) \quad \text{for } t \in (0, T], \quad y(0) = y_\circ. \tag{3.5}$$

Remark that $\hat{y}$ is independent of the control u. Now, let $\mathcal{S} : U \to Y$ be the linear solution operator of the state equation such that $y = \mathcal{S}u \in Y$ solves

$$\dot{y}(t) = A(t)y(t) + B(t)u(t) \quad \text{for } t \in (0, T], \quad y(0) = 0. \tag{3.6}$$

By linearity of the differential equation we can conclude that $\mathcal{S}u + \hat{y}$ solves that state equation (3.1a).

Definition 3.3 (Reduced problem of $(\mathbf{P}_{\mathrm{ODE}})$)
The reduced cost functional $\hat{\mathcal{J}} : U \to \mathbb{R}$ of $(\mathbf{P}_{\mathrm{ODE}})$ is given by

$$\begin{aligned} \hat{\mathcal{J}}(u) &= \frac{1}{2} \int_0^T \|C(t)(\mathcal{S}u(t)) + C(t)\hat{y}(t) - z_{\mathsf{d}}(t)\|_Q^2 \, \mathrm{d}t + \frac{\sigma}{2} \int_0^T \|u(t)\|_R^2 \, \mathrm{d}t \\ &= \frac{1}{2} \int_0^T \|(\mathcal{G}u)(t) - \hat{z}_{\mathsf{d}}(t)\|_Q^2 \, \mathrm{d}t + \frac{\sigma}{2} \int_0^T \|u(t)\|_R^2 \, \mathrm{d}t, \end{aligned} \tag{3.7}$$

where $(\mathcal{G}u)(t) := C(t)(\mathcal{S}u(t))$ and $\hat{z}_d(t) = z_d(t) - C(t)\hat{y}(t)$. Consequently, the reduced problem is given by

$$\min_{u \in U} \hat{\mathcal{J}}(u). \tag{$\hat{\mathbf{P}}_{\mathrm{ODE}}$}$$

Next, we want to compute the gradient of the reduced cost functional. Therefore, we introduce the following lemmata.

Lemma 3.4 (Gâteaux derivative of the reduced cost functional)
The reduced cost function is Gâteaux-differentiable in U with Gâteaux derivative

$$\hat{\mathcal{J}}'(u) = \int_0^T \langle (\mathcal{G}^\star(\mathcal{G}u - \hat{z}_\mathsf{d}) + \sigma u)(t), \cdot \rangle_R \, \mathrm{d}t \quad \textit{for } u \in U, \tag{3.8}$$

where $\mathcal{G}^\star : L^2(0,T;\mathbb{R}^{n_z}) \to U$ represents the adjoint operator of $\mathcal{G}$. Especially, the map $u \mapsto \hat{\mathcal{J}}'(u) \in U'$ is linear and continuous.

Proof. Let $u \in U$ and $u^\delta \in U \backslash \{0\}$ arbitrary. Then for $\tau > 0$ it holds

$$\begin{aligned}
\frac{\hat{\mathcal{J}}(u+\tau u^\delta) - \hat{\mathcal{J}}(u)}{\tau} &= \frac{1}{2\tau}\left(\int_0^T \|(\mathcal{G}(u+\tau u^\delta))(t) - \hat{z}_\mathsf{d}(t)\|_Q^2 - \|(\mathcal{G}u)(t) - \hat{z}_\mathsf{d}(t)\|_Q^2 \, \mathrm{d}t \right. \\
&\quad \left. + \sigma \int_0^T \|(u+\tau u^\delta)(t)\|_R^2 - \|u(t)\|_R^2 \, \mathrm{d}t \right) \\
&= \frac{1}{2\tau}\left(\int_0^T 2\tau \langle (\mathcal{G}u)(t) - \hat{z}_\mathsf{d}(t), (\mathcal{G}u^\delta)(t) \rangle_Q + \tau^2 \|(\mathcal{G}u^\delta)(t)\|_Q^2 \, \mathrm{d}t \right. \\
&\quad \left. + \sigma \int_0^T 2\tau \langle u(t), u^\delta(t) \rangle_R + \tau^2 \|u^\delta(t)\|_R^2 \, \mathrm{d}t \right) \\
&= \frac{1}{2\tau}\left(\int_0^T 2\tau \langle (G^\star(\mathcal{G}u - \hat{z}_\mathsf{d}) + \sigma u)(t), u^\delta(t) \rangle_R + \tau^2 \|(\mathcal{G}u^\delta)(t)\|_Q^2 \right. \\
&\quad \left. + \sigma\tau^2 \|u^\delta(t)\|_R^2 \, \mathrm{d}t \right).
\end{aligned}$$

In the limit for $\tau \to 0$ we get

$$\lim_{\tau \to 0} \frac{\hat{\mathcal{J}}(u+\tau u^\delta) - \hat{\mathcal{J}}(u)}{\tau} = \int_0^T \langle (\mathcal{G}^\star(\mathcal{G}u - \hat{z}_\mathsf{d}) + \sigma u)(t), u^\delta(t) \rangle_R \, \mathrm{d}t.$$

Consequently, we define the operator $\mathcal{A}_u : U \to \mathbb{R}$ as

$$\mathcal{A}_u(u^\delta) := \int_0^T \langle (\mathcal{G}^\star(\mathcal{G}u - \hat{z}_\mathsf{d}) + \sigma u)(t), u^\delta(t) \rangle_R \, \mathrm{d}t.$$

The operator $\mathcal{A}_u$ is linear, because of the linearity of the inner product $\langle \cdot, \cdot \rangle_R$ and the integral. Indeed $\mathcal{A}_u$ is also bounded, since it holds

$$\begin{aligned}
|\mathcal{A}_u(u^\delta)| &\le \int_0^T \|\mathcal{G}^\star(\mathcal{G}u - \hat{z}_\mathsf{d}) + \sigma u)(t)\|_R \|u^\delta(t)\|_R \, \mathrm{d}t \\
&\le \left(\int_0^T \|\mathcal{G}^\star(\mathcal{G}u - \hat{z}_\mathsf{d}) + \sigma u)(t)\|_R^2 \, \mathrm{d}t \right)^{1/2} \left(\int_0^T \|u^\delta(t)\|_R^2 \, \mathrm{d}t \right)^{1/2} \\
&\le C \left(\int_0^T \|u^\delta(t)\|_R^2 \, \mathrm{d}t \right)^{1/2} < \infty.
\end{aligned}$$

Therefore, (3.8) is the Gâteaux derivative of $\hat{\mathcal{J}}$ at $u \in U$. The linearity and continuity of the Gâteaux derivative follows directly by the linearity and continuity of the operator $\mathcal{G}$ (see proof of Theorem 3.12) and the linearity, continuity of the inner product. □

Definition 3.5 (Gradient of the reduced cost functional)
For the control space U and the symmetric positive definite matrix $R \in \mathbb{R}^{n_u \times n_u}$ we define the weighted inner product by

$$\langle u, \tilde{u} \rangle_U = \int_0^T \langle u(t), \tilde{u}(t) \rangle_R \, \mathrm{d}t.$$

In the following we will introduce the gradient of $\hat{\mathcal{J}}$ as Riesz representation of the Gâteaux derivative with respect to $\langle \cdot, \cdot \rangle_U$, i.e.

$$\langle \hat{\mathcal{J}}'(u), \tilde{u} \rangle_{U',U} = \langle \nabla \hat{\mathcal{J}}(u), \tilde{u} \rangle_U \quad \text{for all } \tilde{u} \in U.$$

Therefore, we get

$$\nabla \hat{\mathcal{J}}(u) = \mathcal{G}^\star(\mathcal{G}u - \hat{z}_\mathsf{d}) + \sigma u \in U.$$

3.1.3 Adjoint Equation

As mentioned before a first order sufficient optimality condition is given by $\nabla \hat{\mathcal{J}}(\bar{u}) = 0 \in U$. Consequently, it is essential that we are able to compute

$$\nabla \hat{\mathcal{J}}(u) = \mathcal{G}^\star(\mathcal{G}u - \hat{z}_\mathsf{d}) + \sigma u,$$

where $\mathcal{G}^\star : L^2(0,T;\mathbb{R}^{n_z}) \to U$ is the adjoint operator. Since $\mathcal{G}$ is a solution operator of a differential equation it is in general not clear how to compute the adjoint operator $\mathcal{G}^\star$. It will be shown that we can find a representation of the terms $\mathcal{G}^\star\mathcal{G}$ and $\mathcal{G}^\star\hat{z}_\mathsf{d}$ by introducing the so-called adjoint equation.

Definition 3.6 (Adjoint equation)
For the ODE (3.1)*, the cost function $\hat{\mathcal{J}}$ and a given control $u \in U$, we call the backward initial value problem*

$$\begin{aligned} -\dot{p}(t) &= W^{-1}(A(t)^\top W p(t) + C(t)^\top Q(\hat{z}_\mathsf{d}(t) - C(t)\mathcal{S}u(t))), && \text{for } t \in [0,T), \\ p(T) &= 0 && \text{in } \mathbb{R}^{n_y} \end{aligned} \tag{3.9}$$

adjoint equation *with solution $p \in Y$, where $\hat{z}_d(t) = z_d(t) - C(t)\hat{y}(t)$ holds.*

Lemma 3.7 (Existence and uniqueness of the adjoint equation)
For every $u \in U$ the adjoint equation (3.9) *has a unique solution $p \in Y$. Moreover, it holds*

$$\|p\|_Y \leq \tilde{c} \|C^\top Q(\hat{z}_\mathsf{d} - C(\cdot)\mathcal{S}u)\|_{L^2(0,T;\mathbb{R}^{n_y})}$$

for a constant $\tilde{c} > 0$.

Proof. We can transform the adjoint equation (3.9) into an initial value problem. Therefore, let $q(t) := p(T-t)$ for all $t \in [0,T]$. Then the function p solves (3.9) if and only if q solves the initial value

$$\begin{aligned} -\dot{q}(t) &= W^{-1}(A(T-t)^\top W q(t) + C(T-t)^\top Q(\hat{z}_\mathsf{d}(T-t) - C(T-t)\mathcal{S}u(T-t))), \\ q(0) &= 0 \end{aligned}$$

for $t \in (0,T]$. Since the right hand side is in $L^2(0,T;\mathbb{R}^{n_y})$ Theorem 2.11 implies directly the existence of a unique solution for every $u \in U$ with

$$\|q\|_Y \leq \tilde{c}\|C^\top Q(\hat{z}_\mathsf{d} - C(\cdot)\mathcal{S}u)\|_{L^2(0,T;\mathbb{R}^{n_y})}$$

Hence, there also exists a unique solution of (3.9), which satisfies the same inequality. □

Analog to the state equation let $\hat{p} \in Y$ be the solution of the inhomogeneous adjoint equation

$$-\dot{\hat{p}}(t) = W^{-1}(A(t)^\top W\hat{p}(t) + C(t)^\top Q\hat{z}_\mathsf{d}(t)) \quad \text{for } t \in [0,T), \quad p(T) = 0. \tag{3.10}$$

Now, let $\mathcal{A} : U \to Y$ be the linear solution operator of the adjoint equation such that $p = \mathcal{A}u \in Y$ solves

$$-\dot{p}(t) = W^{-1}(A(t)^\top Wp(t) - C(t)^\top QC(t)\mathcal{S}u(t)) \quad \text{for } t \in [0,T), \quad p(T) = 0. \tag{3.11}$$

By linearity of the adjoint equation we can conclude that $\mathcal{A}u + \hat{p}$ solves that adjoint equation (3.9).
For the next two lemmata we refer to [2, Chapter 5.4].

Lemma 3.8
Let $u, \tilde{u} \in U$ be arbitrary. Define $y := \mathcal{S}u \in Y$, $\tilde{y} := \mathcal{S}\tilde{u} \in Y$ and $p := \mathcal{A}\tilde{u} \in Y$. Then it holds

$$\int_0^T \langle B(t)u(t), p(t)\rangle_W \,\mathrm{d}t = -\int_0^T \langle C(t)^\top QC(t)\tilde{y}(t), y(t)\rangle_{\mathbb{R}^{n_y}} \,\mathrm{d}t.$$

Proof. Since $y := \mathcal{S}u$ is the solution of (3.5) with right hand side Bu it holds

$$\int_0^T \langle B(t)u(t), p(t)\rangle_W \,\mathrm{d}t = \int_0^T \langle \partial_t y(t), p(t)\rangle_W - \langle A(t)y(t), p(t)\rangle_W \,\mathrm{d}t.$$

With integration by parts and the initial-, end-condition $p(T) = y(0) = 0$ we obtain

$$\begin{aligned}\int_0^T \langle \dot{y}(t), p(t)\rangle_W - \langle A(t)y(t), p(t)\rangle_W \,\mathrm{d}t &= \int_0^T -\langle y(t), \dot{p}(t)\rangle_W - \langle A(t)y(t), p(t)\rangle_W \,\mathrm{d}t \\ &\quad + \langle y(T), p(T)\rangle_W - \langle y(0), p(0)\rangle_W \\ &= \int_0^T -\langle y(t), \dot{p}(t)\rangle_W - \langle A(t)y(t), p(t)\rangle_W \,\mathrm{d}t \\ &= \int_0^T -\langle \dot{p}(t) + A(t)^\top p(t), y(t)\rangle_W \,\mathrm{d}t.\end{aligned}$$

Using that $p = \mathcal{A}\tilde{u}$ is the solution of (3.11) we obtain

$$\int_0^T -\langle \dot{p}(t) + A(t)^\top p(t), y(t)\rangle_W \,\mathrm{d}t = \int_0^T \langle -C(t)^\top QC(t)(\mathcal{S}\tilde{u}(t)), y(t)\rangle_{\mathbb{R}^{n_y}} \,\mathrm{d}t.$$

With $\tilde{y} = \mathcal{S}\tilde{u}$ the claim follows. □

Lemma 3.9
It holds $\mathcal{G}^\star\mathcal{G} = -R^{-1}B(\cdot)^\top W\mathcal{A} \in L(U)$ *as well as* $\mathcal{G}^\star\hat{z}_\mathsf{d} = R^{-1}B(\cdot)^\top W\hat{p} \in U$.

Proof. Let $u, \tilde{u} \in U$ be arbitrary. Define $y := \mathcal{S}u \in Y$ and $\tilde{y} := \mathcal{S}\tilde{u} \in Y$. Using Lemma 3.8 leads to

$$\begin{aligned}
\langle \mathcal{G}^\star\mathcal{G}\tilde{u}, u\rangle_U &= \langle \mathcal{G}\tilde{u}, \mathcal{G}u\rangle_{L^2(0,T;\mathbb{R}^{n_z})} = \int_0^T \langle C(t)(\mathcal{S}\tilde{u}(t)), C(t)(\mathcal{S}u(t))\rangle_Q \,\mathrm{d}t \\
&= \int_0^T \langle C(t)^\top QC(t)\tilde{y}(t), y(t)\rangle_{\mathbb{R}^{n_y}} \,\mathrm{d}t \\
&= -\int_0^T \langle B(t)u(t), p(t)\rangle_W \,\mathrm{d}t \\
&= -\int_0^T \langle u(t), R^{-1}B(t)^\top Wp(t)\rangle_R \,\mathrm{d}t \\
&= -\langle u, R^{-1}B^\top Wp\rangle_U \\
&= -\langle R^{-1}B^\top W\mathcal{A}\tilde{u}, u\rangle_U.
\end{aligned}$$

For the second claim let $u \in U$ be arbitrary. Firstly, using equation (3.10). Then integration by parts with $\hat{p}(T) = y(0) = 0$ and finally the state equation (3.5) yields

$$\begin{aligned}
\langle \mathcal{G}^\star\hat{z}_\mathsf{d}, u\rangle_U &= \langle \hat{z}_\mathsf{d}, \mathcal{G}u\rangle_{L^2(0,T;\mathbb{R}^{n_z})} = \int_0^T \langle \hat{z}_\mathsf{d}(t), C(t)y(t)\rangle_Q \,\mathrm{d}t \\
&= \int_0^T \langle C(t)^\top Q\hat{z}_\mathsf{d}(t), y(t)\rangle_{\mathbb{R}^{n_y}} \,\mathrm{d}t \\
&= \int_0^T \langle -\dot{\hat{p}}(t) - A(t)^\top\hat{p}(t), y(t)\rangle_W \,\mathrm{d}t \\
&= \int_0^T \langle \dot{y}(t), \hat{p}(t)\rangle_W - \langle A(t)y(t), \hat{p}(t)\rangle_W \,\mathrm{d}t \\
&= \int_0^T \langle B(t)u(t), \hat{p}(t)\rangle_W \,\mathrm{d}t \\
&= \langle u, R^{-1}B^\top Wp\rangle_U.
\end{aligned}$$

and thus the second claim follows. □

Corollary 3.10 (Gradient of the reduced cost functional)
The gradient of the reduced cost functional has the following representation

$$\nabla\hat{\mathcal{J}}(u)(t) = \sigma u(t) - R^{-1}B(t)^\top W(\mathcal{A}u(t) + \hat{p}(t)) \in \mathbb{R}^{n_u}. \tag{3.12}$$

Proof. Lemma 3.4 yields

$$\nabla\hat{\mathcal{J}}(u) = \mathcal{G}^\star(\mathcal{G}u - \hat{z}_\mathsf{d}) + \sigma u \in U.$$

Using the representation of Lemma 3.9 the claim follows directly. □

3.1.4 Optimality System

As we see in the previous section, to compute the gradient of the reduced cost functional we firstly need to solve the state equation (3.1a), afterwards solve the adjoint equation

(3.9) and finally we can compute the reduced gradient $\nabla\hat{\mathcal{J}}$. Together with the first order condition $\nabla\hat{\mathcal{J}}(u) = 0 \in U$ this results in the following optimality system

$$\begin{aligned}
\dot{y}(t) &= A(t)y(t) + B(t)u(t) + f(t), && t \in (0,T], \quad y(0) = y_\circ, && \text{(3.13a)}\\
-\dot{p}(t) &= W^{-1}\Big(A(t)^\top W p(t) + C(t)^\top Q(z_\mathsf{d}(t) - C(t)y(t))\Big), && t \in [0,T), \quad p(T) = 0, && \text{(3.13b)}\\
u(t) &= \frac{1}{\sigma} R^{-1} B(t)^\top W p(t), && t \in (0,T). && \text{(3.13c)}
\end{aligned}$$

Equation (3.13a) represents the state equation, equation (3.13b) the adjoint and (3.13c) is the sufficient first order optimality condition.

Remark 3.11
If we use stationarity conditions of the Lagrangian, i.e. $\nabla\mathcal{L}(y,u,p) = 0$ *one find out*

- $\nabla_y\mathcal{L}(y,u,p) = 0$ *corresponds to the adjoint equation* (3.13b).
- $\nabla_u\mathcal{L}(y,u,p) = 0$ *corresponds to the stationarity condition of the reduced cost functional, i.e. equation* (3.13c).
- $\nabla_p\mathcal{L}(y,u,p) = 0$ *corresponds to the state equation* (3.13a).

We see that the Lagrangian approach is equivalent to our approach and can be found in detail for example in [4].

Note, that we can solve ($\mathbf{P}_{\text{ODE}}$) by using a gradient scheme or solving the optimality system (3.13) directly. In this thesis we focus on the second approach. To do so, we want to simplify the optimality system and transform it into a kind of dynamical system. Inserting (3.13c) in (3.13a) yields to the following optimality system.

$$\begin{aligned}
\dot{y}(t) &= A(t)y(t) + \frac{1}{\sigma} B(t)R^{-1}B(t)^\top W p(t) + f(t), && t \in (0,T], \quad y(0) = y_\circ, && \text{(3.14a)}\\
-\dot{p}(t) &= W^{-1}\Big(A(t)^\top W p(t) + C(t)^\top Q(z_\mathsf{d}(t) - C(t)y(t))\Big), && t \in [0,T), \quad p(T) = 0. && \text{(3.14b)}
\end{aligned}$$

System (3.14) can be rewritten to get the form of a dynamical system, although formally it is not. We introduce the transformed variable $q(t) := p(T-t)$. For brevity, we set

$$\tilde{z}_\mathsf{d}(t) = z_\mathsf{d}(T-t) \in \mathbb{R}^{n_z}, \quad \tilde{A}(t) = A(T-t) \in \mathbb{R}^{n_y \times n_y}, \quad \tilde{C}(t) = C(T-t) \in \mathbb{R}^{n_z \times n_y}$$

for $t \in [0,T]$. This allows us to express (3.14) as

$$\begin{aligned}
\dot{y}(t) &= A(t)y(t) + \frac{1}{\sigma} B(t)R^{-1}B(t)^\top W q(T-t) + f(t), && t \in (0,T], \quad y(0) = y_\circ,\\
\dot{q}(t) &= W^{-1}\Big(\tilde{A}(t)^\top W q(t) + \tilde{C}(t)^\top Q\Big(\tilde{z}_\mathsf{d}(t) - \tilde{C}(t)y(T-t)\Big)\Big), && t \in [0,T), \quad q(0) = 0.
\end{aligned} \tag{3.15}$$

For $t \in (0,T)$ let

$$x(t) = \begin{bmatrix} y(t) \\ q(t) \end{bmatrix} \in \mathbb{R}^{2n_y}, \ \mathcal{F}(t) = \begin{bmatrix} f(t) \\ \tilde{C}(t)^\top Q \tilde{z}_\mathsf{d}(t) \end{bmatrix} \in \mathbb{R}^{2n_y}, \ x_\circ = \begin{bmatrix} y_\circ \\ 0 \end{bmatrix} \in \mathbb{R}^{2n_y}.$$

Then, (3.15) can be written as

$$\dot{x}(t) = \mathcal{A}(t)x(t) + \tilde{\mathcal{A}}(t)x(T-t) + \mathcal{F}(t) \text{ for } t \in (0,T), \quad x(0) = x_\circ \tag{3.16}$$

with the two $(2n_y) \times (2n_y)$-matrices

$$\mathcal{A}(t) = \begin{bmatrix} A(t) & 0 \\ 0 & W^{-1}\tilde{A}(t)^\top W \end{bmatrix}, \; \tilde{\mathcal{A}}(t) = \begin{bmatrix} 0 & \frac{1}{\sigma}B(t)R^{-1}B(t)^\top W \\ -W^{-1}\tilde{C}(t)^\top Q\tilde{C}(t) & 0 \end{bmatrix}.$$

Although (3.16) has not a canonical form, it can be seen as a dynamical system, because the solution x evolves from 0 to the final time T. In conclusion, solving ($\mathbf{P}_{\text{ODE}}$) is equivalent to just solving a dynamical system. Thus, we can apply a model-order reduction scheme on the coupled system (3.16) (see Chapter 4).

3.2 The PDE Extension

We consider an optimal control problem in space and time. Therefore, let $\Omega \subset \mathbb{R}^n$ be an open and bounded Lipschitz domain, $T > 0$. We set

$$\Omega_T := (0,T) \times \Omega \quad \text{and} \quad \Sigma_T := (0,T) \times \partial\Omega.$$

Furthermore, we define the function spaces $V := H^1(\Omega)$, $H := L^2(\Omega)$, the control space $\mathcal{U} := L^2(0,T;H)$, the state space $\mathcal{Y} := W(0,T)$ and the measurement space by $\mathcal{H} := L^2(0,T;H)$. For $\sigma > 0$ we define the cost functional

$$\mathcal{J}(\mathfrak{y},\mathfrak{u}) = \frac{1}{2}\int_0^T \|\mathcal{C}(t)\mathfrak{y}(t) - \mathfrak{z}_\mathfrak{d}(t)\|_H^2 + \sigma\,\|\mathfrak{u}(t)\|_H^2 \,\mathrm{d}t \tag{3.17}$$

with $\mathcal{C} \in L^2(0,T;\mathcal{L}(V,H))$ and $\mathfrak{z}_\mathfrak{d} \in \mathcal{H}$. For the dynamical system we are interested in a general evolution equation

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\langle \mathfrak{y}(t), \varphi\rangle_H + a(t;\mathfrak{y}(t),\varphi) &= \langle \mathfrak{f}(t) + \mathcal{B}(t)\mathfrak{u}(t), \varphi\rangle_{V',V} && \text{for all } \varphi \in V \text{ and } t \in (0,T], \\ \mathfrak{y}(0) &= \mathfrak{y}_\circ && \text{in } H, \end{aligned} \tag{3.18}$$

where $\mathfrak{f} \in L^2(0,T;V')$, $\mathcal{B} \in L^2(0,T;\mathcal{L}(H,V'))$ holds and $a(t;\cdot,\cdot) : V \times V \to \mathbb{R}$ is a time-dependent bilinear form, which satisfies

$$\left|a(t;\varphi,\phi)\right| \le \gamma\,\|\varphi\|_V\|\phi\|_V \qquad \text{for all } \varphi,\phi \in V, \tag{3.19a}$$

$$a(t;\varphi,\varphi) \ge \eta_1\,\|\varphi\|_V^2 - \eta_2\,\|\varphi\|_H^2 \qquad \text{for all } \varphi \in V, \tag{3.19b}$$

for almost all $t \in [0,T]$ and t-independent constants $\gamma > 0$, $\eta_1 > 0$ and $\eta_2 \ge 0$.
The optimal control problem reads as follows

$$\min \mathcal{J}(\mathfrak{y},\mathfrak{u}) \quad \text{subject to} \quad (\mathfrak{y},\mathfrak{u}) \text{ satisfies (3.18).} \tag{$\mathbf{P}_{\text{PDE}}$}$$

Theorem 3.12 (Existence and uniqueness of (3.18))
For every $\mathfrak{u} \in \mathcal{U}$ the evolution equation (3.18) has a unique solution $\mathfrak{y} = \mathfrak{y}(\mathfrak{u}) \in \mathcal{Y}$, which satisfies

$$\|\mathfrak{y}\|_{\mathcal{Y}} \leq \tilde{c}\Big(\|\mathfrak{y}_\circ\|_H + \|\mathfrak{f} + \mathcal{B}(\cdot)\mathfrak{u}\|_{L^2(0,T;V')}\Big).$$

Proof. The proof follows directly from Theorem 2.11. □

Theorem 3.13 (Existence and uniqueness of $(\mathbf{P}_{\mathrm{PDE}})$)
The optimal control problem $(\mathbf{P}_{\mathrm{PDE}})$ has a unique solution $(\bar{\mathfrak{y}}, \bar{\mathfrak{u}}) \in \mathcal{Y} \times \mathcal{U}$.

Proof. The proof follows directly from Theorem 2.16. □

3.2.1 First-Discretize-Then-Optimize

The First-Discretize-Then-Optimize approach works as follows: The state equation (3.18) is discretized a-priori in space by a finite dimensional space, resulting in a semi-discrete optimal control problem, such as $(\mathbf{P}_{\mathrm{ODE}})$. As a result, we obtain an optimality system of the form (3.16). For the spatial discretization, let $\mathcal{T}_h$ be a triangulation of the spatial domain Ω, i.e. $\bar{\Omega} = \bigcup_{T \in \mathcal{T}_h} \bar{\mathcal{T}}$. We define the following finite element (FE) spaces by

$$V^{\mathsf{fe}} := \{v_h \in C^0(\bar{\Omega}) : v_h|_T \in P_k(\mathcal{T}) \text{ for all } \mathcal{T} \in \mathcal{T}_h\} = \operatorname{span}(\varphi_1, ..., \varphi_{n_y}) \subset V, \tag{3.20a}$$

$$H^{\mathsf{fe}} := \{u_h \in L^2(\bar{\Omega}) : v|_T \in P_k(\mathcal{T}) \text{ for all } \mathcal{T} \in \mathcal{T}_h\} = \operatorname{span}(\phi_1, ..., \phi_{n_u}) \subset H, \tag{3.20b}$$

where P_k is a polynomial space of degree $k \in \mathbb{N}$ and the FE ansatz functions $\varphi_1, ..., \varphi_{n_y}$ and $\phi_1, ..., \phi_{n_u}$ build a basis of V^{fe} or H^{fe}, respectively. For $(t, \boldsymbol{x}) \in \Omega_T$ we approximate the state and control as

$$\mathfrak{y}(t, \boldsymbol{x}) \approx \mathfrak{y}^{\mathsf{fe}}(t, \boldsymbol{x}) = \sum_{i=1}^{n_y} y_i(t)\varphi_i(\boldsymbol{x}), \quad \mathfrak{u}(t, \boldsymbol{x}) \approx \mathfrak{u}^{\mathsf{fe}}(t, \boldsymbol{x}) = \sum_{i=1}^{n_u} u_i(t)\phi_i(\boldsymbol{x}), \tag{3.21}$$

respectively. Moreover, let

$$\mathfrak{z}_{\mathsf{d}}(t, \boldsymbol{x}) \approx \sum_{i=1}^{n_y} z_{\mathsf{d}i}(t)\mathcal{C}(t)\varphi_i(\boldsymbol{x}), \ \mathfrak{f}(t, \boldsymbol{x}) \approx \sum_{i=1}^{n_y} f_i(t)\varphi_i(\boldsymbol{x}), \ \mathfrak{y}_\circ(\boldsymbol{x}) \approx \sum_{i=1}^{n_y} y_{\circ i}\varphi_i(\boldsymbol{x}) \tag{3.22}$$

for $(t, \boldsymbol{x}) \in \Omega_T$. We assemble the vectors $y(t) := (y_1(t), ..., y_{n_y}(t))^\top \in \mathbb{R}^{n_y}$, $u(t) := (u_1(t), ..., u_{n_u}(t))^\top \in \mathbb{R}^{n_u}$, $f(t) := (f_1(t), ..., f_{n_y}(t))^\top \in \mathbb{R}^{n_y}$, $z_{\mathsf{d}}(t) := (z_{\mathsf{d}1}(t), ..., z_{\mathsf{d}n_y}(t))^\top \in \mathbb{R}^{n_y}$ and $y_\circ := (y_{\circ 1}, ..., y_{\circ n_y})^\top \in \mathbb{R}^{n_y}$. Finally, we introduce the finite element matrices

$$M \in \mathbb{R}^{n_y \times n_y} \quad M_{i,j} = \int_\Omega \varphi_i(\boldsymbol{x})\varphi_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \qquad \text{for } 1 \leq i, j \leq n_y, \tag{3.23a}$$

$$R \in \mathbb{R}^{n_u \times n_u} \quad R_{i,j} = \int_\Omega \phi_i(\boldsymbol{x})\phi_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \qquad \text{for } 1 \leq i, j \leq n_u, \tag{3.23b}$$

$$A(t) \in \mathbb{R}^{n_y \times n_y} \quad A(t)_{i,j} = a(t; \varphi_i, \varphi_j) \qquad \text{for } 1 \leq i, j \leq n_y, \tag{3.23c}$$

$$B(t) \in \mathbb{R}^{n_y \times n_u} \quad B(t)_{i,j} = \langle \mathcal{B}(t)\phi_j, \varphi_i \rangle_{V',V} \qquad \text{for } 1 \leq i \leq n_y, 1 \leq j \leq n_u, \tag{3.23d}$$

$$C(t) \in \mathbb{R}^{n_y \times n_y} \quad C(t)_{i,j} = \langle \mathcal{C}(t)\varphi_i, \mathcal{C}(t)\varphi_j \rangle_H \qquad \text{for } 1 \leq i, j \leq n_y. \tag{3.23e}$$

Lemma 3.14
The matrices $M \in \mathbb{R}^{n_y \times n_y}$, $R \in \mathbb{R}^{n_u \times n_u}$ and $C(t) \in \mathbb{R}^{n_y \times n_y}$ are symmetric and postive definite for all $t \in [0,T]$. Moreover, $A(t) \in \mathbb{R}^{n_y \times n_y}$ is coercive for all $t \in [0,T]$, i.e.

$$\langle A(t)y, y\rangle_{\mathbb{R}^{n_y}} \geq \eta_1 \|y_h\|_V^2 - \eta_2 \|y_h\|_H^2 \quad \text{for all } y \in \mathbb{R}^{n_y},$$

where $y_h = \sum_{i=1}^{n_y} y_i\varphi_i$ and $\eta_1 > 0$, $\eta_2 \geq 0$ are the respective coercivity constant of the bilinear form.

Proof. Let $y = (y_i) \in \mathbb{R}^{n_y}$ be arbitrary and we set $y_h = \sum_{i=1}^{n_y} y_i\varphi_i \in V^{\mathsf{fe}}$.

(i) The symmetry of the matrices M, R and $C(t)$ is clear by definition. Firstly, we investigate the mass matrix M. It holds

$$y^\top My = \sum_{i=1}^{n_y}\sum_{j=1}^{n_y} y_i M_{ij} y_j = \sum_{i=1}^{n_y}\sum_{j=1}^{n_y} y_i y_j \langle\varphi_i, \varphi_j\rangle_H = \left\langle \sum_{i=1}^{n_y} y_i\varphi_i, \sum_{j=1}^{n_y} y_j\varphi_j \right\rangle_H = \|y_h\|_H^2 \geq 0.$$

Recall that the φ_i's are linearly independent, it holds $y^\top My > 0$ if and only if $y \neq 0 \in \mathbb{R}^{n_y}$, which implies the positive definiteness. The positive definiteness of the matrices R and $C(t)$ follows analogously.

(ii) Secondly, let $t \in [0,T]$, applying (3.19b) yields

$$\begin{aligned} y^\top A(t)y = \sum_{i=1}^{n_y}\sum_{j=1}^{n_y} y_i A(t)_{ij} y_j = \sum_{i=1}^{n_y}\sum_{j=1}^{n_y} y_i y_j a(t;\varphi_i,\varphi_j) &= a\left(t; \sum_{i=1}^{n_y} y_i\varphi_i, \sum_{j=1}^{n_y} y_j\varphi_j\right) \\ &\geq \eta_1 \|y_h\|_V^2 - \eta_2 \|y_h\|_H^2, \end{aligned}$$

where $\eta_1 > 0$ and $\eta_2 \geq 0$.

□

For the semi-discrete state equation we obtain

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\langle \mathfrak{y}^{\mathsf{fe}}(t), \varphi\rangle_H + a(t; \mathfrak{y}^{\mathsf{fe}}(t), \varphi) &= \langle \mathfrak{f}(t) + \mathcal{B}(t)\mathfrak{u}^{\mathsf{fe}}(t), \varphi\rangle_{V',V} \quad \text{for all } \varphi \in V^{\mathsf{fe}} \text{ and } t \in (0,T], \\ \mathfrak{y}^{\mathsf{fe}}(0) &= \mathfrak{y}_\circ \qquad\qquad \text{in } H^{\mathsf{fe}}, \end{aligned} \tag{3.24}$$

which is equivalent to

$$\begin{aligned} M\dot{y}(t) &= -A(t)y(t) + B(t)u(t) + Mf(t), \quad t \in (0,T], \\ y(0) &= y_\circ \qquad\qquad \text{in } \mathbb{R}^{n_y}. \end{aligned} \tag{3.25}$$

For the approximated finite element functions it holds

$$\begin{aligned} \mathcal{J}(\mathfrak{y}^{\mathsf{fe}}, \mathfrak{u}^{\mathsf{fe}}) &= \frac{1}{2}\int_0^T \|\mathcal{C}(t)\mathfrak{y}^{\mathsf{fe}}(t) - \mathfrak{z}_{\mathsf{d}}(t)\|_H^2 + \sigma\|\mathfrak{u}^{\mathsf{fe}}(t)\|_H^2 \,\mathrm{d}t \\ &= \frac{1}{2}\int_0^T \|y(t) - z_{\mathsf{d}}(t)\|_{C(t)}^2 + \sigma\|u(t)\|_R^2 \,\mathrm{d}t =: \mathcal{J}^{\mathsf{fe}}(y,u), \end{aligned}$$

where we used

$$\|\mathcal{C}(t)\mathfrak{y}^{\mathsf{fe}}(t)\|_H^2 = \left\langle \sum_{j=1}^{n_y} y_j(t)\mathcal{C}(t)\varphi_j, \sum_{i=1}^{n_y} y_i(t)\mathcal{C}(t)\varphi_i \right\rangle_H = \sum_{j=1}^{n_y} y_j(t) \left(\left\langle \mathcal{C}(t)\varphi_j, \sum_{i=1}^{n_y} \mathcal{C}(t) y_i(t)\varphi_i \right\rangle_H \right)$$
$$= \sum_{j=1}^{n_y} y_j(t)(C(t)y(t))_j = y(t)^\top C(t)y(t) = \|y(t)\|_{C(t)}^2.$$

Consequently, problem $(\mathbf{P}_{\mathrm{PDE}})$ is approximated by the following FE optimization problem

$$\begin{aligned} &\min \mathcal{J}^{\mathsf{fe}}(y,u) = \frac{1}{2}\int_0^T \|y(t) - z_{\mathsf{d}}(t)\|_{C(t)}^2 + \sigma\|u(t)\|_R^2 \,\mathrm{d}t \\ &\text{s.t. } u \in \mathcal{U}^{\mathsf{fe}} \text{ and } y \in \mathcal{Y}^{\mathsf{fe}} \text{ solves the initial value problem} \\ &\quad M\dot{y}(t) \;\; = -A(t)y(t) + B(t)u(t) + Mf(t) \text{ for } t \in (0,T], \quad y(0) = y_\circ, \end{aligned} \tag{$\mathbf{P}_{\mathrm{PDE}}^{\mathsf{fe}}$}$$

where $\mathcal{Y}^{\mathsf{fe}} = H^1(0,T;\mathbb{R}^{n_y})$ and $\mathcal{U}^{\mathsf{fe}} = L^2(0,T;\mathbb{R}^{n_u})$. Notice that in this semidiscrete setting, problem $(\mathbf{P}_{\mathrm{PDE}}^{\mathsf{fe}})$ is equivalent to problem $(\mathbf{P}_{\mathrm{ODE}})$, which we have studied in Section 3.1. Therefore, sufficient first order optimality conditions for $(\mathbf{P}_{\mathrm{ODE}})$ can be directly utilized for $(\mathbf{P}_{\mathrm{PDE}}^{\mathsf{fe}})$. Therefore, we obtain a dynamical system of the form (3.16) to solve the optimal control problem $(\mathbf{P}_{\mathrm{PDE}}^{\mathsf{fe}})$, i.e.

$$\dot{x}(t) = \mathcal{A}(t)x(t) + \tilde{\mathcal{A}}(t)x(T-t) + \mathcal{F}(t) \text{ for } t \in (0,T), \quad x(0) = x_\circ \tag{3.26}$$

with

$$x(t) = \begin{bmatrix} y(t) \\ p(T-t) \end{bmatrix} \in \mathbb{R}^{2n_y}, \quad \mathcal{F}(t) = \begin{bmatrix} f(t) \\ \tilde{C}(t)\tilde{z}_{\mathsf{d}}(t) \end{bmatrix} \in \mathbb{R}^{2n_y}, \quad x_\circ = \begin{bmatrix} y_\circ \\ 0 \end{bmatrix} \in \mathbb{R}^{2n_y}.$$

and the two $(2n_y)\times(2n_y)$-matrices

$$\mathcal{A}(t) = \begin{bmatrix} -M^{-1}A(t) & 0 \\ 0 & -M^{-1}\tilde{A}(t)^\top \end{bmatrix}, \; \tilde{\mathcal{A}}(t) = \begin{bmatrix} 0 & \frac{1}{\sigma}M^{-1}B(t)R^{-1}B(t)^\top \\ -M^{-1}\tilde{C}(t) & 0 \end{bmatrix}.$$

Remind, the tilde operators are the time-shifted operators, i.e. $\tilde{A}(t) = A(T-t)$. In conclusion, we can state if we discretize the problem $(\mathbf{P}_{\mathrm{PDE}})$ a-priorily in space using finite elements, we end up in the ODE case and finally need to solve a dynamical system with initial condition.

3.2.2 First-Optimize-Then-Discretize

The starting point for this approach is to derive the sufficient first-order optimality system for problem $(\mathbf{P}_{\mathrm{PDE}})$. Next, we discretize the optimality system in space using finite elements. We do this because later on, we will see that the First-Optimize-Then-Discretize gives us more degrees of freedom when computing a discrete solution. Additionally, we have more flexibility when dealing with numerical error analysis. It should be noted that this approach is not completely equivalent to the approach First-Discretize-Then-Optimize" (see Section 3.2.1).

Reduced Problem

Analog to Section 3.1 we derive the first order optimality system for the reduced problem. Therefore, let $\hat{\mathfrak{y}} \in Y$ the inhomogeneous part of the state equation, i.e.

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\left\langle \hat{\mathfrak{y}}(t), \varphi \right\rangle_H + a(t; \hat{\mathfrak{y}}(t), \varphi) &= \left\langle \mathfrak{f}(t), \varphi \right\rangle_{V',V} \quad \text{for all } \varphi \in V \text{ and } t \in (0,T], \\ \hat{\mathfrak{y}}(0) &= \mathfrak{y}_\circ \qquad\qquad \text{in } H. \end{aligned} \tag{3.27}$$

Moreover, we define for every control $\mathfrak{u} \in \mathcal{U}$ the solution operator $\mathcal{S} : \mathcal{U} \to \mathcal{Y}$ by $\mathfrak{y} = \mathcal{S}\mathfrak{u}$, where $\mathfrak{y} \in \mathcal{Y}$ is the solution of

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\left\langle \mathfrak{y}(t), \varphi \right\rangle_H + a(t; \mathfrak{y}(t), \varphi) &= \left\langle \mathcal{B}(t)u(t), \varphi \right\rangle_{V',V} \quad \text{for all } \varphi \in V \text{ and } t \in (0,T], \\ \mathfrak{y}(0) &= 0 \qquad\qquad \text{in } H. \end{aligned} \tag{3.28}$$

From the linearity of the differential equation we conclude that $\mathcal{S}\mathfrak{u} + \hat{\mathfrak{y}} \in Y$ solves (3.18).

Definition 3.15 (Reduced problem of $(\mathbf{P}_{\mathrm{PDE}})$)
The reduced cost functional $\hat{\mathcal{J}} : \mathcal{U} \to \mathbb{R}$ of $(\mathbf{P}_{\mathrm{PDE}})$ is given by

$$\begin{aligned} \hat{\mathcal{J}}(\mathfrak{u}) &= \frac{1}{2}\int_0^T \left\| \mathcal{C}(t)(\mathcal{S}\mathfrak{u}(t)) + \mathcal{C}(t)\hat{\mathfrak{y}}(t) - \mathfrak{z}_\mathsf{d}(t) \right\|_H^2 \mathrm{d}t + \frac{\sigma}{2}\int_0^T \left\| \mathfrak{u}(t) \right\|_H^2 \mathrm{d}t \\ &= \frac{1}{2}\int_0^T \left\| (\mathcal{G}\mathfrak{u})(t) - \hat{\mathfrak{z}}_\mathsf{d}(t) \right\|_H^2 \mathrm{d}t + \frac{\sigma}{2}\int_0^T \left\| \mathfrak{u}(t) \right\|_H^2 \mathrm{d}t, \end{aligned} \tag{3.29}$$

where $\mathcal{G} : \mathcal{U} \to \mathcal{H}$ with $(\mathcal{G}\mathfrak{u})(t) := \mathcal{C}(t)(\mathcal{S}\mathfrak{u}(t))$ and $\hat{\mathfrak{z}}_d(t) = \mathfrak{z}_d(t) - \mathcal{C}(t)\hat{\mathfrak{y}}(t)$. Consequently, the reduced problem is given by

$$\min_{\mathfrak{u}\in\mathcal{U}} \hat{\mathcal{J}}(\mathfrak{u}). \tag{$\hat{\mathbf{P}}_{\mathrm{PDE}}$}$$

Analog to Section 3.1.2 the gradient of the reduced cost functional is given by

$$\nabla\hat{\mathcal{J}}(\mathfrak{u}) = \mathcal{G}^\star(\mathcal{G}\mathfrak{u} - \hat{\mathfrak{z}}_\mathsf{d}) + \sigma\mathfrak{u} \in U, \tag{3.30}$$

where $\mathcal{G}^\star : \mathcal{H} \to \mathcal{U}$ is the adjoint operator of $\mathcal{G}$.

Adjoint Equation

To compute the action of the adjoint operator $\mathcal{G}^\star$ we will introduce the adjoint equation.

Definition 3.16 (Adjoint equation)
For the evolution problem (3.18)*, the cost function $\hat{\mathcal{J}}$ and a given control $\mathfrak{u} \in \mathcal{U}$, we call the backward initial value problem*

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \mathfrak{p}(t), \varphi \right\rangle_H + a(t; \varphi, \mathfrak{p}(t)) &= \left\langle \hat{\mathfrak{z}}_\mathsf{d}(t) - (\mathcal{G}\mathfrak{u})(t), \mathcal{C}(t)\varphi \right\rangle_H \quad \textit{for all } \varphi \in V, t \in [0,T), \\ \mathfrak{p}(T) &= 0 \qquad\qquad \textit{in } H \end{aligned} \tag{3.31}$$

adjoint equation *with solution $\mathfrak{p} \in \mathcal{Y}$, where $\hat{\mathfrak{z}}_\mathsf{d} := \mathfrak{z}_\mathsf{d} - \mathcal{C}\hat{\mathfrak{y}}$.*

Lemma 3.17 (Existence and uniqueness of the adjoint equation)
For every $\mathfrak{u} \in \mathcal{U}$ the adjoint equation (3.9) *has a unique solution $\mathfrak{p} \in \mathcal{Y}$. Moreover, it holds*

$$\|\mathfrak{p}\|_{\mathcal{Y}} \leq \tilde{c}\|\mathcal{C}^\star(\hat{\mathfrak{z}}_\mathfrak{d} - \mathcal{G}\mathfrak{u})\|_{L^2(0,T;V')}$$

for a constant $\tilde{c} > 0$.

Proof. Transform the adjoint equation (3.31) into an initial value problem, like in Lemma 3.8. Then the claim follows directly utilizing Theorem 2.11. □

Similar to Section 3.1.3, we define the inhomogeneous part of the adjoint equation by $\hat{\mathfrak{p}} \in \mathcal{Y}$ as solution of

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\langle \mathfrak{p}(t), \varphi\rangle_H + a(t;\varphi,\mathfrak{p}(t)) &= \langle \hat{\mathfrak{z}}_\mathfrak{d}(t), \mathcal{C}(t)\varphi\rangle_H \quad \text{for all } \varphi \in V, t \in [0,T), \\ \mathfrak{p}(T) &= 0 \qquad\qquad\qquad\quad \text{in } H. \end{aligned} \tag{3.32}$$

Now, let $\mathcal{A} : \mathcal{U} \to \mathcal{Y}$ be the linear, bounded solution operator of the adjoint equation such that $\mathfrak{p} = \mathcal{A}\mathfrak{u} \in Y$ solves

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\langle \mathfrak{p}(t), \varphi\rangle_H + a(t;\varphi,\mathfrak{p}(t)) &= -\langle (\mathcal{G}\mathfrak{u})(t), \mathcal{C}(t)\varphi\rangle_H \quad \text{for all } \varphi \in V, t \in [0,T), \\ \mathfrak{p}(T) &= 0 \qquad\qquad\qquad\quad \text{in } H. \end{aligned} \tag{3.33}$$

From linearity of the adjoint equation it follows that $\hat{\mathfrak{p}} + \mathcal{A}\mathfrak{u} \in \mathcal{Y}$ solves the adjoint equation (3.31). With help of the adjoint equation we can now compute the action of the adjoint solution operator $\mathcal{G}^\star : L^2(0,T;H) \to \mathcal{U}$. The proof of the following lemmata are based on [42].

Lemma 3.18
Let $\mathfrak{u}, \tilde{\mathfrak{u}} \in U$ be arbitrary. Define $\mathfrak{y} := \mathcal{S}\mathfrak{u} \in \mathcal{Y}$, $\tilde{\mathfrak{y}} := \mathcal{S}\tilde{\mathfrak{u}} \in \mathcal{Y}$ and $\mathfrak{p} := \mathcal{A}\tilde{\mathfrak{u}} \in \mathcal{Y}$. Then it holds

$$\int_0^T \langle \mathcal{B}(t)\mathfrak{u}(t), \mathfrak{p}(t)\rangle_{V',V}\,\mathrm{d}t = -\int_0^T \langle \mathcal{C}(t)\tilde{\mathfrak{y}}(t), \mathcal{C}(t)\mathfrak{y}(t)\rangle_H\,\mathrm{d}t.$$

Proof. Since $\mathfrak{y} := \mathcal{S}\mathfrak{u}$ is the solution of (3.28) with right hand side $\mathcal{B}\mathfrak{u}$ it holds

$$\int_0^T \langle \mathcal{B}(t)\mathfrak{u}(t), \mathfrak{p}(t)\rangle_{V',V}\,\mathrm{d}t = \int_0^T \langle \partial_t\mathfrak{y}(t), \mathfrak{p}(t)\rangle_H + a(t;\mathfrak{y}(t),\mathfrak{p}(t))\,\mathrm{d}t.$$

With integration by parts and the initial-, end-condition $\mathfrak{p}(T) = \mathfrak{y}(0) = 0$ we obtain

$$\begin{aligned} \int_0^T \langle \partial_t\mathfrak{y}(t), \mathfrak{p}(t)\rangle_H + a(t;\mathfrak{y}(t),\mathfrak{p}(t))\,\mathrm{d}t &= \int_0^T -\langle \mathfrak{y}(t), \partial_t\mathfrak{p}(t)\rangle_H + a(t;\mathfrak{y}(t),\mathfrak{p}(t))\,\mathrm{d}t \\ &\quad + \langle \mathfrak{y}(T), \mathfrak{p}(T)\rangle_H - \langle \mathfrak{y}(0), \mathfrak{p}(0)\rangle_H \\ &= \int_0^T -\langle \mathfrak{y}(t), \partial_t\mathfrak{p}(t)\rangle_H + a(t;\mathfrak{y}(t),\mathfrak{p}(t))\,\mathrm{d}t. \end{aligned}$$

Using that $\mathfrak{p} = \mathcal{A}\tilde{\mathfrak{u}}$ is the solution of (3.33) we obtain

$$\int_0^T -\langle \mathfrak{y}(t), \partial_t \mathfrak{p}(t)\rangle_H + a(t; \mathfrak{y}(t), \mathfrak{p}(t))\,\mathrm{d}t = \int_0^T -\langle (\mathcal{G}\tilde{\mathfrak{u}})(t), \mathcal{C}\mathfrak{y}(t)\rangle_H\,\mathrm{d}t.$$

With $\tilde{\mathfrak{y}} = \mathcal{S}\tilde{\mathfrak{u}}$ the claim follows. □

Lemma 3.19
It holds $\mathcal{G}^\star\mathcal{G} = -\mathcal{B}(\cdot)^\star\mathcal{A} \in \mathcal{L}(\mathcal{U})$ as well as $\mathcal{G}^\star\hat{\mathfrak{z}}_\mathsf{d} = \mathcal{B}(\cdot)^\star\hat{\mathfrak{p}} \in \mathcal{U}$.

Proof. Let $\mathfrak{u}, \tilde{\mathfrak{u}} \in \mathcal{U}$ be arbitrary. Define $\mathfrak{y} := \mathcal{S}\mathfrak{u} \in \mathcal{Y}$ and $\tilde{\mathfrak{y}} := \mathcal{S}\tilde{\mathfrak{u}} \in \mathcal{Y}$. Using Lemma 3.18 leads to

$$\begin{aligned}
\langle \mathcal{G}^\star\mathcal{G}\tilde{\mathfrak{u}}, \mathfrak{u}\rangle_\mathcal{U} = \langle \mathcal{G}\tilde{\mathfrak{u}}, \mathcal{G}\mathfrak{u}\rangle_\mathcal{H} &= \int_0^T \langle \mathcal{C}(t)\mathcal{S}\tilde{\mathfrak{u}}(t), \mathcal{C}(t)\mathcal{S}\mathfrak{u}(t)\rangle_H\,\mathrm{d}t \\
&= \int_0^T \langle \mathcal{C}(t)\tilde{\mathfrak{y}}(t), \mathcal{C}(t)\mathfrak{y}(t)\rangle_H\,\mathrm{d}t \\
&= -\int_0^T \langle \mathcal{B}(t)\mathfrak{u}(t), \mathfrak{p}(t)\rangle_{V',V}\,\mathrm{d}t \\
&= -\langle \mathfrak{u}, \mathcal{B}(\cdot)^\star\mathfrak{p}\rangle_\mathcal{U} \\
&= -\langle \mathcal{B}(\cdot)^\star\mathcal{A}\tilde{\mathfrak{u}}, \mathfrak{u}\rangle_\mathcal{U}.
\end{aligned}$$

For the second claim let $\mathfrak{u} \in \mathcal{U}$ be arbitrary. Firstly, using equation (3.32). Then integration by parts with $\hat{\mathfrak{p}}(T) = \mathfrak{y}(0) = 0$ and finally the state equation (3.28) yields

$$\begin{aligned}
\langle \mathcal{G}^\star\hat{\mathfrak{z}}_\mathsf{d}, \mathfrak{u}\rangle_U = \langle \hat{\mathfrak{z}}_\mathsf{d}, \mathcal{G}\mathfrak{u}\rangle_\mathcal{H} &= \int_0^T \langle \hat{\mathfrak{z}}_\mathsf{d}(t), \mathcal{C}(t)\mathfrak{y}(t)\rangle_H\,\mathrm{d}t \\
&= \int_0^T \langle -\partial_t\hat{\mathfrak{p}}(t), \mathfrak{y}(t)\rangle_H + a(t; \mathfrak{y}(t), \hat{\mathfrak{p}}(t))\,\mathrm{d}t \\
&= \int_0^T \langle \partial_t\mathfrak{y}(t), \hat{\mathfrak{p}}(t)\rangle_H + a(t; \mathfrak{y}(t), \hat{\mathfrak{p}}(t))\,\mathrm{d}t \\
&= \int_0^T \langle \mathcal{B}(t)\mathfrak{u}(t), \hat{\mathfrak{p}}(t)\rangle_{V',V}\,\mathrm{d}t = \langle \mathcal{B}(\cdot)^\star\hat{\mathfrak{p}}, \mathfrak{u}\rangle_\mathcal{U}
\end{aligned}$$

and thus the second claim follows. □

Corollary 3.20 (Gradient of the reduced cost functional)
The gradient of the reduced cost functional has the following representation

$$\nabla\hat{\mathcal{J}}(u)(t) = \sigma\mathfrak{u}(t) - \mathcal{B}(t)^\star(\mathcal{A}\mathfrak{u}(t) + \hat{\mathfrak{p}}(t)) \in H. \tag{3.34}$$

Proof. Lemma 3.4 yields

$$\nabla\hat{\mathcal{J}}(\mathfrak{u}) = \mathcal{G}^\star(\mathcal{G}\mathfrak{u} - \hat{\mathfrak{z}}_\mathsf{d}) + \sigma\mathfrak{u} \in \mathcal{U}.$$

Using Lemma 3.19 the claim follows directly. □

Optimality System

We have derived a representation of the gradient, we can state the first order optimality system. Again, it is necessary to solve the state and the adjoint equation to get the gradient. The optimality system reads as follows

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \mathfrak{y}(t),\varphi\right\rangle_H + a(t;\mathfrak{y}(t),\varphi) = \left\langle \mathfrak{f}(t)+\mathcal{B}(t)\mathfrak{u}(t),\varphi\right\rangle_{V',V}, \qquad t\in(0,T], \tag{3.35a}$$

$$-\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \mathfrak{p}(t),\varphi\right\rangle_H + a(t;\varphi,\mathfrak{p}(t)) = \left\langle \mathfrak{z}_\mathsf{d}(t)-\mathcal{C}(t)\mathfrak{y}(t),\mathcal{C}(t)\varphi\right\rangle_H, \qquad t\in[0,T), \tag{3.35b}$$

$$\int_0^T \left\langle \mathfrak{u}(t),\phi(t)\right\rangle_H \mathrm{d}t = \int_0^T \left\langle \frac{1}{\sigma}\mathcal{B}(t)^\star\mathfrak{p}(t),\phi(t)\right\rangle_H \mathrm{d}t, \qquad t\in(0,T) \tag{3.35c}$$

for all $\varphi\in V$ and $\phi\in\mathcal{U}$ with $\mathfrak{y}(0)=\mathfrak{y}_\circ$ and $\mathfrak{p}(T)=0$.

Finite Element Discretization

Now we discretize the first order optimality system (3.35) related to the state $\mathfrak{y}$, the adjoint $\mathfrak{p}$, the control $\mathfrak{u}$ and to functionals and dualities. Therefore, let $V^\mathsf{fe}=\operatorname{span}(\varphi_1,...,\varphi_{n_y})\subset V$ and $H^\mathsf{fe}=\operatorname{span}(\phi_1,...,\varphi_{n_u})\subset H$ be the finite element spaces introduced in (3.20). We quickly recall the approximated state and control given by

$$\mathfrak{y}(t,\boldsymbol{x})\approx\mathfrak{y}^\mathsf{fe}(t,\boldsymbol{x})=\sum_{i=1}^{n_y}y_i(t)\varphi_i(\boldsymbol{x}),\quad \mathfrak{u}(t,\boldsymbol{x})\approx\mathfrak{u}^\mathsf{fe}(t,\boldsymbol{x})=\sum_{i=1}^{n_u}u_i(t)\phi_i(\boldsymbol{x})$$

for $(t,\boldsymbol{x})\in\Omega_T$. We discretize the adjoint as

$$\mathfrak{p}(t,\boldsymbol{x})\approx\mathfrak{p}^\mathsf{fe}(t,\boldsymbol{x})=\sum_{i=1}^{n_y}p_i(t)\varphi_i(\boldsymbol{x}),$$

and define $p(t):=(p_1(t),...,p_{n_y}(t))\in\mathbb{R}^{n_y}$. The approximated optimality system reads as follow

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \mathfrak{y}^\mathsf{fe}(t),\varphi\right\rangle_H + a(t;\mathfrak{y}^\mathsf{fe}(t),\varphi) = \left\langle \mathfrak{f}(t)+\mathcal{B}(t)\mathfrak{u}^\mathsf{fe}(t),\varphi\right\rangle_{V',V}, \qquad t\in(0,T], \tag{3.36a}$$

$$-\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \mathfrak{p}^\mathsf{fe}(t),\varphi\right\rangle_H + a(t;\varphi,\mathfrak{p}^\mathsf{fe}(t)) = \left\langle \mathfrak{z}_\mathsf{d}(t)-\mathcal{C}(t)\mathfrak{y}^\mathsf{fe}(t),\mathcal{C}(t)\varphi\right\rangle_H, \qquad t\in[0,T), \tag{3.36b}$$

$$\left\langle \mathfrak{u}^\mathsf{fe}(t),\phi\right\rangle_H = \left\langle \frac{1}{\sigma}\mathcal{B}(t)^\star\mathfrak{p}^\mathsf{fe}(t),\phi\right\rangle_H, \qquad t\in(0,T) \tag{3.36c}$$

for all $\varphi\in V^\mathsf{fe}$ and $\phi\in H^\mathsf{fe}$ with the initial values $\mathfrak{y}^\mathsf{fe}(0)=\mathfrak{y}_\circ$ and $\mathfrak{p}^\mathsf{fe}(T)=0$. Introducing the FE vectors (5.7) and the FE matrices (3.23) the semi-discrete optimality system (3.36a) turns to

$$M\dot{y}(t)+A(t)y(t)=Mf(t)+B(t)u(t), \qquad t\in(0,T], \tag{3.37a}$$

$$-M\dot{p}(t)+A(t)^\top p(t)=C(t)(z_d(t)-y(t)), \qquad t\in[0,T), \tag{3.37b}$$

$$Ru(t)=\frac{1}{\sigma}B(t)^\top p(t), \qquad t\in(0,T) \tag{3.37c}$$

with $y(t) = y_\circ$ and $p(T) = 0$. Analog to Section 3.1.4 we make the unusual transformation and transform (3.37) into a dynamical system. Firstly, we insert equation (3.37c) in (3.37a) and obtain

$$M\dot{y}(t) + A(t)y(t) = Mf(t) + \frac{1}{\sigma}B(t)R^{-1}B(t)^\top p(t), \qquad t \in (0,T], \tag{3.38}$$

$$-M\dot{p}(t) + A(t)^\top p(t) = C(t)(z_d(t) - y(t)), \qquad t \in [0,T) \tag{3.39}$$

with $y(t) = y_\circ$ and $p(T) = 0$. We introduce the transformed variable $q(t) := p(T-t)$. For $t \in (0,T)$ let

$$x(t) = \begin{bmatrix} y(t) \\ q(t) \end{bmatrix} \in \mathbb{R}^{2n_y},\ \mathcal{F}(t) = \begin{bmatrix} f(t) \\ \tilde{C}(t)\tilde{z}_{\mathsf{d}}(t) \end{bmatrix} \in \mathbb{R}^{2n_y},\ x_\circ = \begin{bmatrix} y_\circ \\ 0 \end{bmatrix} \in \mathbb{R}^{2n_y},$$

where $\tilde{C}(t) = C(T-t)$ and $\tilde{z}_d(t) = z_d(T-t)$. Then, (3.38) can be written as

$$\dot{x}(t) = \mathcal{A}(t)x(t) + \tilde{\mathcal{A}}(t)x(T-t) + \mathcal{F}(t) \text{ for } t \in (0,T), \quad x(0) = x_\circ \tag{3.40}$$

with the two $(2n_y) \times (2n_y)$-matrices

$$\mathcal{A}(t) = \begin{bmatrix} M^{-1}A(t) & 0 \\ 0 & M^{-1}\tilde{A}(t)^\top \end{bmatrix},\ \tilde{\mathcal{A}}(t) = \begin{bmatrix} 0 & \frac{1}{\sigma}M^{-1}B(t)R^{-1}B(t)^\top \\ -M^{-1}\tilde{C}(t) & 0 \end{bmatrix},$$

where $\tilde{A}(t) = A(T-t)$.

Discussion

Now, let us discuss the two previous approaches. It is clear that choosing the same FE ansatz space V^{fe} for the state $\mathfrak{y}$ and the adjoint variable $\mathfrak{p}$ the First-Optimize-Then-Discretize approach is identical to the First-Discretize-Then-Optimize approach, as seen previously. Nevertheless, in some cases, it makes sense to choose a different FE ansatz space for the adjoint variable $\mathfrak{p}$. In many cases of optimal control problems with PDE constraints, the adjoint variable $\mathfrak{p}$ has more regularity than the state $\mathfrak{y}$. While the regularity of the state $\mathfrak{y}$ is affected by the regularity of the control $\mathfrak{u}$, which is typically nonsmooth and only in L^2, the adjoint $\mathfrak{p}$ has the desired state $\mathfrak{z}_d$ and the state $\mathfrak{y}$ itself as its right-hand side. If $\mathfrak{z}_d$ is sufficient smooth, it can be shown (more details in [19]) that the adjoint $\mathfrak{p}$ admits two more weak derivatives than the state $\mathfrak{y}$. Consequently, it can be worthwhile to use a FE ansatz space with higher polynomial degree for $\mathfrak{p}$ than for $\mathfrak{y}$. This can only be achieved using the First-Optimize-Then-Discretize approach and if a different FE ansatz space is chosen for the adjoint, the two approaches are not coinciding.

Thus far, there is no general rule specifying which approach should be preferred. Instead, it should depend on the application and the available computational resources.

3.2.3 Additional Control Constraints

In most applications, it is not typical to have access to an unbounded control function. For example, in a heating problem where the goal is to control a system to reach a desired

room temperature, it is not possible to have an unbounded heater. There is a natural limit. In this section, we take this bound in consideration and make the following assumptions.

Let $\mathfrak{u}_a, \mathfrak{u}_b \in \mathcal{U}$ are given functions with $\mathfrak{u}_a(t,\boldsymbol{x}) \leq \mathfrak{u}_b(t,\boldsymbol{x})$ for almost all $(t,\boldsymbol{x}) \in \Omega_T$. We define the admissible set by

$$\mathcal{U}_{\text{ad}} = \{\mathfrak{u} \in \mathcal{U} : \mathfrak{u}_a \leq \mathfrak{u} \leq \mathfrak{u}_b\}.$$

Notice that the set $\mathcal{U}_{\text{ad}}$ is convex, bounded, closed and nonempty. The new optimal control problem reads as follow

$$\min \mathcal{J}(\mathfrak{y},\mathfrak{u}) \quad \text{s.t.} \quad (\mathfrak{y},\mathfrak{u}) \text{ satisfies (3.18) and } \mathfrak{u} \in \mathcal{U}_{\text{ad}}, \tag{$\mathbf{P}^c_{\text{PDE}}$}$$

where $\mathcal{J} : \mathcal{Y} \times \mathcal{U} \to \mathbb{R}$ is given in (3.17). For the reduced problem we obtain

$$\min_{\mathfrak{u}\in\mathcal{U}_{\text{ad}}} \hat{\mathcal{J}}(\mathfrak{u}), \tag{$\hat{\mathbf{P}}^c_{\text{PDE}}$}$$

where $\hat{\mathcal{J}} : \mathcal{U} \to \mathbb{R}$ is given in (3.29). The following theorem guarantees us the existence of a unique solution of the optimal control problem.

Theorem 3.21 (Existence and uniqueness of ($\mathbf{P}^c_{\text{PDE}}$))
The optimal control problem ($\mathbf{P}^c_{\text{PDE}}$) *has a unique solution* $(\bar{\mathfrak{y}},\bar{\mathfrak{u}}) \in \mathcal{Y} \times \mathcal{U}$.

Proof. The proof follows directly from Theorem 2.15. □

The sufficient first order optimality condition of ($\hat{\mathbf{P}}^c_{\text{PDE}}$) reads as the variational inequality

$$\langle \nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}), \mathfrak{u}-\bar{\mathfrak{u}}\rangle_{\mathcal{U}} \geq 0 \quad \text{for all } \mathfrak{u} \in \mathcal{U}_{\text{ad}}, \tag{3.41}$$

where $\bar{\mathfrak{u}} \in \mathcal{U}_{\text{ad}}$ is the optimal solution of ($\mathbf{P}^c_{\text{PDE}}$). A proof can be found in [19, Theorem 1.46]. Next we want to simplify the variational inequality (3.41) by transforming it into an equality. Therefore, we need the following lemma.

Lemma 3.22
Let $\bar{\mathfrak{u}} \in \mathcal{U}_{ad}$*, then the following conditions are equivalent:*

(i) It holds

$$\langle \nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}), \mathfrak{u}-\bar{\mathfrak{u}}\rangle_{\mathcal{U}} \geq 0 \quad \textit{for all } \mathfrak{u} \in \mathcal{U}_{ad}.$$

(ii) We have

$$\nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}})(t,\boldsymbol{x}) \begin{cases} = 0 & \textit{if } \mathfrak{u}_a(t,\boldsymbol{x}) < \bar{\mathfrak{u}}(t,\boldsymbol{x}) < \mathfrak{u}_b(t,\boldsymbol{x}), \\ \geq 0 & \textit{if } \mathfrak{u}_a(t,\boldsymbol{x}) = \bar{\mathfrak{u}}(t,\boldsymbol{x}) < \mathfrak{u}_b(t,\boldsymbol{x}), \\ \leq 0 & \textit{if } \mathfrak{u}_a(t,\boldsymbol{x}) < \bar{\mathfrak{u}}(t,\boldsymbol{x}) = \mathfrak{u}_b(t,\boldsymbol{x}), \end{cases} \quad \textit{for a.a. } (t,\boldsymbol{x}) \in \Omega_T.$$

(iii) There exists $\bar{\lambda} \in U$ *with*

$$\nabla\hat{\mathcal{J}}(\mathfrak{u}) + \bar{\lambda} = 0,$$

where $\bar{\lambda} = \min(0, -\nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) + \eta(\bar{\mathfrak{u}} - \mathfrak{u}_a)) + \max(0, -\nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) + \eta(\bar{\mathfrak{u}} - \mathfrak{u}_b)) \in \mathcal{U}$ *for all* $\eta > 0$*. The* min-, max- *operators should be understood pointwise for* $(t,\boldsymbol{x}) \in \Omega_T$.

Proof. $(i) \Rightarrow (ii)$: Clearly (ii) is the same as

$$\nabla \hat{\mathcal{J}}(\bar{\mathfrak{u}})(t,\boldsymbol{x}) \begin{cases} \geq 0 & \text{a.e. on } I_a, \\ \leq 0 & \text{a.e. on } I_b, \end{cases}$$

where

$$I_a = \{(t,\boldsymbol{x}) \in \Omega_T : u_a(t,\boldsymbol{x}) \leq \bar{\mathfrak{u}}(t,\boldsymbol{x}) < \mathfrak{u}_b(t,\boldsymbol{x})\},$$
$$I_b = \{(t,\boldsymbol{x}) \in \Omega_T : u_a(t,\boldsymbol{x}) < \bar{\mathfrak{u}}(t,\boldsymbol{x}) \leq \mathfrak{u}_b(t,\boldsymbol{x})\}.$$

Assume (ii) is not true. Without loss of generality there exists a set $M \subset I_a$ with positive (Lebesgue) measure and $\nabla \hat{\mathcal{J}}(\bar{\mathfrak{u}})(t,\boldsymbol{x}) < 0$ on M. Let $\mathfrak{u} = \bar{\mathfrak{u}} + \chi_M(\mathfrak{u}_b - \bar{\mathfrak{u}}) \in \mathcal{U}_{ad}$, where χ is the indicator function. Then it holds $\mathfrak{u} - \bar{\mathfrak{u}} = \mathfrak{u}_b - \bar{\mathfrak{u}} > 0$ on M and $\mathfrak{u} - \bar{\mathfrak{u}} = 0$ elsewhere. But then we get the contradiction

$$\langle \nabla \hat{\mathcal{J}}(\bar{\mathfrak{u}}), \mathfrak{u} - \bar{\mathfrak{u}} \rangle_{\mathcal{U}} = \int_M \underbrace{\hat{\mathcal{J}}(\bar{\mathfrak{u}})}_{<0} \underbrace{(\mathfrak{u} - \bar{\mathfrak{u}})}_{>0} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t < 0.$$

$(ii) \Rightarrow (i)$: 1) If $\mathfrak{u}_a < \bar{\mathfrak{u}} < \mathfrak{u}_b$ in $\mathcal{U}$ it holds

$$\underbrace{\nabla \hat{\mathcal{J}}(\bar{\mathfrak{u}})(t,\boldsymbol{x})}_{=0}(\mathfrak{u}(t,\boldsymbol{x}) - \bar{\mathfrak{u}}(t,\boldsymbol{x})) = 0$$

almost everywhere for all $\mathfrak{u} \in \mathcal{U}_{\mathrm{ad}}$.
2) If $\mathfrak{u}_a = \bar{\mathfrak{u}}$ in $\mathcal{U}$ it holds

$$\nabla \hat{\mathcal{J}}(\bar{\mathfrak{u}})(t,\boldsymbol{x})(\mathfrak{u}(t,\boldsymbol{x}) - \bar{\mathfrak{u}}(t,\boldsymbol{x})) = \underbrace{\nabla \hat{\mathcal{J}}(\bar{\mathfrak{u}})(t,\boldsymbol{x})}_{\geq 0} \underbrace{(\mathfrak{u}(t,\boldsymbol{x}) - \mathfrak{u}_a(t,\boldsymbol{x}))}_{\geq 0} \geq 0$$

almost everywhere for all $\mathfrak{u} \in \mathcal{U}_{\mathrm{ad}}$.
3) If $\mathfrak{u}_b = \bar{\mathfrak{u}}$ in $\mathcal{U}$ it holds

$$\nabla \hat{\mathcal{J}}(\bar{\mathfrak{u}})(t,\boldsymbol{x})(\mathfrak{u}(t,\boldsymbol{x}) - \bar{\mathfrak{u}}(t,\boldsymbol{x})) = \underbrace{\nabla \hat{\mathcal{J}}(\bar{\mathfrak{u}})(t,\boldsymbol{x})}_{\leq 0} \underbrace{(\mathfrak{u}(t,\boldsymbol{x}) - \mathfrak{u}_b(t,\boldsymbol{x}))}_{\leq 0} \geq 0$$

almost everywhere for all $\mathfrak{u} \in \mathcal{U}_{\mathrm{ad}}$.
Consequently, we get

$$\langle \nabla \hat{\mathcal{J}}(\bar{\mathfrak{u}}), \mathfrak{u} - \bar{\mathfrak{u}} \rangle_{\mathcal{U}} \geq 0 \quad \text{for all } \mathfrak{u} \in \mathcal{U}_{\mathrm{ad}}.$$

$(ii) \Rightarrow (iii)$: Let us define $\bar{\lambda}_a(\eta) := \min(0, -\nabla \hat{\mathcal{J}}(\mathfrak{u}) + \eta(\bar{\mathfrak{u}} - \mathfrak{u}_a))$ and $\bar{\lambda}_b(\eta) := \max(0, -\nabla \hat{\mathcal{J}}(\mathfrak{u}) + \eta(\bar{\mathfrak{u}} - \mathfrak{u}_b))$.
1) If $\mathfrak{u}_a < \bar{\mathfrak{u}} < \mathfrak{u}_b$ in $\mathcal{U}$ it holds $\nabla \hat{\mathcal{J}}(\bar{\mathfrak{u}}) = 0$ and therefore

$$\begin{aligned} \nabla \hat{\mathcal{J}}(\bar{\mathfrak{u}}) + \bar{\lambda} &= \nabla \hat{\mathcal{J}}(\bar{\mathfrak{u}}) + \bar{\lambda}_a(\eta) + \bar{\lambda}_b(\eta) \\ &= \min(0, \underbrace{\eta(\bar{\mathfrak{u}} - \mathfrak{u}_a)}_{>0}) + \max(0, \underbrace{\eta(\bar{\mathfrak{u}} - \mathfrak{u}_b)}_{<0}) = 0. \end{aligned}$$

2) If $\mathfrak{u}_a = \bar{\mathfrak{u}}$ in $\mathcal{U}$ it holds

$$\begin{aligned}\nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) + \bar{\lambda} &= \nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) + \bar{\lambda}_a(\eta) + \bar{\lambda}_b(\eta)\\ &= \nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) + \min(0, \underbrace{-\nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}})}_{\leq 0}) + \max(0, \underbrace{-\nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}})}_{\leq 0} + \underbrace{\eta(\mathfrak{u}_a - \mathfrak{u}_b)}_{\leq 0})\\ &= \nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) - \nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) = 0.\end{aligned}$$

3) If $\mathfrak{u}_b = \bar{\mathfrak{u}}$ in $\mathcal{U}$ it holds

$$\begin{aligned}\nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) + \bar{\lambda} &= \nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) + \bar{\lambda}_a(\eta) + \bar{\lambda}_b(\eta)\\ &= \nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) + \min(0, \underbrace{-\nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}})}_{\geq 0} + \underbrace{\eta(\mathfrak{u}_b - \mathfrak{u}_a)}_{\geq 0}) + \max(0, \underbrace{-\nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}})}_{\geq 0})\\ &= \nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) - \nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) = 0.\end{aligned}$$

$(iii) \Rightarrow (ii)$: 1) If $\mathfrak{u}_a < \bar{\mathfrak{u}} < \mathfrak{u}_b$ in $\mathcal{U}$ it holds

$$0 = \nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) + \bar{\lambda} = \nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) + \bar{\lambda}_a(\eta) + \bar{\lambda}_b(\eta).$$

Since the equality holds for all $\eta > 0$, the terms $\eta(\bar{\mathfrak{u}} - \mathfrak{u}_a)$ and $\eta(\bar{\mathfrak{u}} - \mathfrak{u}_b)$ can be made arbitrarily large or small, which implies $\bar{\lambda}_a(\eta) = \bar{\lambda}_b(\eta) = 0$ and therefore $\nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) = 0$.
2) If $\mathfrak{u}_a = \bar{\mathfrak{u}}$ in $\mathcal{U}$ it holds

$$0 = \nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) + \bar{\lambda} = \nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) + \min(0, -\nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}})) + \bar{\lambda}_b(\eta).$$

Since the equality holds for all $\eta > 0$, it implies $\bar{\lambda}_b(\eta) = 0$ and therefore $\nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) = -\min(0, -\nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}})) \geq 0$.
3) If $\mathfrak{u}_b = \bar{\mathfrak{u}}$ in $\mathcal{U}$ it holds

$$0 = \nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) + \bar{\lambda} = \nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) + \bar{\lambda}_a(\eta) + \max(0, -\nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}})).$$

Since the equality holds for all $\eta > 0$, it implies $\bar{\lambda}_a(\eta) = 0$ and therefore $\nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}}) = -\max(0, -\nabla\hat{\mathcal{J}}(\bar{\mathfrak{u}})) \leq 0$. □

Next, if we use the representation of the gradient $\nabla\hat{\mathcal{J}}(\mathfrak{u}) = \sigma\mathfrak{u} - \mathcal{B}^\star\mathfrak{p}$ and choose $\eta = \sigma$ the first order optimality condition (3.41) becomes

$$\sigma\bar{\mathfrak{u}} - \mathcal{B}^\star\bar{\mathfrak{p}} + \min(0, \mathcal{B}^\star\bar{\mathfrak{p}} - \sigma\mathfrak{u}_a) + \max(0, \mathcal{B}^\star\bar{\mathfrak{p}} - \sigma\mathfrak{u}_a) = 0 \in \mathcal{U}. \tag{3.42}$$

Optimality System

Let us define the active sets by

$$\begin{aligned}\mathcal{A}_a(\mathfrak{p}) &:= \{(t, \boldsymbol{x}) \in \Omega_T : \mathcal{B}^\star\mathfrak{p} - \sigma\mathfrak{u}_a < 0 \text{ a.e.}\},\\ \mathcal{A}_b(\mathfrak{p}) &:= \{(t, \boldsymbol{x}) \in \Omega_T : \mathcal{B}^\star\mathfrak{p} - \sigma\mathfrak{u}_b > 0 \text{ a.e.}\}.\end{aligned}$$

The inactive set is given by

$$\mathcal{I}(\mathfrak{p}) = \Omega_T \backslash (\mathcal{A}_a(\mathfrak{p}) \cup \mathcal{A}_b(\mathfrak{p})).$$

Moreover, we define

$$\mathfrak{g}_a(t,\mathfrak{p}) := \chi_{\mathcal{A}_a(\mathfrak{p})}(\mathcal{B}(t)^\star\mathfrak{p}(t) - \sigma\mathfrak{u}_a(t)) \quad \text{and} \quad \mathfrak{g}_b(t,\mathfrak{p}) := \chi_{\mathcal{A}_b(\mathfrak{p})}(\mathcal{B}(t)^\star\mathfrak{p}(t) - \sigma\mathfrak{u}_b(t)).$$

Then the optimality system reads as follow.

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle\mathfrak{y}(t),\varphi\rangle_H + a(t;\mathfrak{y}(t),\varphi) = \langle\mathfrak{f}(t)+\mathcal{B}(t)\mathfrak{u}(t),\varphi\rangle_{V',V}, \qquad t\in(0,T], \quad (3.43a)$$

$$-\frac{\mathrm{d}}{\mathrm{d}t}\langle\mathfrak{p}(t),\varphi\rangle_H + a(t;\varphi,\mathfrak{p}(t)) = \langle\mathfrak{z}_\mathfrak{d}(t)-\mathcal{C}(t)\mathfrak{y}(t),\mathcal{C}(t)\varphi\rangle_H, \qquad t\in[0,T), \quad (3.43b)$$

$$\int_0^T \langle\mathfrak{u}(t),\phi(t)\rangle_H\,\mathrm{d}t = \int_0^T \left\langle\frac{1}{\sigma}(\mathcal{B}(t)^\star\mathfrak{p}(t)-\mathfrak{g}(t,\mathfrak{p})),\phi(t)\right\rangle_H \mathrm{d}t, t\in(0,T) \quad (3.43c)$$

for all $\varphi\in V$ and $\phi\in\mathcal{U}$ with $\mathfrak{y}(0)=\mathfrak{y}_\circ$ and $\mathfrak{p}(T)=0$, where $\mathfrak{g}(t,\mathfrak{p}) = \mathfrak{g}_a(t,\mathfrak{p}) + \mathfrak{g}_b(t,\mathfrak{p})$.

Finite Element Discretization

Again, we discretize the optimality system (3.43) related to the state $\mathfrak{y}$, the adjoint $\mathfrak{p}$, the control $\mathfrak{u}$. Let $V^{\mathsf{fe}} = \mathrm{span}(\varphi_1,...,\varphi_{n_y}) \subset V$ and $H^{\mathsf{fe}} = \mathrm{span}(\phi_1,...,\varphi_{n_u}) \subset H$ be the finite element spaces introduced in (3.20). All FE functions, matrices and vectors are defined than before (for the details see equation (5.6), (5.7) and (3.23)). The boundary functions are approximated by

$$\mathfrak{u}_a(t,\boldsymbol{x}) \approx \mathfrak{u}_a^{\mathsf{fe}}(t,\boldsymbol{x}) = \sum_{i=1}^{n_u} u_{ai}(t)\phi_i(\boldsymbol{x}), \quad \mathfrak{u}_b(t,\boldsymbol{x}) \approx \mathfrak{u}_b^{\mathsf{fe}}(t,\boldsymbol{x}) = \sum_{i=1}^{n_u} u_{bi}(t)\phi_i(\boldsymbol{x}), \qquad (3.44)$$

and we set $u_a(t) := (u_{a1},...,u_{an_u})^\top \in \mathbb{R}^{n_u}$ and $u_b(t) := (u_{b1},...,u_{bn_u})^\top \in \mathbb{R}^{n_u}$ We define the discrete sets by

$$\mathcal{A}_a^{\mathsf{fe}}(p) := \{(t,i)\in(0,T)\times\{1,...,n_u\} : (B(t)^\top p(t) - \sigma u_a(t))_i < 0\},$$
$$\mathcal{A}_b^{\mathsf{fe}}(p) := \{(t,i)\in(0,T)\times\{1,...,n_u\} : (B(t)^\top p(t) - \sigma u_b(t))_i < 0\}.$$

Moreover, we define

$$g_a(t,p) := \chi_{\mathcal{A}_a^{\mathsf{fe}}(p)}(B(t)^\top p(t) - \sigma u_a(t)) \quad \text{and} \quad g_b(t,p) := \chi_{\mathcal{A}_b^{\mathsf{fe}}(p)}(B(t)^\top p(t) - \sigma u_b(t)).$$

Then the optimality system (3.43) is approximated by

$$M\dot{y}(t) + A(t)y(t) = Mf(t) + B(t)u(t), \qquad t\in(0,T], \qquad (3.45a)$$

$$-M\dot{p}(t) + A(t)^\top p(t) = C(t)(z_d(t) - y(t)), \qquad t\in[0,T), \qquad (3.45b)$$

$$Ru(t) = \frac{1}{\sigma}(B(t)^\top p(t) - g(t,p)), \qquad t\in(0,T) \qquad (3.45c)$$

with $y(t) = y_\circ$ and $p(T) = 0$, where $g(t,p) = g_a(t,b) + g_b(t,p)$. Inserting equation (3.45c) in (3.45a) leads to

$$M\dot{y}(t) + A(t)y(t) = Mf(t) + \frac{1}{\sigma}B(t)R^{-1}(B(t)^\top p(t) - g(t,p)), \quad t\in(0,T], \quad (3.46a)$$

$$-M\dot{p}(t) + A(t)^\top p(t) = C(t)(z_d(t) - y(t)), \qquad t\in[0,T) \quad (3.46b)$$

with $y(t) = y_\circ$ and $p(T) = 0$.

Remark 3.23
Since the optimality system (3.46) *is nonlinear in* p*, because of the term* $g(t,p)$*, we can not transform it into a linear dynamical system of the form* (3.16)*.*

Anyhow, we can transform it into a nonlinear dynamical system by introducing $q(t) := p(T-t)$ and for $t \in (0,T)$ let

$$x(t) = \begin{bmatrix} y(t) \\ q(t) \end{bmatrix} \in \mathbb{R}^{2n_y},\ \mathcal{F}(t) = \begin{bmatrix} f(t) \\ \tilde{C}(t)\tilde{z}_{\mathsf{d}}(t) \end{bmatrix} \in \mathbb{R}^{2n_y},\ x_\circ = \begin{bmatrix} y_\circ \\ 0 \end{bmatrix} \in \mathbb{R}^{2n_y},$$

where $\tilde{C}(t) = C(T-t)$ and $\tilde{z}_d(t) = z_d(T-t)$. Then, (3.46) can be written as

$$\dot{x}(t) = \mathcal{A}(t)x(t) + \tilde{\mathcal{A}}(t)x(T-t) + \mathcal{G}(t, x(T-t)) + \mathcal{F}(t) \text{ for } t \in (0,T), \quad x(0) = x_\circ \tag{3.47}$$

with the two $(2n_y) \times (2n_y)$-matrices

$$\mathcal{A}(t) = \begin{bmatrix} M^{-1}A(t) & 0 \\ 0 & M^{-1}\tilde{A}(t)^\top \end{bmatrix},\ \tilde{\mathcal{A}}(t) = \begin{bmatrix} 0 & \frac{1}{\sigma}M^{-1}B(t)R^{-1}B(t)^\top \\ -M^{-1}\tilde{C}(t) & 0 \end{bmatrix},$$

where $\tilde{A}(t) = A(T-t)$ and the nonlinear term

$$\mathcal{G}(t, x(T-t)) = \begin{bmatrix} -\frac{1}{\sigma}B(t)R^{-1}g(t, x_2(T-t)) \\ 0 \end{bmatrix} \in \mathbb{R}^{2n_y}.$$

4 | Data-Driven Model-Order Reduction

In this chapter, we introduce the concept of *model-order reduction* (MOR) for optimal control problems considered in Chapter 3. MOR is a rapidly advancing field of research that has seen significant growth in both theory and applications in recent years. For an introduction and comprehensive overview, we refer to the references [15, 35], for instance. Without loss of generality, we focus here on the PDE constrained optimal control problem ($\mathbf{P}_{\mathrm{PDE}}$), but same computations can be done analogously for ($\mathbf{P}_{\mathrm{ODE}}$) and ($\mathbf{P}^{c}_{\mathrm{PDE}}$). Firstly, we introduce two different approaches to obtain a reduced model of ($\mathbf{P}_{\mathrm{PDE}}$). We start with the well-known approach, where the state and adjoint equation are reduced separately. Here, we can derive a-posteriori error estimates. Secondly, we introduce the new approach, where the whole optimality system is reduced in one shot, which is also the focus of this thesis. Afterwards, we introduce three different MOR techniques, proper orthogonal decomposition, empirical gramians and extended dynamic mode decomposition. All of these techniques are data-driven, which means that we require some starting snapshots of the full model in advance before we can execute the model-order reduction.

4.1 Model-Order Reduction

Model-order reduction is a technique that deals with the problem of simplification of mathematical models without significantly sacrificing its accuracy. The goal of MOR is to simplify the model while still maintaining its ability to predict the behavior of the system over a certain range of conditions. The reduced order model is then used in place of the original model to perform simulations or control design. The main advantage of MOR is that it can significantly reduce the computational resources and complexity of simulating or controlling large-scale systems.

Example 4.1 (MOR for dynamical systems)
We consider the high-dimensional autonomous dynamical system

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} = \mathcal{F}(y(t), u(t)) \tag{4.1}$$

with $\mathcal{F} : \mathbb{R}^{n_y} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_y}$, $y(t) \in \mathbb{R}^{n_y}$ *and* $u(t) \in \mathbb{R}^{n_u}$. *The goal of model-order reduction is to approximate* (4.1) *with another reduced dynamical system*

$$\frac{\mathrm{d}y^{\ell}(t)}{\mathrm{d}t} = \mathcal{F}^{\ell}(y^{\ell}(t), u(t)) \tag{4.2}$$

with $\mathcal{F}^\ell : \mathbb{R}^\ell \times \mathbb{R}^{n_u} \to \mathbb{R}^\ell$, $y^\ell(t) \in \mathbb{R}^\ell$ *and* $u(t) \in \mathbb{R}^{n_u}$, *such that* $\ell \ll n_y$. *Obviously, solving system* (4.2) *requires less computational effort than solving the original system* (4.1).

A satisfactory reduced model should have the following characteristics:

(i) The approximation error should be small.

(ii) The reduced model should be computationally efficient.

(iii) The procedure to get the reduced model should be automatic and also computationally efficient.

4.1.1 The Separation Approach

For the separation approach, we assume that we need to solve the state and the adjoint equation iteratively for different inputs $\mathfrak{u} \in \mathcal{U}$. This is the case, for example, if we want to solve the optimal control problem ($\mathbf{P}_{\mathrm{PDE}}$) by a gradient descent method. The separation approach approximates the two equations by reduced equations, i.e. we introduce the two low-dimensional subspaces $V_\mathfrak{y}^\ell, V_\mathfrak{p}^\ell \subset V$ and consider the state equation

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\left\langle \mathfrak{y}^\ell(t), \varphi\right\rangle_H + a(t; \mathfrak{y}^\ell(t), \varphi) &= \left\langle \mathfrak{f}(t) + \mathcal{B}(t)\mathfrak{u}(t), \varphi\right\rangle_{V',V} \quad \text{for all } \varphi \in V_\mathfrak{y}^\ell \text{ and } t \in (0,T], \\ \mathfrak{y}^\ell(0) &= \mathcal{P}^\ell \mathfrak{y}_\circ \qquad\qquad\qquad\qquad \text{in } H \end{aligned} \tag{4.3}$$

with a linear and bounded projection $\mathcal{P}^\ell : H \to V_\mathfrak{y}^\ell$ and the adjoint equation by

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \mathfrak{p}^\ell(t), \varphi\right\rangle_H + a(t; \varphi, \mathfrak{p}^\ell(t)) &= \left\langle \mathfrak{z}_\mathsf{d}(t) - \mathcal{C}(t)\mathfrak{y}^\ell(t), \mathcal{C}(t)\varphi\right\rangle_H \quad \text{for all } \varphi \in V_\mathfrak{p}^\ell, t \in [0,T) \\ \mathfrak{p}^\ell(T) &= 0 \qquad\qquad\qquad\qquad \text{in } H. \end{aligned} \tag{4.4}$$

Remark 4.2
Choosing $V_\mathfrak{y}^\ell = V_\mathfrak{p}^\ell = V^{\mathsf{fe}}$ *will lead to the standard FE approach we have discussed in Section 3.2.2. Typically, FE spaces are high-dimensional because they are designed to be a good approximation of the entire space* V. *On the other hand, the MOR spaces* $V_\mathfrak{y}^\ell$ *and* $V_\mathfrak{p}^\ell$ *should only approximate a space that contains, in some sense, the characteristics of the solutions* $\mathfrak{y} = \mathfrak{y}(\mathfrak{u})$, *respectively* $\mathfrak{p} = \mathfrak{p}(\mathfrak{u})$. *Therefore, the FE space is not a good choice for the MOR, because it is computationally too expensive. However, in the most applications one chooses* $V_\mathfrak{y}^\ell, V_\mathfrak{p}^\ell \subsetneq V^{\mathsf{fe}}$.

A-posteriori Error

Next, we want to find an a-posteriori error, which measures the error between the MOR and FE solution. We define the primal and the dual errors as

$$e^{\mathsf{pr}}(t) = \mathfrak{y}(t) - \mathfrak{y}^\ell(t) \in V \quad \text{and} \quad e^{\mathsf{du}}(t) = \mathfrak{p}(t) - \mathfrak{p}^\ell(t) \in V$$

for almost all $t \in (0,T]$. Furthermore, we define the primal and dual residuum as

$$\langle \mathsf{res}^{\mathsf{pr}}(t), \varphi \rangle_{V',V} := \langle \mathfrak{f}(t) + \mathcal{B}(t)\mathfrak{u}(t), \varphi \rangle_{V',V} - \frac{\mathrm{d}}{\mathrm{d}t} \langle \mathfrak{y}^\ell(t), \varphi \rangle_H - a(t; \mathfrak{y}^\ell(t), \varphi),$$
$$\langle \mathsf{res}^{\mathsf{du}}(t), \varphi \rangle_{V',V} := \langle \mathfrak{z}_d(t) - \mathcal{C}(t)\mathfrak{y}^\ell(t), \mathcal{C}(t)\varphi \rangle_H + \frac{\mathrm{d}}{\mathrm{d}t} \langle \mathfrak{p}^\ell(t), \varphi \rangle_H - a(t; \varphi, \mathfrak{p}^\ell(t))$$

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

Theorem 4.3 (A-posteriori primal error)
Let $\mathfrak{u} \in \mathcal{U}$, $\mathfrak{y}_\circ \in H$ be arbitrary and denote by $\mathfrak{y} \in \mathcal{Y}, \mathfrak{y}^\ell \in W(0,T;V_\mathfrak{y}^\ell)$ the unique solution of (3.18) *and* (4.3)*, respectively. Then,*

$$\|e^{\mathsf{pr}}(t)\|_H^2 \leq \exp(2\eta_2 t) \left(\|y_\circ - \mathcal{P}^\ell y_\circ\|_H^2 + \frac{1}{\eta_1} \int_0^t \|\mathsf{res}^{\mathsf{pr}}(t)\|_{V'}^2 \,\mathrm{d}s \right) \quad \textit{f.a.a. } t \in [0,T]$$

and

$$\int_0^t \|e^{\mathsf{pr}}(s)\|_V^2 \,\mathrm{d}s \leq c_1 \|y_\circ - \mathcal{P}^\ell y_\circ\|_H^2 + c_2 \int_0^t \|\mathsf{res}^{\mathsf{pr}}(s)\|_{V'}^2 \,\mathrm{d}s \quad \textit{f.a.a. } t \in [0,T]$$

with $c_1 = \exp(2\eta_2 t)/\eta_1$, $c_2 = 2(\exp(2\eta_2 t) - 1)/\eta_1^2$, $\eta_1 > 0$ and $\eta_2 \geq 0$ are the coercivity constants of a.

Proof. Let $\varphi \in V$ be arbitrary. Note that

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t} \langle e^{\mathsf{pr}}(t), \varphi \rangle_H + a(t; e^{\mathsf{pr}}(t), \varphi) \\
&\quad = \frac{\mathrm{d}}{\mathrm{d}t} \langle \mathfrak{y}(t), \varphi \rangle_H + a(t; \mathfrak{y}(t), \varphi) - \frac{\mathrm{d}}{\mathrm{d}t} \langle \mathfrak{y}^\ell(t), \varphi \rangle_H - a(t; \mathfrak{y}^\ell(t), \varphi) \\
&\quad = \langle \mathfrak{f}(t) + \mathcal{B}(t)\mathfrak{u}(t), \varphi \rangle_{V',V} - \frac{\mathrm{d}}{\mathrm{d}t} \langle \mathfrak{y}^\ell(t), \varphi \rangle_H - a(t; \mathfrak{y}^\ell(t), \varphi) = \langle \mathsf{res}^{\mathsf{pr}}(t), \varphi \rangle_{V',V}
\end{aligned}$$

for almost all $t \in (0,T]$. Now, choosing $\varphi = e^{\mathsf{pr}}(t)$ and using the coercivity of a (3.19b) yields

$$\begin{aligned}
\langle \mathsf{res}^{\mathsf{pr}}(t), e^{\mathsf{pr}}(t) \rangle_{V',V} &= \langle \mathfrak{f}(t) + \mathcal{B}(t)\mathfrak{u}(t), e^{\mathsf{pr}}(t) \rangle_{V',V} - \frac{\mathrm{d}}{\mathrm{d}t} \langle \mathfrak{y}^\ell(t), e^{\mathsf{pr}}(t) \rangle_H - a(t; \mathfrak{y}^\ell(t), e^{\mathsf{pr}}(t)) \\
&= \frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \|e^{\mathsf{pr}}(t)\|_H^2 + a(t; e^{\mathsf{pr}}(t), e^{\mathsf{pr}}(t)) \\
&\geq \frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \|e^{\mathsf{pr}}(t)\|_H^2 + \eta_1 \|e^{\mathsf{pr}}(t)\|_V^2 - \eta_2 \|e^{\mathsf{pr}}(t)\|_H^2
\end{aligned}$$

for almost all $t \in [0,T]$ for $\eta_1 > 0$ and $\eta_2 \geq 0$. This in combination with Young's inequality [11, p. 622] implies

$$\begin{aligned}
\frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \|e^{\mathsf{pr}}(t)\|_H^2 + \eta_1 \|e^{\mathsf{pr}}(t)\|_V^2 &\leq \eta_2 \|e^{\mathsf{pr}}(t)\|_H^2 + \|\mathsf{res}^{\mathsf{pr}}(t)\|_{V'} \|e^{\mathsf{pr}}(t)\|_V \\
&\leq \eta_2 \|e^{\mathsf{pr}}(t)\|_H^2 + \frac{1}{2\eta_1} \|\mathsf{res}^{\mathsf{pr}}(t)\|_{V'}^2 + \frac{\eta_1}{2} \|e^{\mathsf{pr}}(t)\|_V^2
\end{aligned}$$

for almost all $t \in [0,T]$. Consequently, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\|e^{\mathsf{pr}}(t)\|_H^2 + \eta_1\|e^{\mathsf{pr}}(t)\|_V^2 \leq 2\eta_2\|e^{\mathsf{pr}}(t)\|_H^2 + \frac{1}{\eta_1}\|\mathsf{res}^{\mathsf{pr}}(t)\|_{V'}^2 \tag{4.5}$$

for almost all $t \in [0,T]$. Utilizing Gronwall's lemma [Theorem 2.6] we find

$$\begin{aligned}\|e^{\mathsf{pr}}(t)\|_H^2 &\leq \exp(2\eta_2 t)\left(\|e^{\mathsf{pr}}(0)\|_H^2 + \frac{1}{\eta_1}\int_0^t \|\mathsf{res}^{\mathsf{pr}}(t)\|_{V'}^2\,\mathrm{d}s\right)\\ &= \exp(2\eta_2 t)\left(\|y_\circ - \mathcal{P}^\ell y_\circ\|_H^2 + \frac{1}{\eta_1}\int_0^t \|\mathsf{res}^{\mathsf{pr}}(t)\|_{V'}^2\,\mathrm{d}s\right)\end{aligned} \tag{4.6}$$

for almost all $t \in [0,T]$. Integrating (4.5) over $(0,t) \subseteq (0,T)$ and applying (4.6) we derive that

$$\begin{aligned}&\|e^{\mathsf{pr}}(t)\|_H^2 + \eta_1\int_0^t \|e^{\mathsf{pr}}(s)\|_V^2\,\mathrm{d}s\\ &\leq \|e^{\mathsf{pr}}(0)\|_H^2 + 2\eta_2\int_0^t \|e^{\mathsf{pr}}(s)\|_H^2\,\mathrm{d}s + \frac{1}{\eta_1}\int_0^t \|\mathsf{res}^{\mathsf{pr}}(s)\|_{V'}^2\,\mathrm{d}s\\ &\leq \|e^{\mathsf{pr}}(0)\|_H^2 + 2\eta_2\int_0^t \exp(2\eta_2 s)\left(\|e^{\mathsf{pr}}(0)\|_H^2 + \frac{1}{\eta_1}\int_0^s \|\mathsf{res}^{\mathsf{pr}}(\tau)\|_{V'}^2\,\mathrm{d}\tau\right)\mathrm{d}s\\ &\quad + \frac{1}{\eta_1}\int_0^t \|\mathsf{res}^{\mathsf{pr}}(s)\|_{V'}^2\,\mathrm{d}s.\end{aligned}$$

Consequently,

$$\begin{aligned}&\|e^{\mathsf{pr}}(t)\|_H^2 + \eta_1\int_0^t \|e^{\mathsf{pr}}(s)\|_V^2\,\mathrm{d}s\\ &\leq \|e^{\mathsf{pr}}(0)\|_H^2 + (\exp(2\eta_2 t)-1)\left(\|e^{\mathsf{pr}}(0)\|_H^2 + \frac{1}{\eta_1}\int_0^t \|\mathsf{res}^{\mathsf{pr}}(\tau)\|_{V'}^2\,\mathrm{d}\tau\right)\\ &\quad + \frac{1}{\eta_1}\int_0^t \|\mathsf{res}^{\mathsf{pr}}(s)\|_{V'}\,\mathrm{d}s\\ &\leq \exp(2\eta_2 t)\|e^{\mathsf{pr}}(0)\|_H^2 + \frac{2(\exp(2\eta_2 t)-1)}{\eta_1}\int_0^t \|\mathsf{res}^{\mathsf{pr}}(s)\|_{V'}^2\,\mathrm{d}s.\end{aligned}$$

□

Theorem 4.4 (A-posteriori dual error)
Let $\mathfrak{y}^\ell \in W(0,T;V_\mathfrak{y}^\ell)$ be arbitrary and denote by $\mathfrak{p} \in \mathcal{Y}, \mathfrak{p}^\ell \in W(0,T;V_\mathfrak{p}^\ell)$ the unique solution of (3.31) and (4.4), respectively. Moreover, we assume that $\|\mathcal{C}(t)\varphi\|_H \leq c_\mathcal{C}\|\varphi\|_H$ for all $\varphi \in V$ and almost all $t \in (0,T)$. Then,

$$\|e^{\mathsf{du}}(t)\|_H^2 \leq \exp(c_3(T-t))\left(\int_t^T \frac{1}{\eta_1}\|\mathsf{res}^{\mathsf{du}}(s)\|_{V'}^2 + c_\mathcal{C}\|e^{\mathsf{pr}}(s)\|_H^2\,\mathrm{d}s\right) \quad \textit{f.a.a. } t \in [0,T]$$

with $c_3 = c_\mathcal{C} + 2\eta_2$, $\eta_1 > 0$ and $\eta_2 \geq 0$ are the coercivity constants of a.

Proof. Let $\varphi \in V$ be arbitrary. Note that

$$\begin{aligned}
&-\frac{\mathrm{d}}{\mathrm{d}t}\left\langle e^{\mathsf{du}}(t),\varphi\right\rangle_H + a(t;\varphi,e^{\mathsf{du}}(t)) \\
&\quad = -\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \mathfrak{p}(t),\varphi\right\rangle_H + a(t;\varphi,\mathfrak{p}(t)) + \frac{\mathrm{d}}{\mathrm{d}t}\left\langle \mathfrak{p}^\ell(t),\varphi\right\rangle_H - a(t;\varphi,\mathfrak{p}^\ell(t)) \\
&\quad = \left\langle \mathfrak{z}(t) - \mathcal{C}(t)\mathfrak{y}(t), \mathcal{C}(t)\varphi\right\rangle_H + \frac{\mathrm{d}}{\mathrm{d}t}\left\langle \mathfrak{p}^\ell(t),\varphi\right\rangle_H - a(t;\varphi,\mathfrak{p}^\ell(t)) \\
&\quad = -\left\langle \mathcal{C}(t)e^{\mathsf{pr}}(t), \mathcal{C}(t)\varphi\right\rangle_H + \left\langle \mathsf{res}^{\mathsf{du}}(t),\varphi\right\rangle_{V',V}
\end{aligned}$$

for almost all $t \in (0,T]$. Now, choosing $\varphi = e^{\mathsf{du}}(t)$ we find

$$\begin{aligned}
&\left\langle \mathsf{res}^{\mathsf{du}}(t), e^{\mathsf{du}}(t)\right\rangle_{V',V} - \left\langle \mathcal{C}(t)e^{\mathsf{pr}}(t), \mathcal{C}(t)e^{\mathsf{du}}(t)\right\rangle_H \\
&\quad = -\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|e^{\mathsf{du}}(t)\|_H^2 + a(t;e^{\mathsf{du}}(t),e^{\mathsf{du}}(t)) \\
&\quad \geq -\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|e^{\mathsf{du}}(t)\|_H^2 + \eta_1\|e^{\mathsf{du}}(t)\|_V^2 - \eta_2\|e^{\mathsf{du}}(t)\|_H^2
\end{aligned}$$

for almost all $t \in [0,T]$ for $\eta_1 > 0$ and $\eta_2 \geq 0$. This in combination with Young's inequality implies

$$\begin{aligned}
&-\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|e^{\mathsf{du}}(t)\|_H^2 + \eta_1\|e^{\mathsf{du}}(t)\|_V^2 \\
&\leq \eta_2\|e^{\mathsf{du}}(t)\|_H^2 + \|\mathsf{res}^{\mathsf{du}}(t)\|_{V'}\|e^{\mathsf{du}}(t)\|_V + \|\mathcal{C}(t)e^{\mathsf{pr}}(t)\|_H\|\mathcal{C}(t)e^{\mathsf{du}}(t)\|_H \\
&\leq \eta_2\|e^{\mathsf{du}}(t)\|_H^2 + \frac{1}{2\eta_1}\|\mathsf{res}^{\mathsf{du}}(t)\|_{V'}^2 + \frac{\eta_1}{2}\|e^{\mathsf{du}}(t)\|_V^2 + \frac{1}{2}c_{\mathcal{C}}(\|e^{\mathsf{pr}}(t)\|_H^2 + \|e^{\mathsf{du}}(t)\|_H^2)
\end{aligned}$$

for almost all $t \in [0,T]$. Therefore,

$$\begin{aligned}
&-\frac{\mathrm{d}}{\mathrm{d}t}\|e^{\mathsf{du}}(t)\|_H^2 + \eta_1\|e^{\mathsf{du}}(t)\|_V^2 \\
&\qquad \leq (c_{\mathcal{C}} + 2\eta_2)\|e^{\mathsf{du}}(t)\|_H^2 + \frac{1}{\eta_1}\|\mathsf{res}^{\mathsf{du}}(t)\|_{V'}^2 + c_{\mathcal{C}}\|e^{\mathsf{pr}}(t)\|_H^2
\end{aligned}$$

for almost all $t \in [0,T]$. Utilizing Gronwall's Lemma [8, p. 559] and $e^{\mathsf{du}}(T) = 0$ we find

$$\begin{aligned}
\|e^{\mathsf{du}}(t)\|_H^2 &\leq \exp((c_{\mathcal{C}} + 2\eta_2)(T-t))\left(\|e^{\mathsf{du}}(T)\|_H^2 + \int_t^T \frac{1}{\eta_1}\|\mathsf{res}^{\mathsf{du}}(s)\|_{V'}^2 + c_{\mathcal{C}}\|e^{\mathsf{pr}}(s)\|_H^2\,\mathrm{d}s\right) \\
&\leq \exp((c_{\mathcal{C}} + 2\eta_2)(T-t))\left(\int_t^T \frac{1}{\eta_1}\|\mathsf{res}^{\mathsf{du}}(s)\|_{V'}^2 + c_{\mathcal{C}}\|e^{\mathsf{pr}}(s)\|_H^2\,\mathrm{d}s\right).
\end{aligned}$$

□

Now, we have the a-posteriori primal and dual error. Thus we can also make a statement for the control error.

Corollary 4.5 (A-posteriori control error)
Let $\bar{\mathfrak{r}} = (\bar{\mathfrak{y}}, \bar{\mathfrak{u}}, \bar{\mathfrak{p}})$ be the unique optimal triple of $(\mathbf{P}_{\mathrm{ODE}})$ satisfying the first order sufficient optimality conditions (3.35). Analogously, the reduced-order optimal control problem admits a unique optimal triple $\bar{\mathfrak{r}}^\ell = (\bar{\mathfrak{y}}^\ell, \bar{\mathfrak{u}}^\ell, \bar{\mathfrak{p}}^\ell)$. Moreover, we assume that $\|\mathcal{B}(t)^\star \varphi\|_H \leq c_{\mathcal{B}} \|\varphi\|_H$ and $\|\mathcal{C}(t)\varphi\|_H \leq c_{\mathcal{C}} \|\varphi\|_H$ for all $\varphi \in V$ and almost all $t \in (0,T)$. Then,

$$\begin{aligned}\|\bar{\mathfrak{u}} - \bar{\mathfrak{u}}^\ell\|_{\mathcal{U}}^2 \leq \frac{c_{\mathcal{B}}^2}{\sigma^2} \int_0^T e^{c_3(T-t)} \Bigg(\int_t^T \frac{1}{\eta_1} \|\mathsf{res}^{\mathsf{du}}(s)\|_{V'}^2 \\ + c_{\mathcal{C}} e^{2\eta_2 s} \left(\|y_\circ - \mathcal{P}^\ell y_\circ\|_H^2 + \frac{1}{\eta_1} \int_0^s \|\mathsf{res}^{\mathsf{pr}}(\tau)\|_{V'}^2 \,\mathrm{d}\tau \right) \mathrm{d}s \Bigg) \,\mathrm{d}t\end{aligned}$$

with $c_3 = c_{\mathcal{C}} + 2\eta_2$, $\eta_1 > 0$ and $\eta_2 \geq 0$ are the coercivity constants of a.

Proof. From the optimality system we have the relation

$$\bar{\mathfrak{u}}(t) = \frac{1}{\sigma} \mathcal{B}(t)^\star \bar{\mathfrak{p}}(t) \quad \text{and} \quad \bar{\mathfrak{u}}^\ell(t) = \frac{1}{\sigma} \mathcal{B}(t)^\star \bar{\mathfrak{p}}^\ell(t) \quad \text{in } H.$$

Consequently, we obtain

$$\|\bar{\mathfrak{u}} - \bar{\mathfrak{u}}^\ell\|_{\mathcal{U}} = \frac{1}{\sigma} \|\mathcal{B}^\star(\bar{\mathfrak{p}} - \bar{\mathfrak{p}}^\ell)\|_{\mathcal{U}},$$

where we can apply the dual error. So we obtain

$$\begin{aligned}\|\bar{\mathfrak{u}} - \bar{\mathfrak{u}}^\ell\|_{\mathcal{U}}^2 &= \frac{1}{\sigma^2} \|\mathcal{B}^\star(\bar{\mathfrak{p}} - \bar{\mathfrak{p}}^\ell)\|_{\mathcal{U}}^2 = \frac{1}{\sigma^2} \int_0^T \|\mathcal{B}(t)^\star(\bar{\mathfrak{p}}(t) - \bar{\mathfrak{p}}^\ell(t))\|_H^2 \,\mathrm{d}t \\ &\leq \frac{c_{\mathcal{B}}^2}{\sigma^2} \int_0^T \|\bar{\mathfrak{p}}(t) - \bar{\mathfrak{p}}^\ell(t)\|_H^2 \,\mathrm{d}t \\ &\leq \frac{c_{\mathcal{B}}^2}{\sigma^2} \int_0^T e^{c_3(T-t)} \left(\int_t^T \frac{1}{\eta_1} \|\mathsf{res}^{\mathsf{du}}(s)\|_{V'}^2 + c_{\mathcal{C}} \|e^{\mathsf{pr}}(s)\|_H^2 \,\mathrm{d}s \right) \mathrm{d}t \\ &\leq \frac{c_{\mathcal{B}}^2}{\sigma^2} \int_0^T e^{c_3(T-t)} \Bigg(\int_t^T \frac{1}{\eta_1} \|\mathsf{res}^{\mathsf{du}}(s)\|_{V'}^2 \\ &\quad + c_{\mathcal{C}} e^{2\eta_2 s} \left(\|y_\circ - \mathcal{P}^\ell y_\circ\|_H^2 + \frac{1}{\eta_1} \int_0^s \|\mathsf{res}^{\mathsf{pr}}(\tau)\|_{V'}^2 \,\mathrm{d}\tau \right) \mathrm{d}s \Bigg) \,\mathrm{d}t.\end{aligned}$$

□

With the a-priori error estimates we have the possibility to measure the performance of the reduced model without computing the full-order solutions. This is really important when choosing or updating the reduced basis.

4.1.2 The All-In-One Approach

The intention of the All-In-One approach is to solve $(\mathbf{P}_{\mathrm{PDE}})$ in one shot, without separating the state and adjoint equations. Unlike the separation approach, which captures the dominant dynamics of the two equations for different control inputs $\mathfrak{u}$, the All-In-One approach is focused on approximating the solution of the optimal control problem for the

optimal control $\bar{\mathfrak{u}}$.

Therefore, we introduce $X = V \times V$ and define $\mathfrak{x}(t) := (\mathfrak{y}(t), \mathfrak{p}(T-t)) \in X$ for almost all $t \in [0,T]$. Furthermore, let $X^\ell \subset X$ be a reduced space. Then the reduced problem reads as follow

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \mathfrak{x}_1^\ell(t), \varphi_1 \right\rangle_H + a(t; \mathfrak{x}_1^\ell(t), \varphi_1) = \left\langle \mathfrak{f}(t) + \frac{1}{\sigma}\mathcal{B}(t)\mathcal{B}(t)^\star \mathfrak{x}_2^\ell(T-t), \varphi_1 \right\rangle_{V',V} \quad t \in (0,T], \tag{4.7a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \mathfrak{x}_2^\ell(t), \varphi_2 \right\rangle_H + a(t; \varphi_2, \mathfrak{x}_2^\ell(t)) = \left\langle \tilde{\mathfrak{z}}_\mathsf{d}(t) - \tilde{\mathcal{C}}(t)\mathfrak{x}_1(T-t), \tilde{\mathcal{C}}(t)\varphi_2 \right\rangle_H \qquad t \in (0,T], \tag{4.7b}$$

for all $\varphi \in X^\ell$ and $\mathfrak{x}^\ell(0) = (\mathcal{P}^\ell \mathfrak{y}_\circ, 0) \in H \times H$. For practical reasons we choose $X^\ell \subset X^\mathsf{fe} = V^\mathsf{fe} \times V^\mathsf{fe}$ and therefore we a interested in a model-order reduction of the dynamical system (3.40). The reduced model is given by

$$\dot{x}^\ell(t) = \mathcal{A}^\ell(t)x^\ell(t) + \tilde{\mathcal{A}}^\ell(t)x^\ell(T-t) + \mathcal{F}^\ell(t) \text{ for } t \in (0,T), \quad x^\ell(0) = x_\circ^\ell, \tag{4.8}$$

where $x^\ell(t) \in \mathbb{R}^\ell$, $\mathcal{A}^\ell(t) \in \mathbb{R}^{\ell \times \ell}$ is the reduced operator of $\mathcal{A}(t) \in \mathbb{R}^{n_y \times n_y}$ and $\mathcal{F}^\ell(t) \in \mathbb{R}^\ell$ is the reduced right hand side of $\mathcal{F}(t) \in \mathbb{R}^\ell$. In doing so, we assume $\ell \ll n_y$.

What we aim to achieve with this approach is to efficiently compute the optimal solution of $(\mathbf{P}_{\mathrm{PDE}})$ for different values of its underlying parameters, such as the right hand side f or the desired state $\mathfrak{z}_\mathsf{d}$. This technique is particularly useful in scenarios where multiple optimal control problems with varying parameters need to be solved. As demonstrated in [29], this approach can be particularly beneficial in the context of multiobjective optimization or model predictive control (MPC), where multiple optimal control problems with changing parameters need to be solved. Additionally, this approach has also proven to be beneficial in the context of reduced MPC with error estimator. Next, we discuss the construction of the low-dimensional subspace X^ℓ that can capture the information contained in the given data.

4.2 Proper Orthogonal Decomposition

In this section, we introduce the *proper orthogonal decomposition* (POD), a well-known, data-driven technique for reducing the model of parameter-dependent optimal control problems. We provide a brief overview of the discrete POD version, which is particularly useful for numerical applications. However, a continuous POD version also exists. For a more in-depth description of POD for linear-quadratic optimal control problems, we refer the reader to [14, Chapter 2], for instance.

4.2.1 The (discrete) POD Method

The (discrete) POD method is based on a given set of vectors $\{y_j^k\}_{j=1}^n \subset \mathbb{R}^m$ for $1 \leq k \leq \wp$, the so-called *snapshots*, cf., e.g. [23] and [14]. In example, the snapshots can be represent already computed finite-dimensional solutions for different parameter. Let

$$\mathcal{V}^n = \operatorname{span}\left\{ y_j^k \,\middle|\, 1 \leq j \leq n \text{ and } 1 \leq k \leq \wp \right\} \subset \mathbb{R}^m \tag{4.9}$$

be the linear subspace spanned by the snapshots with dimension $n_{\mathcal{V}} = \dim \mathcal{V}^n \in \{1, ..., n\wp\} < \infty$. For $\ell \leq n_{\mathcal{V}}$ the POD method generates pairwise orthonormal functions $\{\psi_i\}_{i=1}^{\ell} \subset \mathbb{R}^m$, known as the *POD basis of rank* ℓ. The goal is for all y_j^k, $1 \leq j \leq n$ and $1 \leq k \leq \wp$ to be represented with sufficient accuracy by the POD basis, i.e. to approximate y_j^k's with a linear combination of the ψ's. This is achieved by minimizing the weighted mean square error between the y_j^k's and their corresponding ℓ-th partial Fourier sum:

$$\begin{cases} \min \sum_{k=1}^{\wp} \sum_{j=1}^{n} \alpha_j^n \left\| y_j^k - \sum_{i=1}^{\ell} \langle y_j^k, \psi_i \rangle_{\tilde{W}} \, \psi_i \right\|_{\tilde{W}}^2 \\ \text{s.t. } \{\psi_i\}_{i=1}^{\ell} \subset \mathbb{R}^m \text{ and } \langle \psi_i, \psi_j \rangle_{\tilde{W}} = \delta_{ij}, \ 1 \leq i, j \leq \ell, \end{cases} \tag{$\mathbf{P}_{\mathrm{POD}}^d$}$$

where $\tilde{W} \in \mathbb{R}^{m \times m}$ is symmetric and positiv definite, such that $\langle \cdot, \cdot \rangle_{\tilde{W}} := \langle W \cdot, \cdot \rangle_{\mathbb{R}^m}$ is an inner product, as well as the α_j^n's are positive weighting parameters for $j = 1, \ldots, n$. The symbol δ_{ij} denotes the Kronecker symbol satisfying $\delta_{ii} = 1$ and $\delta_{ij} = 0$ for $i \neq j$. Therefore, the POD basis of rank ℓ is an optimal solution to ($\mathbf{P}_{\mathrm{POD}}^d$).

Next, we discuss how to solve the constrained optimization problem ($\mathbf{P}_{\mathrm{POD}}^d$).

Definition and Theorem 4.6
Let $\{y_j^k\}_{j=1}^n \subset \mathbb{R}^m$ for $1 \leq k \leq \wp$ be discrete snapshots. Then the operator

$$\mathcal{R}^n : \mathbb{R}^m \to \mathcal{V}^n, \quad \psi \mapsto \sum_{k=1}^{\wp} \sum_{j=1}^{n} \alpha_j^n \langle y_j^k, \psi \rangle_{\tilde{W}} y_j^k$$

is linear, continuous, compact nonnegative and self-adjoint.

Proof. A proof is given in [14, Lemma 2.2]. □

Now, it can be proven that a solution of ($\mathbf{P}_{\mathrm{POD}}^d$) is given by eigenvectors of the operator $\mathcal{R}^n$. In detail we have the following theorem.

Theorem 4.7
Let $\{y_j^k\}_{j=1}^n \subset \mathbb{R}^m$ for $1 \leq k \leq \wp$ be discrete snapshots and $\mathcal{R}^n$ be given as in Definition and Theorem (4.6). Then there exist nonnegative eigenvalues $\{\bar{\lambda}_k^n\}_{k=1}^{n_{\mathcal{V}}}$ and associated orthonormal eigenvectors $\{\bar{\psi}_k^n\}_{k=1}^{n_{\mathcal{V}}} \subset \mathbb{R}^m$ satisfying

$$\mathcal{R}^n \bar{\psi}_k^n = \bar{\lambda}_k^n \bar{\psi}_k^n \quad \text{and} \quad \bar{\lambda}_1^n \geq ... \geq \bar{\lambda}_{n_{\mathcal{V}}}^n > 0 \tag{4.10}$$

for $k = 1, ..., n_{\mathcal{V}}$. For every $\ell \in \{1, ..., n_{\mathcal{V}}\}$ the first ℓ eigenfunctions $\{\bar{\psi}_k^n\}_{k=1}^{\ell}$ solve ($\mathbf{P}_{\mathrm{POD}}^d$). Furthermore, the following approximation error formula holds true:

$$\sum_{k=1}^{\wp} \sum_{j=1}^{n} \alpha_j^n \left\| y_j^k - \sum_{i=1}^{\ell} \langle y_j^k, \bar{\psi}_i^n \rangle_{\tilde{W}} \bar{\psi}_i^n \right\|_{\tilde{W}}^2 = \sum_{k=\ell+1}^{n_{\mathcal{V}}} \bar{\lambda}_k^n.$$

Proof. A proof is given in [14, Theorem 2.7]. □

The next question is how to solve the eigenvalue problem (4.10). Without loss of generality, we consider the general case of two sets of snapshots, i.e. $\wp = 2$. The extension for $\wp > 2$ is straightforward. Therefore, we introduce the snapshot matrix $Y = [y_1^1|...|y_n^1|y_1^2|...|y_n^2] \in \mathbb{R}^{m\times(2n)}$ with rank $n_{\mathcal{V}} \leq \min(m, 2n)$. Note that in general, we can assume $m < n\wp$, as we typically have a larger number of snapshot sets, resulting in a large value of $\wp$. Additionally, we define the diagonal weighting matrix $D = \operatorname{diag}(\alpha_1^n, ..., \alpha_n^n) \in \mathbb{R}^{n\times n}$. Then it holds

$$\begin{aligned}\mathcal{R}^n\psi &= \sum_{j=1}^{n}\left(\alpha_j^n\langle y_j^1, \psi\rangle_{\tilde{W}} y_j^1 + \alpha_j^n\langle y_j^2, \psi\rangle_{\tilde{W}} y_j^2\right)\\ &= Y\underbrace{\begin{pmatrix} D & 0\\ 0 & D\end{pmatrix}}_{=:\tilde{D}\in\mathbb{R}^{(np)\times(np)}} Y^T\tilde{W}\psi = Y\tilde{D}Y^T\tilde{W}\psi \quad \text{for } \psi\in\mathbb{R}^m.\end{aligned}$$

Hence, the eigenvalue problem (4.10) reads as

$$Y\tilde{D}Y^T\tilde{W}\bar{\psi}_k^n = \bar{\lambda}_k^n\bar{\psi}_k^n \quad \text{and} \quad \bar{\lambda}_1^n \geq ... \geq \bar{\lambda}_{n_{\mathcal{V}}}^n > 0. \tag{4.11}$$

If we set $\psi_k^n = \tilde{W}^{1/2}\bar{\psi}_k^n$ and multiplying (4.11) by $\tilde{W}^{1/2}$ from the left we get

$$\tilde{W}^{1/2}Y\tilde{D}Y^\top\tilde{W}^{1/2}\psi_k^n = \bar{\lambda}_k^n\psi_k^n. \tag{4.12}$$

Since $\tilde{W}$ and $\tilde{D}$ are symmetric and positive definite the matrix $\hat{Y} = \tilde{W}^{1/2}Y\tilde{D}^{1/2} \in \mathbb{R}^{m\times(n\wp)}$ is well-defined and the POD basis $\{\bar{\psi}_k^n\}_{k=1}^{\ell}$ of rank ℓ is given by the symmetric $m\times m$ eigenvalue problem

$$\hat{Y}\hat{Y}^\top\psi_k^n = \bar{\lambda}_k^n\psi_k^n \quad \text{for } 1\leq k\leq\ell \quad \text{and } \langle\psi_i^n, \psi_j^n\rangle_{\mathbb{R}^m} = \delta_{ij} \quad \text{for } 1\leq i,j\leq\ell \tag{4.13}$$

and $\bar{\psi}_k^n = \tilde{W}^{-1/2}\psi_k^n$ for $1\leq k\leq\ell$. Instead of solving (4.13) we can also obtain the POD basis by solving the $(n\wp)\times(n\wp)$ eigenvalue problem

$$\hat{Y}^\top\hat{Y}\phi_k^n = \bar{\lambda}_k^n\phi_k^n \quad \text{for } 1\leq k\leq\ell \quad \text{and } \langle\phi_i^n, \phi_j^n\rangle_{\mathbb{R}^m} = \delta_{ij} \quad \text{for } 1\leq i,j\leq\ell \tag{4.14}$$

and set

$$\bar{\psi}_k^n = \tilde{W}^{1/2}\psi_k^n = \frac{1}{(\bar{\lambda}_k^n)^{\frac{1}{2}}}\tilde{W}^{1/2}\hat{Y}\psi_k^n = \frac{1}{(\bar{\lambda}_k^n)^{\frac{1}{2}}}Y\tilde{D}^{1/2}\psi_k^n$$

for $1\leq k\leq\ell$. If one solves (4.13) or (4.14) depends on the dimension of the problem. Typically, one solves the eigenvalue problem with the lower dimension, which corresponds to less numerical effort. Eigenvalue problems of the form (4.13) and (4.14) can be solved by utilizing the *singular value decomposition* (SVD). For more details we refer to [32].

Definition and Theorem 4.8 (Singular value decomposition)
Let $Y\in\mathbb{R}^{m\times n}$ of rank $d\leq\min(m,n)$. Then there exists uniquely scalars $\sigma_1\geq\sigma_2\geq...\geq\sigma_d>0$ and orthogonal matrices $U\in\mathbb{R}^{m\times m}$ with columns $u_i\in\mathbb{R}^m$ for $1\leq i\leq m$ and $V\in\mathbb{R}^{n\times n}$ with columns $v_i\in\mathbb{R}^n$ for $1\leq i\leq n$ such that

$$U^\top YV = \begin{pmatrix} D & 0\\ 0 & 0\end{pmatrix} =: \Sigma\in\mathbb{R}^{m\times n},$$

where $D = \operatorname{diag}(\sigma_1, ..., \sigma_d)\in\mathbb{R}^{d\times d}$. Furthermore, the vectors $\{u_i\}_{i=1}^m$ and $\{v_i\}_{i=1}^n$ satisfy

$$Yv_i = \sigma_i u_i \quad \text{and} \quad Y^Tu_i = \sigma_i v_i \quad \text{for } i = 1, ..., d.$$

Proof. A proof is given in [32]. □

Corollary 4.9
Let $\hat{Y} \in \mathbb{R}^{m \times n_\wp}$ be the weighted snapshot matrix of rank $n_\mathcal{V} \leq \min(m, n_\wp)$ and $\hat{Y} \overset{SVD}{=} U\Sigma V^\top$ the singular value decomposition of $\hat{Y}$ with columns $\{u_i\}_{i=1}^m$ and $\{v_i\}_{i=1}^n$. Then it holds

(i) $\hat{Y}\hat{Y}^\top u_i = \sigma_i^2 u_i$ *for all* $1 \leq i \leq n_\mathcal{V}$.

(ii) $\hat{Y}^\top\hat{Y} v_i = \sigma_i^2 v_i$ *for all* $1 \leq i \leq n_\mathcal{V}$.

Proof. Applying Definition and Theorem (4.8) it holds for $1 \leq i \leq n_\mathcal{V}$

$$\hat{Y}\hat{Y}^\top u_i = \sigma_i \hat{Y} v_i = \sigma_i^2 u_i \quad \text{and} \quad \hat{Y}^\top\hat{Y} v_i = \sigma_i \hat{Y}^\top u_i = \sigma_i^2 v_i.$$

□

As a result, instead of solving one of the eigenvalue problems (4.13) or (4.14), we can compute a POD basis by computing the SVD of the weighted snapshot matrix $\hat{Y} \in \mathbb{R}^{m \times n_\wp}$. If the matrix $\hat{Y}$ is poorly scaled, it is advisable to avoid building the matrix product $\hat{Y}\hat{Y}^\top$ or $\hat{Y}^\top\hat{Y}$, as the SVD is more stable in this case. If we are interested in a POD basis of rank $\ell < n_\mathcal{V}$, we can also compute a *truncated singular value decomposition* (tSVD) with truncation value $\ell \in \mathbb{N}$, $\hat{Y} \overset{tSVD}{=} U_\ell \Sigma_\ell V_\ell^\top$. This involves computing only the largest ℓ singular values (the remaining singular values are set to zero), resulting in $U_\ell \in \mathbb{R}^{m \times \ell}$, $\Sigma_\ell = \operatorname{diag}(\sigma_1, ..., \sigma_\ell) \in \mathbb{R}^{\ell \times \ell}$ with $\sigma_1 \geq ... \geq \sigma_\ell > 0$ and $V_\ell \in \mathbb{R}^{n_\wp \times \ell}$. The matrix U_ℓ correspond to the POD basis of rank ℓ.

4.2.2 POD for Dynamical Systems

Next, we will explain how the discrete POD framework is used to perform model-order reduction for dynamical systems. Considering the nonlinear discrete input-output system of the form

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} = \mathcal{F}(t, y(t), u(t)) \qquad \text{for } t \in (0,T), \quad y(0) = y_\circ, \tag{4.15a}$$

$$z(t) = \mathcal{G}(y(t)) \qquad \text{for } t \in (0,T), \tag{4.15b}$$

where $y_\circ \in \mathbb{R}^{n_y}$ is the initial condition and $\mathcal{F} : [0,T] \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_y}$, $\mathcal{G} : \mathbb{R}^{n_y} \to \mathbb{R}^{n_z}$ are given functions. Moreover, we assume that we have access to a set of snapshots $\{y_j^k\}_{j=1}^n \subset \mathbb{R}^{n_y}$, $1 \leq k \leq \wp$, which can be an approximation of the state y at different time points for different controls u. Now, we can derive the reduced POD model in a standard way (e.g., as described in [1]).

Algorithm 1: (Discrete) POD MOR for input-output systems

Require: Set of snapshots $\{y_j^k\}_{j=1}^n \subset \mathbb{R}^{n_y}$ $(1 \le k \le \wp)$, dynamical system $(\mathcal{F}, \mathcal{G})$, truncation value $\ell \ge 1$, POD weighting matrices $\tilde{W} \in \mathbb{R}^{n_y \times n_y}$ and $D = \operatorname{diag}(\alpha_1^n, ..., \alpha_n^n) \in \mathbb{R}^{n_y \times n_y}$.

1: Set $Y = [y_1^1|...|y_n^1|y_1^2|...|y_n^2|...|y_1^\wp|...|y_n^\wp] \in \mathbb{R}^{n_y \times (n\wp)}$, $\tilde{D} = \operatorname{diag}(D, ..., D) \in \mathbb{R}^{(n\wp)\times(np)}$;
2: Compute $\hat{Y} = \tilde{W}^{1/2} Y \tilde{D}^{1/2} \in \mathbb{R}^{n_y \times (n\wp)}$;
3: Compute the truncated SVD $\hat{Y} \overset{tSVD}{=} U_\ell \Sigma_\ell V_\ell^\top$ with truncation value $\ell \ll n_y$;
4: Approximate $y(t) \approx U y^\ell(t)$, where $y^\ell(t) \in \mathbb{R}^\ell$ solves together with $z^\ell(t) \in \mathbb{R}^{n_z}$

$$\begin{aligned} \dot{y}^\ell(t) &= U^\top \mathcal{F}(t, U y^\ell(t), u(t)) && \text{for } t \in (0,T), \quad y^\ell(0) = U^\top y_\circ, \\ z^\ell(t) &= \mathcal{G}(U y^\ell(t)) && \text{for } t \in (0,T); \end{aligned}$$

4.3 Gramian Model-Order Reduction

In this section, we explain the concepts of empirical gramians and how they can be used to perform model-order reduction for nonlinear time-variant input-output systems of the form

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} = \mathcal{F}(t, y(t), u(t)) \qquad \text{for } t \in (0,T), \quad y(0) = y_\circ, \tag{4.17a}$$

$$z(t) = \mathcal{G}(y(t)) \qquad \text{for } t \in (0,T), \tag{4.17b}$$

where $y_\circ \in \mathbb{R}^{n_y}$ is the initial condition and $\mathcal{F} : [0,T] \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_y}$, $\mathcal{G} : \mathbb{R}^{n_y} \to \mathbb{R}^{n_z}$ are the given dynamic. We mainly follow the work [16, 25] and assume $T = \infty$. In addition, we assume that (4.17) is exponentially stable, meaning that for a given control $u \in L^\infty(0,\infty;\mathbb{R}^{n_u})$, the state and output of the system satisfy $y \in L^\infty(0,\infty;\mathbb{R}^{n_y})$ and $z \in L^\infty(0,\infty;\mathbb{R}^{n_z})$ (as described in [25]). We first introduce the well-known gramian setting for linear time-invariant systems, and then extend it to nonlinear systems.

4.3.1 Gramians for LTI systems

Firstly, we consider the LTI system

$$\begin{aligned} \dot{y}(t) &= Ay(t) + Bu(t) && \text{for } t \in (0,\infty), \quad y(0) = y_\circ, \\ z(t) &= Cy(t) && \text{for } t \in (0,\infty), \end{aligned} \tag{4.18}$$

where $A \in \mathbb{R}^{n_y \times n_y}$, $B \in \mathbb{R}^{n_y \times n_u}$ and $C \in \mathbb{R}^{n_z \times n_y}$. For the following section we suppose that the LTI system (4.18) is stable, i.e. all eigenvalues of A have strictly negative real part; cf. [49].

Definition 4.10 (Controllability gramian)
The linear controllability gramian *is defined as the symmetric matrix*

$$L_c := \mathcal{C}\mathcal{C}^* = \int_0^\infty e^{At} B B^\top e^{A^\top t} \,\mathrm{d}t = \int_0^\infty \left(e^{At}B\right)\left(e^{At}B\right)^\top \mathrm{d}t \in \mathbb{R}^{n_y \times n_y},$$

where $\mathcal{C} : L^2(0,\infty) \to \mathbb{R}^{n_y}$, $u \mapsto \int_0^\infty e^{At} B u(t) \,\mathrm{d}t$ *is the (linear)* controllability operator.

The controllability gramian indicates how well the state y of an underlying LTI system is driven by the control u.

Definition 4.11 (Observability gramian)
The linear observability gramian *is given as the symmetric matrix*

$$L_o := \mathcal{O}^*\mathcal{O} = \int_0^\infty e^{A^\top t} C^\top C e^{At}\,\mathrm{d}t = \int_0^\infty \left(e^{A^\top t}C^\top\right)\left(e^{A^\top t}C^\top\right)^\top \mathrm{d}t \in \mathbb{R}^{n_y \times n_y},$$

where $\mathcal{O} : \mathbb{R}^{n_y} \to L^2(0,\infty;\mathbb{R}^{n_z})$, $x \mapsto Ce^{At}x$ *is the (linear and bounded)* observability operator.

The observability gramian indicates how well a change in the state y of an underlying LTI system is recognizable in the output. The next lemma shows, that the gramians L_c and L_o can be computed by solving (linear) matrix equations.

Lemma 4.12
The controllability gramian L_c *is a positive semidefinite solution to the* Lyapunov equation

$$AL_c + L_cA^\top + BB^\top = 0. \tag{4.19}$$

If $Im(\mathcal{C}) = \mathbb{R}^{n_y}$, *then* (4.18) *is controllable and* (4.19) *admits a unique solution* L_c *which has full rank* n_y *and is positive definite. Moreover, the observability gramian* L_o *is a positive semidefinite solution to the Lyapunov equation*

$$A^\top L_o + L_oA + C^\top C = 0. \tag{4.20}$$

If $\ker(\mathcal{O}) = \{0\}$, *then* (4.18) *is observable and system* (4.20) *admits a unique solution* L_o. *Moreover,* L_o *is positive definite.*

Proof. To get an idea of the proof, we show the assertion for the controllability gramian L_c if $\mathrm{Im}(\mathcal{C}) = \mathbb{R}^{n_y}$. The rest of the proof is given in [21].
For the Lyapunov equation we obtain

$$\begin{aligned} AL_c + L_cA^\top &= \int_0^\infty Ae^{At}BB^\top e^{A^\top t}\,\mathrm{d}t + \int_0^\infty e^{At}BB^\top e^{A^\top t}A^\top\,\mathrm{d}t \\ &= \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}t}(e^{At}BB^\top e^{A^\top t})\,\mathrm{d}t \\ &= e^{At}BB^\top e^{A^\top t}\Big|_{t=0}^{\infty} \\ &= -BB^\top, \end{aligned}$$

where we used that $e^{At} \to 0$ as $t \to \infty$ for stable A (all eigenvalues of A have negative real part). Since $BB^\top$ is a symmetric matrix, L_c is symmetric as well.
For the uniqueness let L_c^1 and L_c^2 be two solutions of (4.19). Then we have

$$A(L_c^1 - L_c^2) + (L_c^1 - L_c^2)A^\top = 0.$$

Multiplying by e^{At} by the left and $e^{A^\top t}$ by the right and afterwards integrating over $(0,\infty)$ leads to

$$\begin{aligned} 0 &= \int_0^\infty e^{At}(A(L_c^1 - L_c^2) + (L_c^1 - L_c^2)A^\top)e^{A^\top t}\,\mathrm{d}t \\ &= \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}t}(e^{At}(L_c^1 - L_c^2)e^{A^\top t})\,\mathrm{d}t \\ &= L_c^1 - L_c^2, \end{aligned}$$

where we used again that $e^{At} \to 0$ as $t \to \infty$ for stable A. For the positive definiteness of L_c let $y \in \mathbb{R}^{n_y}$ be arbitrary. Indeed, it holds

$$y^\top L_c y = \int_0^\infty y^\top e^{At} B B^\top e^{A^\top t} y \, \mathrm{d}t = \int_0^\infty \|B^\top e^{A^\top t} y\|_2^2 \, \mathrm{d}t > 0.$$

□

The central idea behind the gramians is that the matrices L_c and L_o contain essential information about which states of the dynamical system can be controlled or observed. This information can be used for model-order reduction. For example, using balanced truncation, almost uncontrollable and unobservable states are neglected; cf. [3]. The so-called *cross gramian matrix* quantifies the controllability and observability in one linear operator. It is required that $n_u = n_z$, and the cross gramian matrix is defined as

$$L_x := \int_0^\infty e^{At} B C e^{At} \, \mathrm{d}t = \int_0^\infty \left(e^{At} B\right)\left(e^{A^\top t} C^\top\right)^\top \mathrm{d}t \in \mathbb{R}^{n_y \times n_y}.$$

4.3.2 Empirical Gramians

In the previous Section 4.3.1 we discussed the representation of gramians for linear time-invariant (LTI) systems. However, these representations cannot be applied to nonlinear systems that vary over time. To address this limitation, we introduce an extension to the classical gramians, the so called *empirical gramians*. Unlike the classical gramians, the empirical gramians are data-driven and do not rely on the structure of LTI systems. They can be computed from measured or simluated data of the dynamic system, and were first introduced in [31] and later extended in [25]. The main idea behind empirical gramians is to average local gramians over varying controls, initial states, parameters or any other time-dependent components. With this in mind, we begin with the following definition:

Definition 4.13 (Gramian perturbation sets)
For given $k \in \mathbb{N}$ let $I_k \in \mathbb{R}^{k \times k}$ be the identity matrix. Define the following sets:

$$\begin{aligned}
\mathfrak{T}^k &:= \left\{ T_1, \dots, T_{n_\mathfrak{T}} \,\middle|\, T_i \in \mathbb{R}^{k \times k},\, T_i^\top T_i = I_k,\, i = 1, \dots, n_\mathfrak{T} \right\} \subset \mathbb{R}^{k \times k}, \\
\mathfrak{M} &:= \left\{ c_1, \dots, c_{n_\mathfrak{M}} \,\middle|\, c_i \in \mathbb{R},\, c_i > 0 \text{ for } i = 1, \dots, n_\mathfrak{M} \right\} \subset \mathbb{R}, \\
\mathfrak{E}^k &:= \left\{ e_1, \dots, e_k \,\middle|\, \text{standard unit vectors in } \mathbb{R}^k \right\} \subset \mathbb{R}^k.
\end{aligned}$$

Here $\mathfrak{T}^k$ is an arbitrary set of $n_\mathfrak{T}$ orthogonal matrices, and $\mathfrak{M}$ is a set of $n_\mathfrak{M}$ positive constants. Furthermore, given $n \in \mathbb{N}$ and a function $w \in L^\infty(0, \infty; \mathbb{R}^n)$ we define the mean $\bar{w} \in \mathbb{R}^n$ as

$$\bar{w} := \lim_{T \to \infty} \frac{1}{T} \int_0^T w(t) \, \mathrm{d}t \in \mathbb{R}^n$$

provided the limit exists.

Now we are in a position to define the empirical gramians.

Definition 4.14 (Empirical controllability gramian)
Given nonempty sets $\mathcal{T}^{n_u}$, $\mathcal{M}$ and $\mathcal{E}^{n_u}$. For (4.17) *the* empirical controllability gramian *$\widehat{L}_c \in \mathbb{R}^{n_y \times n_y}$ is defined as*

$$\widehat{L}_c := \sum_{i=1}^{n_\mathcal{T}} \sum_{j=1}^{n_\mathcal{M}} \sum_{l=1}^{n_u} \frac{1}{c_j^2 n_\mathcal{T} n_\mathcal{M}} \int_0^\infty Y^{ijl}(t)\,\mathrm{d}t$$

with

$$Y^{ijl}(t) = \left(y^{ijl}(t) - \bar{y}^{ijl}\right)\left(y^{ijl}(t) - \bar{y}^{ijl}\right)^\top,$$

where $y^{ijl}(t) \in \mathbb{R}^{n_y}$ solves (4.17) *corresponding to the impulse input $u(t) = c_j T_i e_l \delta(t)$ and $\bar{y}^{ijl}$ stands for its mean.*

Definition 4.15 (Empirical observability gramian)
Given nonempty sets $\mathcal{T}^{n_y}$, $\mathcal{M}$ and $\mathcal{E}^{n_y}$. For (4.17) *the* empirical observability gramian *$\widehat{L}_o \in \mathbb{R}^{n_y \times n_y}$ is defined as*

$$\widehat{L}_o := \sum_{i=1}^{n_\mathcal{T}} \sum_{j=1}^{n_\mathcal{M}} \frac{1}{c_j^2 n_\mathcal{T} n_\mathcal{M}} \int_0^\infty T_i Z^{ij}(t) T_i^\top \,\mathrm{d}t$$

with

$$Z^{ij}_{\nu\mu}(t) = \left(z^{ij\nu}(t) - \bar{z}^{ij\nu}\right)^\top \left(z^{ij\mu}(t) - \bar{z}^{ij\mu}\right),$$

where $z^{ij\nu}(t) \in \mathbb{R}^{n_z}$, $\nu = 1, \ldots, n_y$, is the output of (4.17) *corresponding to the initial condition $y_\circ = c_j T_i e_\nu$ and $\bar{z}^{ij\nu}$ denotes its mean.*

With the following lemma we see the motivation behind the empirical gramians. The empirical gramian are equal to the usual gramians for stable LTI systems.

Lemma 4.16
For any nonempty sets $\mathcal{T}^{n_u}$, $\mathcal{T}^{n_y}$, $\mathcal{E}^{n_u}$, $\mathcal{E}^{n_y}$ and $\mathcal{M}$, the empirical gramians $\widehat{L}_c$ and $\widehat{L}_o$ of the stable linear system $\dot{x}(t) = Ax(t) + Bu(t)$, $z(t) = Cx(t)$ is equal to the usual gramians L_c and L_o.

Proof. A proof is given in [25, Lemma 5 and Lemma 7]. □

Given the symmetric empirical gramians $\widehat{L}_c$ and $\widehat{L}_o$, it is possible to define a reduced-order model through a balancing transformation which is based on the SVD. We follow the approach proposed in [12]. First we compute the SVD of the symmetric matrices $\widehat{L}_c$ and $\widehat{L}_o$:

$$\widehat{L}_c = U_c \Sigma_c U_c^\top \in \mathbb{R}^{n_y \times n_y} \quad \text{and} \quad \widehat{L}_o = U_o \Sigma_o U_o^\top \in \mathbb{R}^{n_y \times n_y} \tag{4.21a}$$

with orthogonal matrices U_c, $U_o \in \mathbb{R}^{n_y \times n_y}$ and diagonal matrices Σ_c, $\Sigma_o \in \mathbb{R}^{n_y \times n_y}$ containing the nonnegative singular values in descending order. Then we can compute the matrices $\widehat{L}_c^{1/2}$ and $\widehat{L}_o^{1/2}$ as well as their product

$$\widehat{L}_{co}^{1/2} := \widehat{L}_c^{1/2} \widehat{L}_o^{1/2} = U_c \Sigma_c^{1/2} U_c^\top U_o \Sigma_o^{1/2} U_o^\top \in \mathbb{R}^{n_y \times n_y}. \tag{4.21b}$$

Finally we derive the SVD of $\widehat{L}_{co}^{1/2}$:

$$\widehat{L}_{co}^{1/2} = U\Sigma V^\top \tag{4.21c}$$

with orthogonal matrices U, $V \in \mathbb{R}^{n_y \times n_y}$ and a diagonal matrix $\Sigma \in \mathbb{R}^{n_y \times n_y}$ containing the nonnegative singular values σ_i, $i = 1, \ldots, n_y$, of $\widehat{L}_{co}^{1/2}$ in descending order.
The next step is to create a simplified version of the model, known as reduced-order modeling (ROM). We use a method called balanced truncation to achieve this. To construct a ROM, we remove the singular values that are smaller than a certain threshold ε, keeping only $\ell \ll n_y$ values so that $\sigma_i < \varepsilon$ holds for $i = \ell + 1, \ldots, n_y$. The SVD is organized in the following way to accomplish this.

$$\widehat{L}_c^{1/2}\widehat{L}_o^{1/2} = U\Sigma V^\top = \begin{bmatrix} U_1 \,|\, U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1 \,|\, V_2 \end{bmatrix}^\top = \begin{pmatrix} U_1\Sigma_1 V_1^\top & 0 \\ 0 & U_2\Sigma_2 V_2^\top \end{pmatrix},$$

where $U_1 \in \mathbb{R}^{n_y \times \ell}$, $U_2 \in \mathbb{R}^{n_y \times (n_y - \ell)}$ and $\Sigma_1 \in \mathbb{R}^{\ell \times \ell}$, $\Sigma_2 \in \mathbb{R}^{(n_y-\ell)\times(n_y-\ell)}$. Now the reduced-order model for (4.17) is derived in a standard way (cf., e.g., [1]): For given large terminal time $T > 0$ and for $t \in [0, T]$ we approximate $y(t) \in \mathbb{R}^{n_y}$ by $U_1 y^\ell(t)$, where $y^\ell(t) \in \mathbb{R}^\ell$ solves together with $z^\ell(t) \in \mathbb{R}^{n_z}$

$$\dot{y}^\ell(t) = V_1^\top \mathcal{F}(t, U_1 y^\ell(t), u(t)) \text{ for } t \in (0, T], \quad y^\ell(0) = V_1^\top y_\circ, \tag{4.22a}$$

$$z^\ell(t) = \mathcal{G}(U_1 y^\ell(t)) \qquad \text{for } t \in [0, T]. \tag{4.22b}$$

Algorithm 2: Gramian-based MOR for input-output systems

Require: Sets $\mathcal{T}^{n_{u_{\mathsf{inp}}}}$, $\mathcal{T}^{n_y}$, $\mathcal{M}$, $\mathcal{E}^{n_{u_{\mathsf{inp}}}}$, $\mathcal{E}^{n_y}$ (cf. Definition 4.13), truncation value $\ell \geq 1$, arbitrary input function u_{inp}, arbitrary output function z_{out}.

1: Compute empirical controllability gramian $\widehat{L}_c \in \mathbb{R}^{n_y \times n_y}$ for (4.17) with inputs

$$\tilde{u}_{\mathsf{inp}}(t) = c_j T_i e_l u_{\mathsf{inp}}(t), \quad j = 1, \ldots, n_{\mathcal{M}},\ i = 1, \ldots, n_{\mathcal{T}},\ l = 1, \ldots, n_{u_{\mathsf{inp}}},\ t \in [0, T],$$

where $c_j \in \mathcal{M}$, $T_i \in \mathcal{T}^{u_{\mathsf{inp}}}$, $e_l \in \mathcal{E}^{u_{\mathsf{inp}}}$.

2: Compute empirical observability gramian $\widehat{L}_o \in \mathbb{R}^{n_y \times n_y}$ for (4.17) with the output z_{out} and random initial guesses

$$x_\circ = c_j T_i e_\nu, \quad j = 1, \ldots, n_{\mathcal{M}},\ i = 1, \ldots, n_{\mathcal{T}},\ \nu = 1, \ldots, n_y$$

where $c_j \in \mathcal{M}$, $T_i \in \mathcal{T}^{n_y}$, $e_\nu \in \mathcal{E}^{n_y}$.

3: Compute the balancing transformation $\widehat{L}_{co}^{1/2} \in \mathbb{R}^{n_y \times n_y}$.

4: Compute truncated SVD of $\widehat{L}_{co}^{1/2} = U_\ell \Sigma_\ell V_\ell^\top$ with truncation value $\ell \in \mathbb{N}$.

5: Approximate $y(t) \approx U y^\ell(t)$, where $y^\ell(t) \in \mathbb{R}^\ell$ solves together with $z^\ell(t) \in \mathbb{R}^{n_z}$

$$\dot{y}^\ell(t) = V_1^\top \mathcal{F}(t, U_1 y^\ell(t), u(t)) \qquad \text{for } t \in (0, T), \quad y^\ell(0) = U^\top y_\circ,$$
$$z^\ell(t) = \mathcal{G}(U_1 y^\ell(t)) \qquad \text{for } t \in (0, T);$$

Remark 4.17 (Computation costs of the empirical gramians)
The primary cost in the creation of empirical Gramians is the calculation of trajectories. However, it is often possible to reduce the number of trajectories by using information about the underlying model or operating area. Additionally, the trajectories can be computed in parallel, which can help to reduce the overall cost.

4.4 Extended Dynamic Mode Decomposition

Dynamic Mode Decomposition (DMD) is a powerful, data-driven technique to identify time-invariant high dimensional dynamical systems by a reduced linearized model [24, 39]. DMD is not a typical MOR technique, but it can interpret as one. For a detailed description of the standard DMD method, we refer to the bachelor thesis [37, Chapter 3]. This section is based on the book [24]. We begin this section introducing the main idea of DMD. The initial setting for DMD is a high dimensional dynamical system of the form

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} = \mathcal{F}(t, y(t)) \quad \text{for } t \in (0,T), \quad y(0) = y_\circ, \tag{4.24}$$

where $y(t) \in \mathbb{R}^{n_y}$ represents the state of our dynamical system at time t, $\mathcal{F} : [0,T] \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_y}$ describes the dynamic behavior and $y_\circ \in \mathbb{R}^{n_y}$ is a given initial condition. Instead of POD and the empirical gramian framework, DMD works in an equation-free perspective, where no knowledge about the the dynamical system $\mathcal{F}$ is provided (it can be even unknown). What DMD needs is a set of given snapshots $\{y_j^k\}_{j=1}^n \subset \mathbb{R}^{n_y}$, $1 \le k \le \wp$, that represents measurements of the continuous state $y(t)$ at certain time instances $0 = t_0 < ... < t_n \le T$, i.e., $y_j^k \approx y(t_j)$ for $k = 0, ..., n$, $k = 1, ..., \wp$. The goal is then to construct a linear dynamical system

$$\frac{\mathrm{d}\tilde{y}(t)}{\mathrm{d}t} = \mathcal{A}\tilde{y}(t) \quad \text{for } t \in (0,T), \quad y(0) = y_\circ \tag{4.25}$$

with appropriate matrices $\mathcal{A} \in \mathbb{R}^{n_y \times n_y}$, so that the solution $\tilde{y}$ to (4.25) approximates the given set of data with best-fit in a least square sense. In a discrete time setting with fixed stepsize Δt problem (4.25) is replaced by the discrete system

$$y_{k+1} = Ay_k \quad \text{for } k \ge 0, \quad y_0 = y_\circ,$$

where $A = \exp(\mathcal{A}\Delta t)$. Hence, the DMD method consists in finding a matrix $A \in \mathbb{R}^{n_y \times n_y}$ so that

$$A = \arg\min_{\tilde{A}} \|Y_1 - \tilde{A}Y_0\|_F$$

for the matrices

$$Y_0 = [y_0^1|...|y_{n-1}^1|...|y_0^\wp|...|y_{n-1}^\wp] \in \mathbb{R}^{n_y \times (n\wp)}, \quad Y_1 = [y_1^1|...|y_n^1|...|y_1^\wp|...|y_n^\wp] \in \mathbb{R}^{n_y \times (n\wp)}.$$

The function $\|\cdot\|_F$ denotes the *Frobenius* norm defined as

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2} \quad \text{for } A \in \mathbb{R}^{m \times n}.$$

The modeling of a complex, high dimensional, dynamical system often requires an external input or control $u(t) \in \mathbb{R}^{n_u}$ and to observe certain quantities of interest. In this case, one is interested in studying the dynamical system with external control

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} = \mathcal{F}(t, y(t), u(t)) \quad \text{for } t \in (0,T), \quad y(0) = y_\circ, \tag{4.26}$$

where $\mathcal{F} : [0,T] \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_y}$ and which has the discrete formulation

$$y_{k+1} = \tilde{\mathcal{F}}(t_k, y_k, u_k) \quad \text{for } k \geq 0, \quad y_0 = y_\circ. \tag{4.27}$$

In contrast to the DMD framework, the *Extended DMD* (EDMD) utilizes the snapshots of the dynamic, inputs (controls) and extra outputs (observables) to extract the underlying dynamics of (4.26), cf. [22, 39]. Let $\psi : \mathbb{R}^{n_y} \to \mathbb{R}^\ell$ be a vector of lifting functions (or observables) $\psi_i : \mathbb{R}^{n_y} \to \mathbb{R}$, i.e.

$$\psi(y) = \begin{pmatrix} \psi_1(y) \\ \vdots \\ \psi_\ell(y) \end{pmatrix}.$$

Moreover, let

$$Y_0 = [y_0^1|...|y_{n-1}^1|...|y_0^\wp|...|y_{n-1}^\wp] \in \mathbb{R}^{n_y \times (n\wp)}, \quad Y_1 = [y_1^1|...|y_n^1|...|y_1^\wp|...|y_n^\wp] \in \mathbb{R}^{n_y \times (n\wp)}.$$

and

$$U = [u_0^1|...|u_{n-1}^1|...|u_0^\wp|...|u_{n-1}^\wp] \in \mathbb{R}^{n_u \times (n\wp)}.$$

Now, we consider the lifted snapshot matrices

$$Y_{0,\text{lift}} = \left[\begin{array}{ccc|c|ccc} \psi_1(y_0^1) & \dots & \psi_1(y_{n-1}^1) & \dots & \psi_1(y_0^\wp) & \dots & \psi_1(y_{n-1}^\wp) \\ \vdots & & \vdots & & \vdots & & \vdots \\ \psi_\ell(y_0^1) & \dots & \psi_\ell(y_{n-1}^1) & \dots & \psi_1(y_0^\wp) & \dots & \psi_1(y_{n-1}^\wp) \end{array}\right] \in \mathbb{R}^{\ell \times (n\wp)}$$

and

$$Y_{1,\text{lift}} = \left[\begin{array}{ccc|c|ccc} \psi_1(y_1^1) & \dots & \psi_1(y_n^1) & \dots & \psi_1(y_1^\wp) & \dots & \psi_1(y_n^\wp) \\ \vdots & & \vdots & & \vdots & & \vdots \\ \psi_\ell(y_1^1) & \dots & \psi_\ell(y_n^1) & \dots & \psi_1(y_1^\wp) & \dots & \psi_1(y_n^\wp) \end{array}\right] \in \mathbb{R}^{\ell \times (n\wp)}.$$

The EDMD consist in identifying the matrices $A \in \mathbb{R}^{\ell \times \ell}$, $B \in \mathbb{R}^{\ell \times n_u}$ and $C \in \mathbb{R}^{n_y \times \ell}$ such that

$$[A, B] = \arg\min_{\tilde{A}, \tilde{B}} \|Y_{1,\text{lift}} - \tilde{A} Y_{0,\text{lift}} - \tilde{B} U\|_F, \tag{4.28a}$$

$$C = \arg\min_{\tilde{C}} \|Y_0 - \tilde{C} Y_{0,\text{lift}}\|_F. \tag{4.28b}$$

In this way, we obtain a discrete input-output linear dynamical system in the observable space

$$y_{k+1}^{\ell} = Ay_k^{\ell} + Bu_k \quad \text{for } k \geq 0, \quad y_0^{\ell} = \begin{pmatrix} \psi_1(y_0^1) \\ \vdots \\ \psi_{\ell}(y_0^1) \end{pmatrix}, \tag{4.29a}$$

$$\hat{y}_{k+1} = Cy_{k+1}^{\ell} \quad \text{for } k \geq 0. \tag{4.29b}$$

Note that the analytical solution of (4.28a) is $[A, B] = Y_{1,\text{lift}}[Y_{0,\text{lift}}, U]^{\dagger}$, where $A^{\dagger}$ denotes the Moore-Penrose pseudoinverse of a matrix A, for more details see [22]. Numerically, this solution can be computed by the singular value decomposition (SVD); cf. Algorithm 3. Similar, $C = Y_0 Y_{0,\text{lift}}^{\dagger}$ solves (4.28b).
The solution $\{\hat{y}_k\}_{k\geq 0}$ of (4.29) is an approximation of the original state y at the time points t_k for $k \geq 0$, which solves the dynamical system (4.24). If we choose the set of lifting functions such that dimension of the image space is much smaller than the dimension of original dynamical system, i.e. $\ell \ll n_y$, EDMD can be also interpret as a model-order reduction technique. Indeed, EDMD provides an extraction of the original dynamical system (4.24) by a lower dimensional input-output system (4.29), which is much more efficient to solve. Since EDMD also performs a linearization of the original problem, we can not expect the best results, especially if we deal with nonlinear dynamical systems.

Algorithm 3: EDMD algorithm

Require: Set of snapshots $\{y_j^k\}_{j=1}^n \subset \mathbb{R}^{n_y}$ and $\{u_j^k\}_{j=1}^n \subset \mathbb{R}^{n_y}$ $(1 \leq k \leq \wp)$, set of lifting functions $\psi : \mathbb{R}^{n_y} \to \mathbb{R}^\ell$.

1: Set

$$Y_0 = [y_0^1|...|y_{n-1}^1|...|y_0^\wp|...|y_{n-1}^\wp] \in \mathbb{R}^{n_y \times (n\wp)}, Y_1 = [y_1^1|...|y_n^1|...|y_1^\wp|...|y_n^\wp] \in \mathbb{R}^{n_y \times (n\wp)},$$
$$U = [u_0^1|...|u_{n-1}^1|...|u_0^\wp|...|u_{n-1}^\wp] \in \mathbb{R}^{n_u \times (n\wp)};$$

2: Compute the matrices

$$Y_{0,\text{lift}} = [\psi(y_0^1)|...|\psi(y_{n-1}^1)|...|\psi(y_0^\wp)|...|\psi(y_{n-1}^\wp)] \in \mathbb{R}^{\ell \times (n\wp)},$$
$$Y_{1,\text{lift}} = [\psi(y_1^1)|...|\psi(y_n^1)|...|\psi(y_1^\wp)|...|\psi(y_n^\wp)] \in \mathbb{R}^{\ell \times (n\wp)};$$

3: Set $r = \text{rank } Y_{0,\text{lift}} \leq \min\{\ell, n\wp\}$;

4: Compute the (truncated) SVD of

$$\begin{bmatrix} Y_{0,\text{lift}} \\ U \end{bmatrix} = W_\mathfrak{r} \Sigma_\mathfrak{r} V_\mathfrak{r}^\top \quad \text{with } \mathfrak{r} \leq \text{rank} \begin{bmatrix} Y_{0,\text{lift}} \\ U \end{bmatrix} \leq \min\{\ell + n_u, n\wp\}$$

$W_\mathfrak{r} \in \mathbb{R}^{(\ell+n_u)\times \mathfrak{r}}$, $\Sigma_\mathfrak{r} = \text{diag}(\sigma_1, ..., \sigma_\mathfrak{r}) \in \mathbb{R}^{\mathfrak{r}\times\mathfrak{r}}$ with $\sigma_1 \geq ... \geq \sigma_\mathfrak{r} > 0$ and $V_\mathfrak{r} \in \mathbb{R}^{(n\wp)\times\mathfrak{r}}$;

5: Split $W_\mathfrak{r}$ in two separate components $W_{\mathfrak{r},1} \in \mathbb{R}^{\ell\times\mathfrak{r}}$ and $W_{\mathfrak{r},2} \in \mathbb{R}^{n_u\times\mathfrak{r}}$;

6: Determine the matrices

$$A = Y_{1,\text{lift}} V_\mathfrak{r} \Sigma_\mathfrak{r}^{-1} W_{\mathfrak{r},1}^\top \in \mathbb{R}^{\ell\times\ell}, \quad B = Y_{1,\text{lift}} V_\mathfrak{r} \Sigma_\mathfrak{r}^{-1} W_{\mathfrak{r},2}^\top \in \mathbb{R}^{\ell\times n_u}$$

7: Compute the (truncated) SVD of $Y_{0,\text{lift}} = W_r \Sigma_r V_r^\top$ with $W_r \in \mathbb{R}^{\ell\times r}$, $\Sigma_r \in \mathbb{R}^{r\times r}$ and $V \in \mathbb{R}^{(n\wp)\times r}$;

8: Compute the matrix $C = Y_0 V_r \Sigma_r^{-1} W_r^\top \in \mathbb{R}^{n_y\times\ell}$;

9: Approximate $y(t_k) \approx \tilde{y}_k$ for $k = 0, ..., n$, where $\hat{y}$ solves the discrete EDMD input-output system (4.29) with the computed matrices A, B, C;

5 | Numerical Results

In this chapter, we examine a particular example of an optimal control problem that was introduced in Chapter 3. We use numerical methods, specifically model predictive control (MPC), to solve this problem. Therefore, we begin the chapter by providing a brief introduction to MPC. The primary goal of this chapter is to demonstrate the effectiveness of the model-order reduction techniques for a reduced MPC with error estimator, and to highlight the pros and cons of the different techniques.

5.1 Model Predictive Control

Model Predictive Control (MPC) is an optimization-based method for the feedback control of dynamical systems, where the time horizon usually taken to be infinite, as described in [13]. The main idea behind MPC is to iteratively perform optimal control on a moving finite horizon. MPC was developed from the theory of optimal control and was first proposed for discrete-time linear systems by Propoi [34] in the early 1960s. One of the main benefits of MPC compared to conventional control strategies is that it allows the optimization of a dynamic system at the current time point while keeping the future states into account.
For large terminal time $T \gg 0$ we are considering problems of the form

$$\begin{aligned} &\min_{(y,u)} \mathcal{J}(y,u) = \int_0^T \Phi(t, y(t), u(t))\,\mathrm{d}t \\ &\text{s.t. } (y,u) \in H^1(0,T;\mathbb{R}^{n_y}) \times L^2(0,T;\mathbb{R}^{n_u}) \text{ solves} \\ &\dot{y}(t) = \mathcal{F}(t, y(t), u(t)) \text{ for } t \in (0,T], \quad y(0) = y_\circ, \end{aligned} \tag{MPC}$$

where $\Phi : [0,T] \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_u} \to \mathbb{R}_0^+$ is a given running cost function and $\mathcal{F} : [0,T] \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_y}$ represents the considered dynamic with initial guess $y_\circ \in \mathbb{R}^{n_y}$.

Instead of solving (MPC) directly for the final time T, MPC is based on solving many subproblems on a smaller time horizon. To do this, we set a *horizon prediction parameter* $T_{pred} > 0$, which gives us a prediction of the system's behavior at the current time point t. Specifically, at time t, we compute an optimal control $\bar{u}_n$ by minimizing the cost function over the interval $[t, t + T_{pred}]$. The desired feedback is obtained by implementing the optimal control $\bar{u}_n$ over the interval $[t, t + T_f]$. Here, $0 < T_f \leq T_{pred}$ is the *feedback parameter*. The procedure is then repeated starting from the new resulting current state at time $t + T_f$, yielding to a new optimal control $\bar{u}_{n+1}$, which contributes to another piece of the final reconstructed suboptimal control for the whole horizon. In summary,

we are solving iteratively optimal control problems, where the finite prediction horizon keeps being shifted forward. The feedback control is obtained because each new computed optimal control for a given time window depends on the current state of the system. To conclude we state the following MPC algorithm.

Algorithm 4: Model predictive control (MPC)

Require: MPC prediction horizon $T_{\mathsf{pred}} > 0$, MPC feedback horizon $T_{\mathsf{f}} > 0$, initial data $y_\circ \in H^1(0,T;\mathbb{R}^{n_y})$, $t_0 = 0$.

1: **for** $n = 0, 1, 2, \ldots$ **do**

2: Set sampling time $t_n = t_0 + nT_{\mathsf{f}}$ and measure the state $\tilde{y}_\circ = y(t_n) \in H^1(0,T;\mathbb{R}^{n_y})$;

3: Solve the optimal control problem

$$\min \int_{t_n}^{t_n+T_{\mathsf{pred}}} \Phi(t, y(t), u(t))\,\mathrm{d}t$$
$$\text{s.t.} \quad \dot{y}(t) = \mathcal{F}(t, y(t), u(t)) \text{ for } t \in (t_n, t_n + T_{\mathsf{pred}}], \quad y(t_n) = \tilde{y}_\circ;$$

4: Store the optimal control $\bar{u}$ over the time interval $[t_n, t_n + T_{\mathsf{f}})$ and use this control value in the next sampling period;

5: **end for**

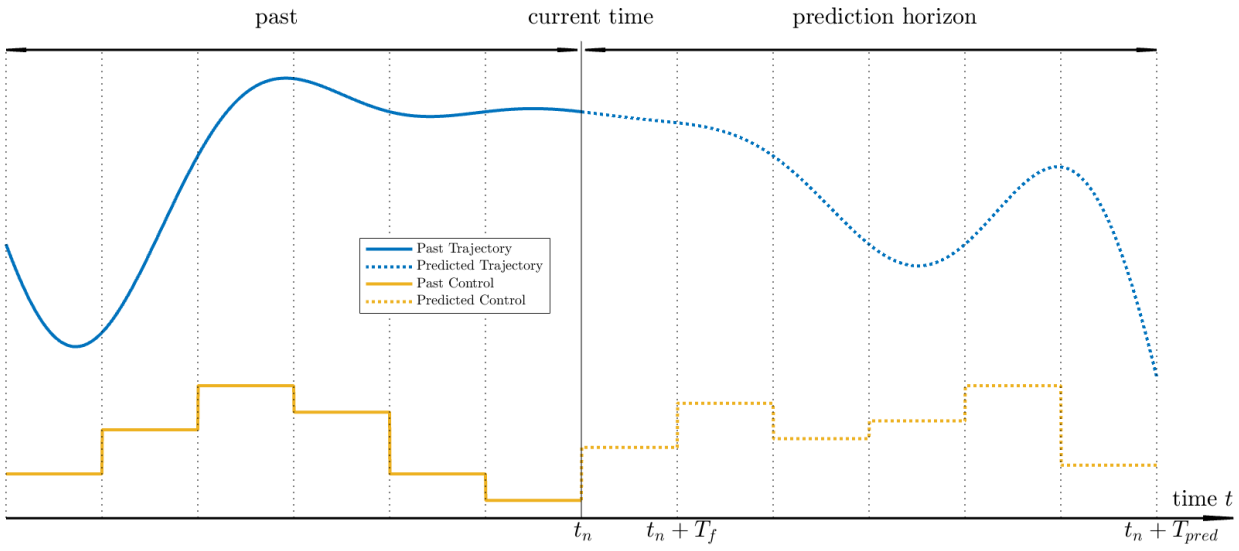


Figure 5.1: Illustration of the MPC step at time t_n.

5.2 Numerical MPC Example

In the following all notations from the previous chapters are used. To quickly recall, let $\Omega \subset \mathbb{R}^n$ be an open and bounded Lipschitz domain, $T > 0$. We set

$$\Omega_T := (0,T) \times \Omega \quad \text{and} \quad \Sigma_T := (0,T) \times \partial\Omega.$$

Furthermore, we define the function spaces $V := H^1(\Omega)$, $H := L^2(\Omega)$, the control space $\mathcal{U} := L^2(0,T;H)$, the state space $\mathcal{Y} := W(0,T)$ and the measurement space by $\mathcal{H} := L^2(0,T;H)$ In this section we study the specific problem of $(\mathbf{P}_{\mathrm{PDE}})$:

$$\min_{\mathfrak{y},\mathfrak{u}} \mathcal{J}(\mathfrak{y},\mathfrak{u}) = \frac{1}{2}\int_0^T \|\mathfrak{y}(t) - \mathfrak{z}_{\mathsf{d}}(t)\|_H^2 + \sigma \|\mathfrak{u}(t)\|_H^2 \,\mathrm{d}t \tag{$\mathbf{P}$}$$

subject to $(\mathfrak{y},\mathfrak{u}) \in \mathcal{Y} \times \mathcal{U}$ solve the linear parabolic advection-diffusion equation

$$\begin{aligned} \frac{\partial \mathfrak{y}}{\partial t}(t,\boldsymbol{x}) - \Delta\mathfrak{y}(t,\boldsymbol{x}) + \alpha(t)\boldsymbol{v}(\boldsymbol{x}) \cdot \nabla\mathfrak{y}(t,\boldsymbol{x}) &= \mathcal{B}(t)\mathfrak{u}(t,\boldsymbol{x}) + \mathfrak{f}(t,\boldsymbol{x}), && (t,\boldsymbol{x}) \in \Omega_T, \\ \frac{\partial \mathfrak{y}}{\partial n}(t,\boldsymbol{s}) &= 0, && (t,\boldsymbol{s}) \in \Sigma_T, \\ \mathfrak{y}(0,\boldsymbol{x}) &= \mathfrak{y}_\circ(\boldsymbol{x}), && \boldsymbol{x} \in \Omega, \end{aligned} \tag{$\mathbf{C}$}$$

where $v \in L^\infty(\Omega;\mathbb{R}^n)$, $\alpha \in L^\infty(0,T)$ is a scalar-valued parameter function, $\mathfrak{f} \in L^2(0,T;H)$, $\mathfrak{y}_\circ \in H$ and $\mathfrak{z}_{\mathsf{d}} \in \mathcal{H}$. Moreover, the operator $\mathcal{B}$ is given as

$$\mathcal{B}(t)\mathfrak{u}(t,\boldsymbol{x}) = \begin{cases} \chi_{\omega_1}(\boldsymbol{x})\mathfrak{u}(t,\boldsymbol{x}) & \text{for } t \le t_{\mathsf{switch}}, \\ \chi_{\omega_2}(\boldsymbol{x})\mathfrak{u}(t,\boldsymbol{x}) & \text{for } t > t_{\mathsf{switch}}, \end{cases}$$

where χ_{ω_i} is the characteristic function of a nonzero subset $\omega_i \subset \Omega$ for $i \in \{1,2\}$ and $t_{\mathsf{switch}} \in (0,T)$.
The bilinear form $a(t;\cdot,\cdot)$ has the specific form

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

and the weak form of $(\mathbf{C})$ reads as follow

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\langle \mathfrak{y}(t),\varphi\rangle_H + a(t;\mathfrak{y}(t),\varphi) &= \langle \mathfrak{f}(t) + \mathcal{B}(t)\mathfrak{u}(t),\varphi\rangle_{V',V} && \text{for all } \varphi \in V \text{ and } t \in (0,T], \\ \mathfrak{y}(0) &= \mathfrak{y}_\circ && \text{in } H. \end{aligned} \tag{5.2}$$

Then, for almost all $t \in [0,T]$ we have

$$\begin{aligned} \big|a(t;\varphi,\phi)\big| &\le \|\varphi\|_V\|\phi\|_V + \|\alpha(t)\|_{L^\infty(0,T)} \int_\Omega \Big|\big(\boldsymbol{v}(\boldsymbol{x}) \cdot \nabla\varphi(\boldsymbol{x})\big)\phi(\boldsymbol{x})\Big| \,\mathrm{d}\boldsymbol{x} \\ &\le \|\varphi\|_V\|\phi\|_V + \|\alpha(t)\|_{L^\infty(0,T)}\|\boldsymbol{v}\|_{L^\infty(\Omega;\mathbb{R}^n)}\|\varphi\|_V\|\phi\|_V \\ &\le \gamma\,\|\varphi\|_V\|\phi\|_V, \end{aligned} \tag{5.3}$$

where we set $\gamma = 1 + \|\boldsymbol{v}\|_{L^\infty(\Omega;\mathbb{R}^n)}\|\alpha(t)\|_{L^\infty(0,T)}$. Moreover, we get by applying Young's inequality

$$\begin{aligned} a(t;\varphi,\varphi) &\geq \|\varphi\|_V^2 - \|\boldsymbol{v}\|_{L^\infty(\Omega;\mathbb{R}^n)}\|\alpha(t)\|_{L^\infty(0,T)}\|\varphi\|_V\|\varphi\|_H \\ &\geq \|\varphi\|_V^2 - \left(\frac{1}{2}\|\varphi\|_V^2 + \frac{1}{2}\|\boldsymbol{v}\|_{L^\infty(\Omega;\mathbb{R}^n)}^2\|\alpha(t)\|_{L^\infty(0,T)}^2\|\varphi\|_H^2\right) \\ &= \eta_1\|\varphi\|_V^2 - \eta_2\|\varphi\|_H^2 \end{aligned} \tag{5.4}$$

with $\eta_1 = 1/2$ and $\eta_2 = \|\boldsymbol{v}\|_{L^\infty(\Omega;\mathbb{R}^n)}^2\|\alpha(t)\|_{L^\infty(0,T)}^2/2$. Thus, the bilinear form a is continuous and coercive. Consequently, for every $\mathfrak{u} \in \mathcal{U}$ there exists a unique solution $\mathfrak{y} \in \mathcal{Y}$, which solves (5.2). Accordingly and together with Theorem 3.12 the problem (**P**)-(**C**) is well defined.

5.2.1 Description of the Implementation

In this subsection, we provide all the details about the numerical implementation. All the following implementations are done in MATLAB. We used geometry functions from MATLAB's PDE-toolbox, and all other calculations were implemented by hand. All computations are carried out on a standard laptop (Apple MacBook Air, with an Apple M1 processor and 8GB of RAM). We begin this subsection by selecting all the necessary parameters and function values for the construction of a permissible test problem. Afterwards, we discuss the discretization of (**P**)-(**C**).

Optimization and PDE Parameter

To reduce computational effort, we consider problem (**P**)-(**C**) only in two dimensions. Therefore, we choose the spatial domain $\Omega = (0,1)^2 \subset \mathbb{R}^2$ and the final time $T = 100$. The goal is to choose a large enough final time so that it makes sense to use an MPC framework.
All other parameter and function values for (**P**)-(**C**) can be looked up in Table 5.1.

Domain parameters	
Spatial domain	$\Omega = (0,1)^2$
End time point	$T = 100$
Optimization parameters	
Cost parameter	$\sigma = 10^{-4}$
Desired state	$\mathfrak{z}_{\mathsf{d}}(t,\boldsymbol{x}) = 0.1 \cdot t \sin(0.1tx)\sin(2\pi x_1)\cos(2\pi x_2)$
PDE parameters	
Advection function	$\alpha(t) = \sin(t)$
Advection velocity	$\boldsymbol{v}(\boldsymbol{x}) = \big(-x_1 - x_2, (x_1+x_2)/2\big)^\top$
Right hand side	$f(t,\boldsymbol{x}) = x_1 x_2 \sin(\pi t)$
Initial data	$\mathfrak{y}_\circ(\boldsymbol{x}) = x_1^2 + x_2^2$
Control domain	$\omega_1 = [0.0.5]^2$ and $\omega_2 = [0, 0.9]^2$
Switch parameter	$t_{\mathsf{switch}} = 50$

Table 5.1: Domain, optimization and PDE data for the numerical implementation of (**P**)-(**C**).

In Figure 5.2 we plot the initial data $\mathfrak{y}_\circ$, the desired state $\mathfrak{z}_{\mathsf{d}}(100,\boldsymbol{x})$ and the advection velocity $\boldsymbol{v}$.

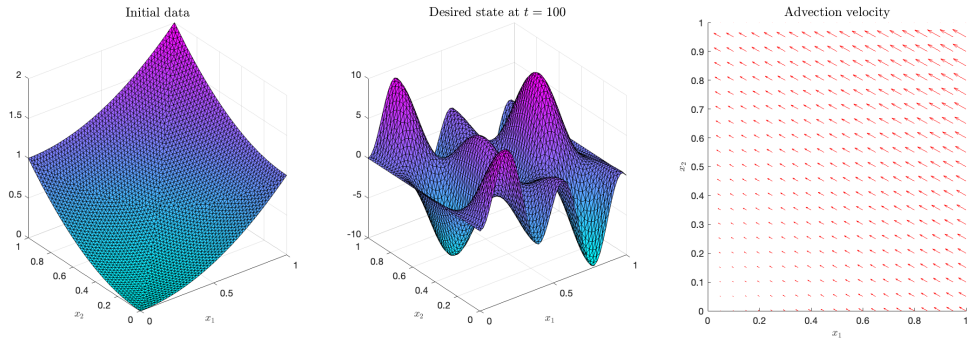


Figure 5.2: Visualization of the initial data $\mathfrak{y}_\circ$ (left), the desired state $\mathfrak{z}_{\mathsf{d}}(100,\boldsymbol{x})$ (middle) and the advection velocity $\boldsymbol{v}$ (right).

Discretization

To solve problem (**P**)-(**C**), we solve the optimality system directly, which has derived in Section 3.2 and is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \mathfrak{y}(t),\varphi\right\rangle_H + a(t;\mathfrak{y}(t),\varphi) = \left\langle \mathfrak{f}(t)+\mathcal{B}(t)\mathfrak{u}(t),\varphi\right\rangle_H, \qquad t\in(0,T], \tag{5.5a}$$

$$-\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \mathfrak{p}(t),\varphi\right\rangle_H + a(t;\varphi,\mathfrak{p}(t)) = \left\langle \mathfrak{z}_{\mathsf{d}}(t)-\mathfrak{y}(t),\varphi\right\rangle_H, \qquad t\in[0,T), \tag{5.5b}$$

$$\int_0^T \left\langle \mathfrak{u}(t),\phi(t)\right\rangle_H \mathrm{d}t = \int_0^T \left\langle \frac{1}{\sigma}\mathcal{B}(t)^\star\mathfrak{p}(t),\phi(t)\right\rangle_H \mathrm{d}t, \qquad t\in(0,T) \tag{5.5c}$$

for all $\varphi\in V$ and $\phi\in\mathcal{U}$ with $\mathfrak{y}(0)=\mathfrak{y}_\circ$ and $\mathfrak{p}(T)=0$. Therefore, we aim to find a suitable discretization for the optimality system in (5.5). To achieve this, we use the method of lines (cf., e.g., [38]), which involves discretizing only the spatial derivatives and leaving the time variable continuous. This results in a system of ordinary differential equations, which we then solve using an implicit Euler scheme.

Finite Element

For the spatial discretization, we briefly review the most important definitions that have been mention in Subsections 3.2.1 and 3.2.2: Let $\mathcal{T}_h$ be a triangulation of the spatial domain Ω, where $\bar{\Omega}=\bigcup_{\mathcal{T}\in\mathcal{T}_h}\bar{\mathcal{T}}$. In our implementation, we set the maximum edge length of one element to $h_{\mathsf{max}}=0.06$, resulting in $n_y=362$ nodes and 654 elements.

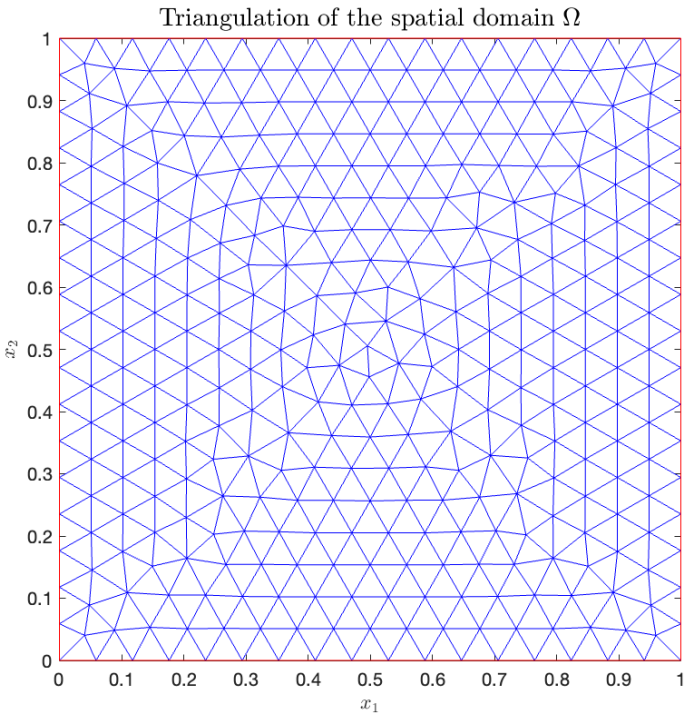


Figure 5.3: Triangulation of the spatial domain Ω with $h_{\mathsf{max}}=0.06$, $n_y=362$ nodes and 654 elements.

In the following let $m = n_y = n_u$ be valid. Furthermore, we use a piecewise linear FE space

$$V^{\mathsf{fe}} := \{v_h \in C^0(\bar{\Omega}) : v_h|_T \in P_1(T) \text{ for all } T \in \mathcal{T}_h\} = \operatorname{span}(\varphi_1, ..., \varphi_m) \subset V.$$

For $(t, \boldsymbol{x}) \in \Omega_T$ we approximate the state and control as

$$\mathfrak{y}(t, \boldsymbol{x}) \approx \mathfrak{y}^{\mathsf{fe}}(t, \boldsymbol{x}) = \sum_{i=1}^{m} y_i(t)\varphi_i(\boldsymbol{x}), \quad \mathfrak{u}(t, \boldsymbol{x}) \approx \mathfrak{u}^{\mathsf{fe}}(t, \boldsymbol{x}) = \sum_{i=1}^{m} u_i(t)\varphi_i(\boldsymbol{x}), \tag{5.6}$$

respectively and the adjoint as

$$\mathfrak{p}(t, \boldsymbol{x}) \approx \mathfrak{p}^{\mathsf{fe}}(t, \boldsymbol{x}) = \sum_{i=1}^{m} p_i(t)\varphi_i(\boldsymbol{x}).$$

Moreover, let

$$\mathfrak{z}_{\mathsf{d}}(t, \boldsymbol{x}) \approx \sum_{i=1}^{m} z_{\mathsf{d}i}(t)\varphi_i(\boldsymbol{x}), \ \mathfrak{f}(t, \boldsymbol{x}) \approx \sum_{i=1}^{m} f_i(t)\varphi_i(\boldsymbol{x}), \ \mathfrak{y}_\circ(\boldsymbol{x}) \approx \sum_{i=1}^{m} y_{\circ i}\varphi_i(\boldsymbol{x}) \tag{5.7}$$

for $(t, \boldsymbol{x}) \in \Omega_T$. We assemble the vectors $y(t) := (y_1(t), ..., y_m(t))^\top \in \mathbb{R}^m$, $u(t) := (u_1(t), ..., u_m(t))^\top \in \mathbb{R}^m$, $p(t) := (p_1(t), ..., p_m(t)) \in \mathbb{R}^m$, $f(t) := (f_1(t), ..., f_m(t))^\top \in \mathbb{R}^m$, $z_{\mathsf{d}}(t) := (z_{\mathsf{d}1}(t), ..., z_{\mathsf{d}m}(t))^\top \in \mathbb{R}^m$ and $y_\circ := (y_{\circ 1}, ..., y_{\circ m})^\top \in \mathbb{R}^m$. The FE matrices are given by

$$M \in \mathbb{R}^{m\times m} \quad M_{i,j} = \int_\Omega \varphi_i(\boldsymbol{x})\varphi_j(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \qquad \text{for } 1 \le i, j \le m, \tag{5.8a}$$

$$A(t) \in \mathbb{R}^{n_y\times n_y} \quad A(t)_{i,j} = a(t; \varphi_i, \varphi_j) \qquad \text{for } 1 \le i, j \le m, \tag{5.8b}$$

$$B(t) \in \mathbb{R}^{n_y\times n_u} \quad B(t)_{i,j} = \langle \mathcal{B}(t)\varphi_j, \varphi_i\rangle_H \qquad \text{for } 1 \le i, j \le m. \tag{5.8c}$$

Then, the semi-discrete optimality system reads as follows

$$\dot{x}(t) = \mathcal{A}(t)x(t) + \tilde{\mathcal{A}}(t)x(T-t) + \mathcal{F}(t) \text{ for } t \in (0, T), \quad x(0) = x_\circ \tag{5.9}$$

with

$$x(t) = \begin{bmatrix} y(t) \\ p(T-t) \end{bmatrix} \in \mathbb{R}^{2m}, \quad \mathcal{F}(t) = \begin{bmatrix} f(t) \\ \tilde{z}_{\mathsf{d}}(t) \end{bmatrix} \in \mathbb{R}^{2m}, \quad x_\circ = \begin{bmatrix} y_\circ \\ 0 \end{bmatrix} \in \mathbb{R}^{2m}.$$

and the two $(2m) \times (2m)$-matrices

$$\mathcal{A}(t) = \begin{bmatrix} -M^{-1}A(t) & 0 \\ 0 & -M^{-1}\tilde{A}(t)^\top \end{bmatrix}, \ \tilde{\mathcal{A}}(t) = \begin{bmatrix} 0 & \frac{1}{\sigma}M^{-1}B(t)M^{-1}B(t)^\top \\ -I_m & 0 \end{bmatrix},$$

where $I_m \in \mathbb{R}^{m\times m}$ is the identity. Remind, the tilde operators are the time-shifted operators, i.e. $\tilde{A}(t) = A(T-t)$. Moreover, we have $u(t) = M^{-1}B(t)^\top p(t)/\sigma$.

Implicit Euler scheme

Next, we will discretize the dynamical system (5.9) in time using an implizit Euler scheme. Therefore, let $N \in \mathbb{N}$, $\Delta t = \frac{T}{N}$, $t_k = k\Delta t$ for $k = 0, ..., N$ and $x_k = x(t_k)$ for $k = 0, ..., N$. In our implementation, we use $\Delta t = 0.04$, which implies $N = 2500$. Using the implicit Euler approach we obtain the following:

$$\frac{x_k - x_{k-1}}{\Delta t} = \mathcal{A}(t_k)x_k + \tilde{\mathcal{A}}(t_k)x_{N-k} + \mathcal{F}(t_k) \quad \text{for } k = 1, ..., N, \tag{5.10}$$

which is equivalent to

$$(I_{2m} - \Delta t\mathcal{A}(t_k))x_k - x_{k-1} - \Delta t\tilde{\mathcal{A}}(t_k)x_{N-k} = \Delta t\mathcal{F}(t_k) \quad \text{for } k = 1, ..., N. \tag{5.11}$$

For simplicity we use in the following notation $I = I_{2m} \in \mathbb{R}^{2m\times 2m}$. If we assemble the vector $x = (x_1, ..., x_N)^\top \in \mathbb{R}^N$, the $(2mN) \times (2mN)$ matrix

$$\mathcal{A} := \begin{bmatrix} I - \Delta t\mathcal{A}(t_1) & & & & & & -\Delta t\tilde{\mathcal{A}}(t_1) \\ -I & I - \Delta t\mathcal{A}(t_2) & & & -\Delta t\tilde{\mathcal{A}}(t_2) & & \\ & -I & \ddots & & & & \\ & & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & \ddots & & \ddots & \ddots & & \\ -\Delta t\tilde{\mathcal{A}}(t_{N-1}) & & & & -I & I - \Delta t\mathcal{A}(t_{N-1}) & \\ & & & & & -I & I - \Delta t\mathcal{A}(t_N) \end{bmatrix}$$

and the right-hand side

$$\mathcal{F} = \Delta t \begin{bmatrix} \mathcal{F}(t_1) + x_0 \\ \mathcal{F}(t_2) \\ \vdots \\ \mathcal{F}(t_{N-1}) \\ \mathcal{F}(t_N) + \tilde{\mathcal{A}}(t_N)x_0 \end{bmatrix} \in \mathbb{R}^{2mN},$$

System (5.11) can be written as

$$\mathcal{A}x = \mathcal{F}. \tag{5.12}$$

Assumption 5.1 (Uniqueness of a solution of the discrete system)
The matrix $\mathcal{A} \in \mathbb{R}^{(2mN)\times(2mN)}$ is regular. Therefore, the system (5.12) has for every right-hand side $\mathcal{F} \in \mathbb{R}^{2mN}$ a unique solution $\bar{x} \in \mathbb{R}^{2mN}$.

We justify Assumption 5.1 with the following argument: Theorem 3.13 states a unique solution $(\bar{\mathfrak{y}}, \bar{\mathfrak{u}})$ of the continuous optimality system. The assumption is based on the fact that the discrete optimality system approximates the continuous one as $\Delta t, h_{\mathsf{max}} \to 0$. Therefore, it is likely to be shown that the optimality systems with corresponding solutions converge in a appropriate norm. This turns out to be a challenging task.

5.2.2 Full Model Results

For the numerical results we solve the optimal control problem (**P**)-(**C**) using the MPC scheme. Our objective is to compare different MPC parameters and point out the respective benefits.
As a first test, we set $\omega_1 = \omega_2 = \Omega$, which means we have full control over the PDE throughout the entire domain Ω. Using the resulting operator $\mathcal{B}$, we get the following results. We can observe that in this setting, the MPC can push the state $\mathfrak{y}$ close to the

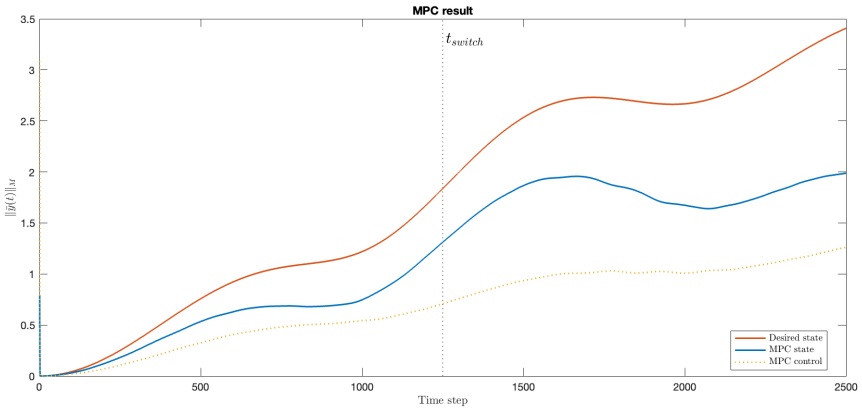


Figure 5.4: MPC result for problem (**P**)-(**C**) with $\omega_1 = \omega_2 = \Omega$. For the visualization we plot the three functions $t \mapsto \|\bar{y}(t)\|_M$, $t \mapsto \|y_d(t)\|_M$ and $t \mapsto 10^{-2} \times \|\bar{u}(t)\|_M$, where the tuple $(\bar{y}, \bar{u})$ denotes the MPC solution with $T_{\mathsf{f}} = 2$ and $T_{\mathsf{pred}} = 4$.

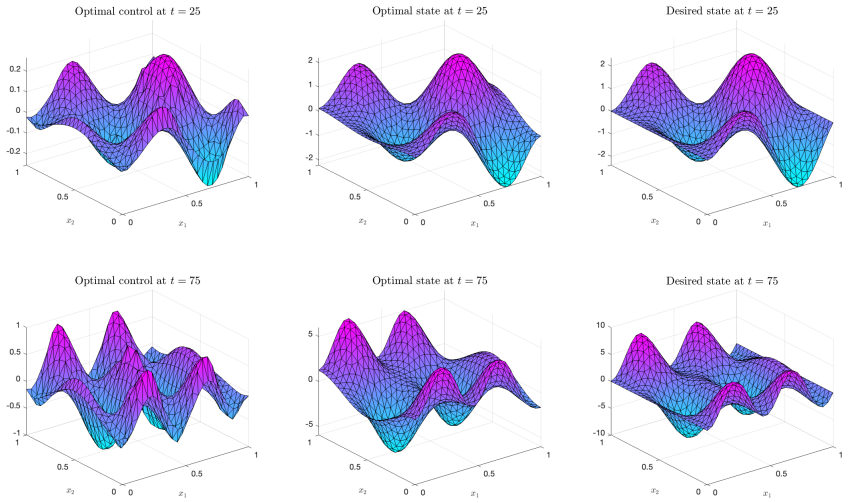


Figure 5.5: MPC result for problem (**P**)-(**C**) with $T_{\mathsf{f}} = 2$ and $T_{\mathsf{pred}} = 4$. The optimal control $\bar{\mathfrak{u}}$ (left), optimal state $\bar{\mathfrak{y}}$ and the desired state $\mathfrak{z}_{\mathsf{d}}$ at the time point $t = 25$ (top line) and $t = 75$ (bottom line).

target state $\mathfrak{z}_{\mathsf{d}}$, but cannot reach it exactly due to the constraint dynamics.

Next, we fix the the operator $\mathcal{B}$ as in Table 5.1, i.e. $\omega_1 = [0, 0.5]^2$ and $\omega_2 = [0, 0.9]^2$. Now, we cannot expect the same level of performance as in Figure 5.4, since we do not have control over the entire domain. Indeed, we can see in Figure 5.6 that if we have access

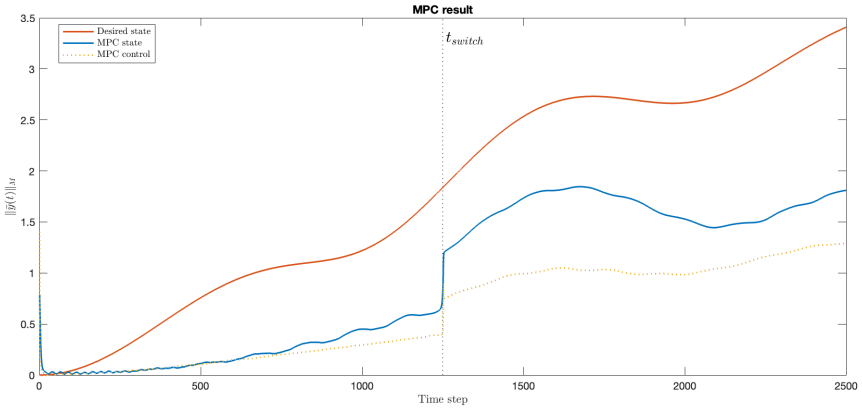


Figure 5.6: MPC result for problem (**P**)-(**C**) with $\omega_1 = [0, 0.5]^2$ and $\omega_2 = [0, 0.9]^2$. For the visualization we plot the three functions $t \mapsto \|\bar{y}(t)\|_M$, $t \mapsto \|y_d(t)\|_M$ and $t \mapsto 10^{-2} \times \|\bar{u}(t)\|_M$, where the tuple $(\bar{y}, \bar{u})$ denotes the MPC solution with $T_\mathsf{f} = 2$ and $T_\mathsf{pred} = 4$.

to only a small control subdomain, i.e. ω_1, it becomes quite difficult to bring the state to the target state $\mathfrak{z}_\mathsf{d}$. However, if we exceed the time point t_switch and can control the PDE on the larger subdomain ω_2, we can bring the state closer to the target from the start.

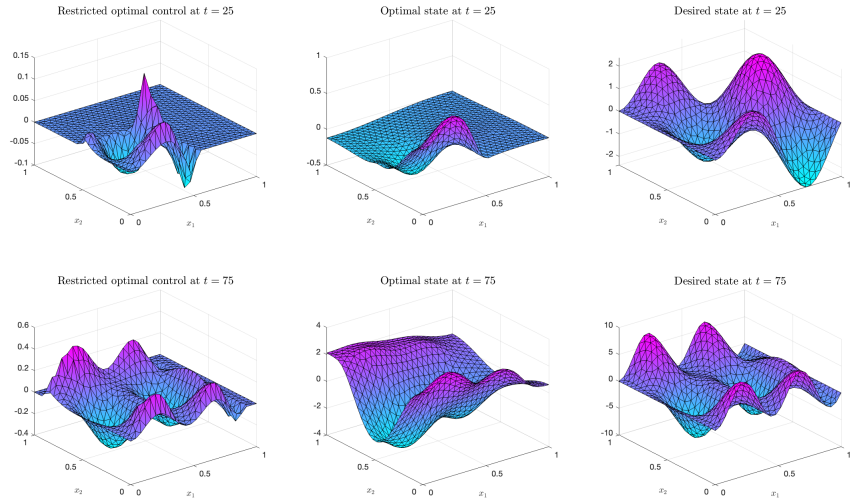


Figure 5.7: MPC result for problem (**P**)-(**C**) with $T_\mathsf{f} = 2$ and $T_\mathsf{pred} = 4$. The restricted optimal control $\mathcal{B}\bar{\mathfrak{u}}$ (left), optimal state $\bar{\mathfrak{y}}$ and the desired state $\mathfrak{z}_\mathsf{d}$ at the time point $t = 25$ (top line) and $t = 75$ (bottom line).

Now, let us discuss how to choose the feedback and prediction parameter T_f and T_pred in our example. To do this, we fix $(\tilde{y}, \tilde{u})$ as the MPC solution of (**P**)-(**C**) for $T_\mathsf{f} = 2$ and

$T_{\mathsf{pred}} = 4$. Moreover, let (y, u) be the MPC solution of (**P**)-(**C**) for different parameter T_{f} and T_{pred}. We define the absolute L^2-difference to $(\tilde{y}, \tilde{u})$ and the FE desired state y_{d} as

$$\mathsf{diff}^{y}_{\mathsf{abs}} = \left(\int_0^T \|\tilde{y}(t) - y(t)\|_M^2 \, \mathrm{d}t \right)^{1/2}, \qquad \mathsf{diff}^{u}_{\mathsf{abs}} = \left(\int_0^T \|\tilde{u}(t) - u(t)\|_M^2 \, \mathrm{d}t \right)^{1/2}$$

and

$$\mathsf{diff}^{y_{\mathsf{d}}}_{\mathsf{abs}} = \left(\int_0^T \|y_{\mathsf{d}}(t) - \bar{y}(t)\|_M^2 \, \mathrm{d}t \right)^{1/2}.$$

The time integrals are realized numerical by applying a standard trapezoidal approximation. The results are shown in Table 5.2. The computation time is an average value after

T_{pred}	T_{f}	MPC steps	Computation time	$\mathsf{diff}^{y}_{\mathsf{abs}}$	$\mathsf{diff}^{u}_{\mathsf{abs}}$	$\mathsf{diff}^{y_{\mathsf{d}}}_{\mathsf{abs}}$
2	1	100	297s	9.3×10^{-6}	5.8×10^{-3}	13.28
3	1	100	680s	2.6×10^{-8}	1.4×10^{-6}	13.28
3	2	50	342s	7.4×10^{-5}	3.9×10^{-3}	13.28
4	2	50	475s	0	0	13.28
5	4	25	339s	5.0×10^{-5}	2.7×10^{-3}	13.28
10	5	20	740s	2.6×10^{-8}	1.4×10^{-6}	13.28
30	25	4	1222s	2.6×10^{-8}	1.4×10^{-6}	13.28

Table 5.2: MPC results for varying feedback and prediction parameter T_{f} and T_{pred}.

5 runs. An increase int the prediction horizon T_{pred} takes into account more future states and therefore the resulting linear system becomes larger, which is directly related to an increase in computational effort. This can be observed in the computational time in Table 5.2. Since all results yields similar outcomes, we fix the prediction parameter to $T_{\mathsf{pred}} = 2$ and the feedback parameter $T_{\mathsf{f}} = 1$. In all tests, we observe a significant computation time to solve the full problem. The next question is how we can reduce the computational effort, while still maintaining a sufficient level of approximation quality.

5.2.3 Reduced Model Results

In this section, we briefly explain how we combine the different model-order reduction techniques, as shown in Chapter 4, with the MPC approach for the optimal control problem (**P**)-(**C**). Firstly, we will introduce a MPC strategy with a reduced basis update strategy. Then, we apply the MOR techniques to the reduced MPC strategy, starting with POD and the empirical gramian approach, followed by the (E)DMD.

Error Measures and reduced MPC with a-posteriori error estimator

In the following experiments, we aim to investigate the quality of solutions and the computational effort of the different reduced models. The quantities $\mathsf{err}^{y}_{\mathsf{abs}}$, $\mathsf{err}^{u}_{\mathsf{abs}}$, $\mathsf{err}^{y}_{\mathsf{rel}}$ and $\mathsf{err}^{u}_{\mathsf{rel}}$ are the absolute errors and the relative errors in approximating the full-order optimal

state $\bar{y}$ and control $\bar{u}$, respectively. More precisely, with $\bar{y}^\ell$ and $\bar{u}^\ell$ being a numerically reduced-order model solution, we define

$$\mathsf{err}^y_{\mathsf{abs}} = \left(\int_0^T \|\bar{y}(t) - \bar{y}^\ell(t)\|_M^2 \, \mathrm{d}t \right)^{1/2}, \qquad \mathsf{err}^y_{\mathsf{rel}} = \frac{\mathsf{err}^y_{\mathsf{abs}}}{\left(\int_0^T \|\bar{y}(t)\|_M^2 \, \mathrm{d}t \right)^{1/2}},$$
$$\mathsf{err}^u_{\mathsf{abs}} = \left(\int_0^T \|\bar{u}(t) - \bar{u}^\ell(t)\|_M^2 \, \mathrm{d}t \right)^{1/2}, \qquad \mathsf{err}^u_{\mathsf{rel}} = \frac{\mathsf{err}^u_{\mathsf{abs}}}{\left(\int_0^T \|\bar{u}(t)\|_M^2 \, \mathrm{d}t \right)^{1/2}},$$

where again the time integrals are realized numerical by applying a standard trapezoidal approximation.
As previously mentioned, to compute a reduced basis, we need snapshots of the full model in advance. The better the snapshots represent the dynamics involved in the problem, the smaller the approximation error in the reduced-order model will be. However, since we are in an MPC setting, solving the full FE optimality system over the entire time-horizon up to T would be computationally too expensive. Therefore, we generate the snapshots by solving the full-order optimality system only up to a fixed time horizon. Since the MPC is performed shifting the time-horizon at each time step, we can expect that the MOR approximation of the initial snapshots may not be accurate for the entire MPC computation. Therefore, one has to decide under which conditions to perform an update of the reduced basis. Now, the following question arises naturally: Is it possible to measure the error between the reduced-order and the full-order solutions without computing the full order one?
In our case, the answer is provided by the a-posteriori error estimate of Corollary 4.5. Applied to the optimization problem (**P**)-(**C**) it has the following form:

$$\begin{aligned} \|\bar{\mathfrak{u}} - \bar{\mathfrak{u}}^\ell\|_{\mathfrak{U}}^2 \leq \frac{1}{\sigma^2} \int_0^T e^{\tilde{c}(T-t)} \Bigg(\int_t^T 2 \|\mathsf{res}^{\mathsf{du}}(s)\|_{V'}^2 \\ + e^{\sqrt{5}s} \left(\|y_\circ - \mathcal{P}^\ell y_\circ\|_H^2 + 2 \int_0^s \|\mathsf{res}^{\mathsf{pr}}(\tau)\|_{V'}^2 \, \mathrm{d}\tau \right) \mathrm{d}s \Bigg) \, \mathrm{d}t, \end{aligned} \tag{5.13}$$

with $\tilde{c} = 1 + \sqrt{5}$ and the residuals

$$\begin{aligned} \langle \mathsf{res}^{\mathsf{pr}}(t), \varphi \rangle_{V',V} &:= \langle \mathfrak{f}(t) + \mathcal{B}(t)\mathfrak{u}(t), \varphi \rangle_{V',V} - \frac{\mathrm{d}}{\mathrm{d}t} \langle \mathfrak{y}^\ell(t), \varphi \rangle_H - a(t; \mathfrak{y}^\ell(t), \varphi), \\ \langle \mathsf{res}^{\mathsf{du}}(t), \varphi \rangle_{V',V} &:= \langle \mathfrak{z}_d(t) - \mathfrak{y}^\ell(t), \varphi \rangle_H + \frac{\mathrm{d}}{\mathrm{d}t} \langle \mathfrak{p}^\ell(t), \varphi \rangle_H - a(t; \varphi, \mathfrak{p}^\ell(t)) \end{aligned}$$

for all $\varphi \in V$ and for almost all $t \in (0, T]$. If the a-posteriori error estimate (5.13) exceeds a predefined tolerance, we update the reduced model by solving the full model for a predefined time-horizon. This procedure is summarized in Algorithm 5. We define the a-posteriori error estimator as

$$\begin{aligned} \Delta_{\mathsf{abs}}(\bar{\mathfrak{u}}, \bar{\mathfrak{u}}^\ell) := \Bigg(\frac{1}{\sigma^2} \int_0^T e^{\tilde{c}(T-t)} \Bigg(\int_t^T 2 \|\mathsf{res}^{\mathsf{du}}(s)\|_{V'}^2 \\ + \, e^{\sqrt{5}s} \left(\|y_\circ - \mathcal{P}^\ell y_\circ\|_H^2 + 2 \int_0^s \|\mathsf{res}^{\mathsf{pr}}(\tau)\|_{V'}^2 \, \mathrm{d}\tau \right) \mathrm{d}s \Bigg) \, \mathrm{d}t \Bigg)^{\frac{1}{2}}. \end{aligned} \tag{5.14}$$

For the numerical realization and, therefore, the semi-discrete FE solutions $\bar{u}$ and $\bar{u}^\ell$ the a-posteriori estimator $\Delta_{\mathsf{abs}}(\bar{u}, \bar{u}^\ell)$ is defined analogously to (5.14) by replacing the function space norms with the respective finite dimensional FE matrix norms, i.e.

$$\|\bar{\mathfrak{u}}^{\mathsf{fe}} - \bar{\mathfrak{u}}^{\mathsf{fe},\ell}\|_{\mathfrak{U}}^2 = \int_0^T \|\bar{u}(t) - \bar{u}^\ell(t)\|_M^2 \, \mathrm{d}t.$$

For a efficient evaluation of the dual norms $\|\mathsf{res}^{\mathsf{du}}(s)\|_{(V^{\mathsf{fe}})'}$ and $\|\mathsf{res}^{\mathsf{pr}}(s)\|_{(V^{\mathsf{fe}})'}$ of the residuals we refer to [15, Section 4.2.5]. Remark, that for the a-posteriori error estimate we do not have to compute the full-order solutions.

Algorithm 5: Reduced MPC with a-posteriori error estimator

Require: MPC prediction horizon $T_{\mathsf{pred}} > 0$, MPC feedback horizon $T_{\mathsf{f}} > 0$, update horizon $T_{\mathsf{train}} > 0$, update tolerance τ_{upd}, initial data $x_\circ \in \mathbb{R}^{2m}$.

1: Compute snapshots $X = [x_1^1|...|x_n^1|x_1^2|...|x_n^2|...|x_1^\wp|...|x_n^\wp] \in \mathbb{R}^{2m\times(N\wp)}$ by solving (5.11) p-times over the finite horizon $[0, T_{\mathsf{train}}]$;
2: Use snapshots X to compute a reduced discrete optimality system $(\mathbf{P}_{\mathsf{red}})$;
3: **for** $n = 0, 1, 2, ...$ **do**
4: Set sampling time $t_n = t_0 + nT_{\mathsf{f}}$ and measure the reduced couples state $\tilde{x}_\circ^\ell = x^\ell(t_n) \in H^1(0, T; \mathbb{R}^{n_y})$;
5: Solve the reduced optimality system $(\mathbf{P}_{\mathsf{red}})$ over the finite horizon $[t_n, t_n + T_{\mathsf{pred}}]$ and obtain $\bar{x}^\ell$;
6: Extract reduced optimal control $\bar{u}^\ell$;
7: **if** $\Delta_{\mathsf{abs}}(\bar{u}, \bar{u}^\ell) > \tau_{\mathsf{upd}}$ **then**
8: Update the reduced optimality system $(\mathbf{P}_{\mathsf{red}})$ using snapshots of (5.11) over the finite horizon $[t_n, t_n + T_{\mathsf{train}}]$;
9: Store the optimal control $\bar{u}$ over the time interval $[t_n, t_n + T_f)$ and use this control value in the next sampling period;
10: **else**
11: Store the reduced optimal control $\bar{u}^\ell$ over the time interval $[t_n, t_n + T_f)$ and use this control value in the next sampling period;
12: **end if**
13: **end for**

Empirical Gramians and POD

Next, we want to discuss how we compute the reduced models for Algorithm 5, see step 2. We will begin by highlighting the main advantages of POD and empirical gramains.

- For the POD model, one would typically run simulations of the state and adjoint equations by varying the control u to construct a POD basis. The more inputs u used for the training that are closer to the optimal control, the better the POD approximation will be. However, this is only possible if the entire dynamics are known a-priori. Since we are in a MPC setting, this is an unlikely assumption, because in a realistic scenario one does not know the exact dynamics while looking far into the future. Furthermore, it is often not possible to predict a-priori what the optimal control will be.

- The empirical gramian approach, on the other hand, is more flexible. There is the possibility of training the controllability gramian (and thus the reduced-order model) by choosing an arbitrary parameter function as input. This could be, for example, the target z_d, which is executed in a multi-objective optimization setting in [29], or any other parameter function. Building a POD model capable of approximating the solution of problems with different parameter functions would require running many simulations varying parameter functions and controls. This is also only possible if the different parameter function are known a-priori. The gramian approach is able to exploit randomness to construct a good approximation of the full-order model, while the POD seriously struggles if the generated snapshots are not close to the optimal solution of the problem. In this way, we can compute basis functions that will be sensitive to perturbations of a parameter function. Furthermore, we have the degree of freedom to choose a output for the observability gramian. One can for example choose the first order optimality condition $z(t) = M^{-1}B(t)^\top p(t)/\sigma$ of the perturbed optimal control problem, which can lead to a reduced reduced-order model that accurately reconstructs the optimal control.

- If we examine the computation time of the trainings phase, we can see that if we have a large number of snapshots, we have to perform a costly SVD on the big snapshot matrix for the POD approach. For the gramian approach, we have the advantage that only a SVD on a matrix that has the same size as the problem needs to be performed.

Training of the reduced-order model

To compute the empirical gramians, we apply Algorithm 2 to the discrete optimality system (5.11) over the finite horizon $[0, T_{\mathsf{train}}]$, with $T_{\mathsf{train}} = 1$. For the input function, we choose the advection coefficient, i.e. $u_{\mathsf{inp}}(t) = \alpha(t)$, and as output we choose $z_{\mathsf{out}}(t) = M^{-1}B(t)^\top p(t)/\sigma$. The gramian coefficient set $\mathcal{E}$ is chosen randomly with $|\mathcal{M}| = 3$ and $1 \in \mathcal{M}$. For the gramian orthogonal matrices sets, we choose $\mathcal{T}^{2m}$ and $\mathcal{T}$ randomly with $|\mathcal{T}^{2m}| = |\mathcal{T}| = 2$, $I_{2m} \in \mathcal{T}^{2m}$ and $1 \in \mathcal{T}$. Further, we restrict $\mathcal{E}^{2m}$ to ten random vectors. In conclusion, we compute $3 \times 2 = 6$ trajectories for the controllability gramian and $3 \times 2 \times 10 = 50$ trajectories for the observability gramian.
For a fair comparison, the snapshots used to generate the POD basis are the trajectories

while computing the controllability and observability gramians in Algorithm 2. The POD basis is then generate with Algorithm 1 with $\tilde{W} = I_{2m}$ and D are trapezoidal weights. It is worth noting that if we increase the number of computed snapshots for the training

	Snapshots	Empirical Gramian	POD
Computation time	59.1s	0.1s (+59.1s)	0.4s (+59.1s)

Table 5.3: CPU time of the offline phase.

time T_{train} the offline CPU time of the POD method increase significantly, while the gramian approach remains fast. To reduce the CPU time of computing the snapshots, one can think of computing the different trajectories in parallel. This is possible since all trajectories are independent of each other.

In the first test, we assume that we know the advection coefficient α. Therefore, we train the controllability gramian while perturbing the original advection coefficient α. Also, for the observability gramian, we use the original advection coefficient. The result is shown in Figure 5.8. Note that the full-order model of the optimality system has dimension 724.

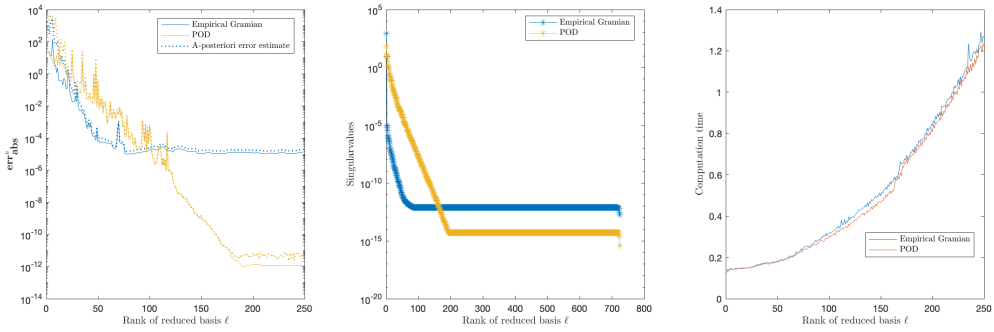


Figure 5.8: Empirical gramian and POD MOR results for (5.11) over the finite horizon $[0, T_{\mathsf{train}}]$ with known advection. Reported are the absolute control error $\mathsf{err}^u_{\mathsf{abs}}$ with the a-posteriori error Δ_{abs} for different number ℓ of basis functions (left), the singular values of the empirical gramian and the POD snapshot matrix (middle) and the different CPU times for different number ℓ of basis functions (right).

As shown in Figure 5.8, the eigenvalues corresponding to the gramian basis functions decrease more rapidly at the beginning compared to those corresponding to the POD approach. This leads to an improved low-rank approximation of the gramian reduced model, as more information about the dynamics is captured by the first gramian basis functions. However, as the number of basis functions is increased, the roles are reversed and the POD approximation becomes more accurate. This is because after approximately the first 60 singular values of the gramian, the decay of the singular values plateaus and further improvement of the reduced model is not possible. However, it should be noted that the plateau is at a level that is lower than the approximation quality for the full model obtained by the implicit Euler method and the FE method. On the other hand, the singular values of the POD snapshot matrix continue to decrease and the POD

approximation becomes very close to the full model. Increasing the rank of the POD basis also leads to an increase in computational effort. Since it takes around 1.2 seconds to compute the full-order model for the horizon $[0, T_{\mathsf{train}}]$ it is not practical to choose a reduced-order basis with a rank ℓ greater than 200. The choice of how large to make the rank ℓ can be also determined using the a-posteriori error estimator, given a tolerance τ_ℓ. In a MPC setting, one can also consider a combination of the two reduced models. The gramian basis can provide a fast low-rank approximation, while the POD basis can be used for more accurate approximation, at the cost of increased numerical costs.

In the second test, we assume that the advection coefficient α is unknown. As a result, we cannot compute snapshots by pertubing the correct value of α. Instead, we use a randomly chosen advection coefficient while computing the gramians. The trajectories required for the gramians are also used as the snapshot matrix for the POD. As shown in Figure

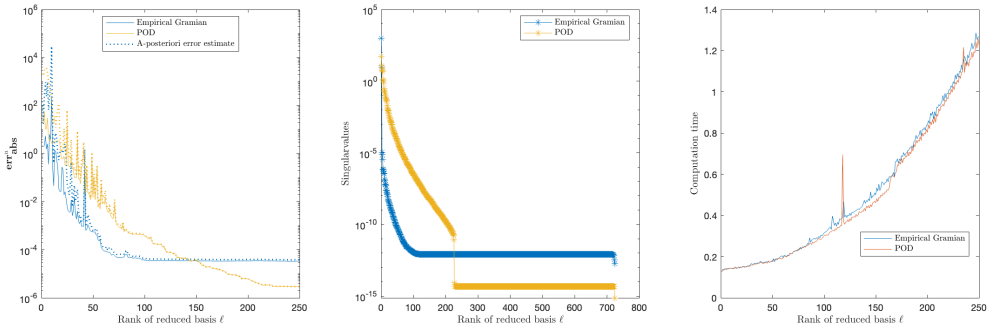


Figure 5.9: Empirical gramian and POD MOR results for (5.11) over the finite horizon $[0, T_{\mathsf{train}}]$ with unknown advection. Reported are the absolute control error $\mathsf{err}^u_{\mathsf{abs}}$ with the a-posteriori error estimate Δ_{abs} for a different number ℓ of basis functions (left), the singular values of the empirical gramian and the POD snapshot matrix (middle) and the different CPU times for different number ℓ of basis functions (right).

5.9, the results are similar to those obtained before, but the POD has more difficulty in reconstructing the solution. As expected, using a random advection coefficient leads to snapshots that are not as close to full solution. It is way harder for the POD-based reduced-order model to reconstruct the optimal solution with these snapshots, whereas the gramian approach produces similar results as before. The gramian-based MOR is capable of capturing the characteristics of the dynamical system. Therefore, individual snapshots do not fall within the weight, instead the entire dynamics is reproduced.

In the last test, we again assume that the advection coefficient α is unknown. In addition, we increase the value of α so that the problem defined in (**P**)-(**C**) becomes more advection dominant. To achieve this, we set $\alpha(t) = 10\sin(t)$. As shown in Figure 5.10, the POD model performs poorly in the presence of advection dominant problems, which is a well-known behavior. In [28], it is also noted that for advection dominant problems, a small number of snapshot that deviate from the original snapshots can lead to inaccuracies. The gramian approach is also less accurate than before, but the deviation of the snapshots does not greatly affect the overall model.
In conclusion the gramian method is more effective in replicating the key characteristics

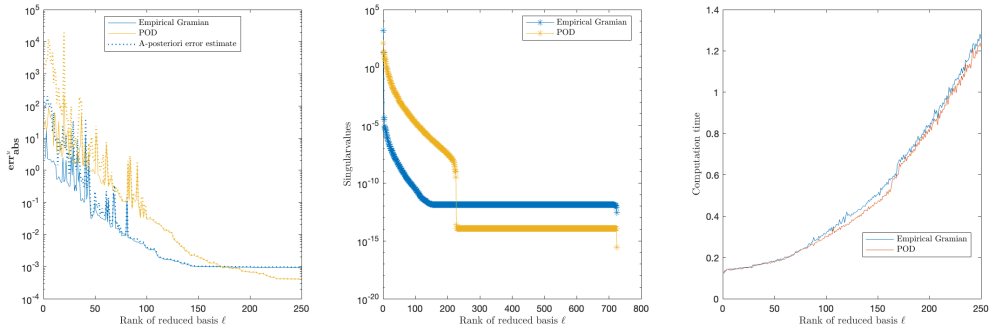


Figure 5.10: Empirical gramian and POD MOR results for (5.11) over the finite horizon $[0, T_{\mathsf{train}}]$ with unknown advection. Reported are the absolute control error $\mathsf{err}^u_{\mathsf{abs}}$ with the a-posteriori error estimate Δ_{abs} for a different number ℓ of basis functions (left), the singular values of the empirical gramian and the POD snapshot matrix (middle) and the different CPU times for different number ℓ of basis functions (right).

of the dynamics, resulting in several advantages, including the ability to handle unknown coefficient functions and advection. Additionally, the gramian-based MOR showed a better low-rank approximation. This superiorities of the gramian method is due to its ability to incorporate control and observability knowledge through the use of empirical gramians, which were specifically trained for the advection coefficient in this instance. However, if the correct full-order model snapshots are available, POD still offers the best reduced-order model for a specific solution. If one is focused on a specific solution, then the POD method is recommended. Anyhow, if the aim is to study the dynamics with varying coefficients, the gramian approach is more beneficial.

Finally, we are interested in the efficiency of the a-posteriori error estimator. For this application, it would be ineffective if the error estimator overestimates the actual error between the full and the reduced model. Therefore, we define the following efficiency constant

$$\mathsf{eff}^{\Delta} := \frac{\mathsf{err}^{\mathsf{u}}_{\mathsf{abs}}}{\Delta_{\mathsf{abs}}(\bar{u}, \bar{u}^{\ell})} \in [0, 1] \tag{5.15}$$

and the error gap

$$\mathsf{gap}^{\Delta} := \Delta_{\mathsf{abs}}(\bar{u}, \bar{u}^{\ell}) - \mathsf{err}^{\mathsf{u}}_{\mathsf{abs}} \geq 0. \tag{5.16}$$

The closer eff^{Δ} is to the one and the smaller is the gap gap^{Δ}, the better is the error estimator. In Figure 5.11, we present the results for all three examples above using gramian basis functions. It can be observed that the better the model-order reduction, the better the a-posteriori error estimator Δ_{abs} performs. When the rank of the reduced basis is sufficiently large, the error estimator Δ_{abs} approximates the actual error very closely. Similar results are obtained using POD basis functions.

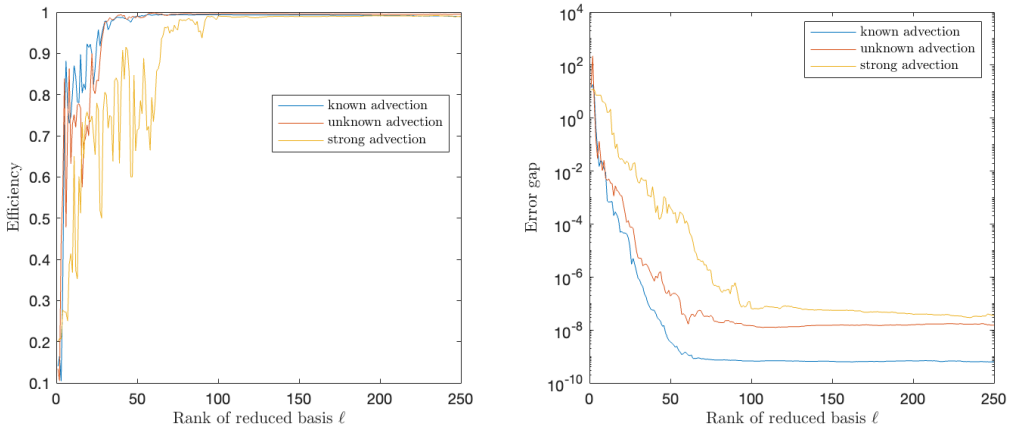


Figure 5.11: Efficiency eff^{Δ} (left) and error gap gap^{Δ} for the error estimator Δ_{abs} for the three examples investigated above with varying advection. The results are based on the empirical gramian MOR.

MPC for the reduced-order models

Now, we present the results for the reduced MPC results. To have a fixed setting, we choose the MPC parameter $T_{\mathsf{pred}} = 2$ and $T_{\mathsf{f}} = 1$, and assume that we know the advection in the trainings phase to build the reduced-order models. In a first test, we compute a reduced POD and gramian basis. Therefore, we compute the empirical gramian for $T_{\mathsf{train}}^{\mathsf{Gramian}} = 2$. For the POD basis, we do not use the snapshots required for the gramian, instead we solve the constraint dynamics once up to the time point $T_{\mathsf{train}}^{\mathsf{POD}} = 8$. Firstly, we investigate the performance of one single basis for the entire MPC. We choose a reduced basis rank of $\ell = 50$, as it has resulted in an accurate reduced model, and the a-posteriori error estimator Δ_{abs} appears to be stable. The results can be found in Figure 5.12 with additional errors and computation times in Table 5.4. In Figure 5.12 we can see that in the

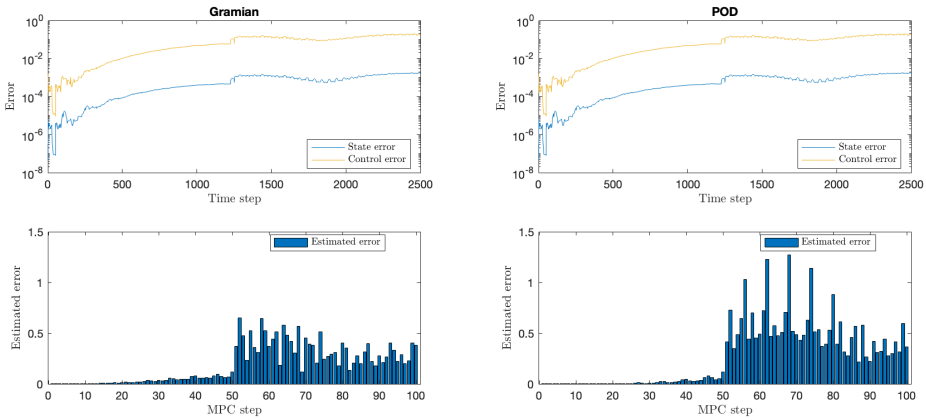


Figure 5.12: Reduced MPC with one reduced gramian (left) and POD (right) basis with reduced basis rank $\ell = 50$. One every column are the absolute L^2-spatial-difference in state and control over the time, i.e. $t \mapsto \|\bar{y}(t) - \bar{y}^{\ell}(t)\|_M$ and $t \mapsto \|\bar{u}(t) - \bar{u}^{\ell}(t)\|_M$ (top) and the particular error estimates for every MPC step (bottom).

beginning, we get a good approximation of the reduced model and the estimated errors are small. This is due to the fact that the reduced basis is trained baseed on these snapshots. However, at least from the time when the operator $\mathcal{B}$ changes, the error estimates increase abruptly and the reduced basis becomes unusable. As a result, in the second test, we make use of the update strategy mention in Algorithm 5. We are not investigating the use a gramian-based MPC with error estimator, as updating the gramains is computational too expensive. If we only use one solution trajectory as snapshots to compute the gramian, the approach is coinciding to the POD approach. Therefore, we restrict ourselves to the POD setting for the update strategy. The basis update is executed as follows:

- Solve the full FE model from the current time up to $T_{\mathsf{train}} = 8$.
- Compute the reduced POD basis with respect to the snapshots of the full FE model.

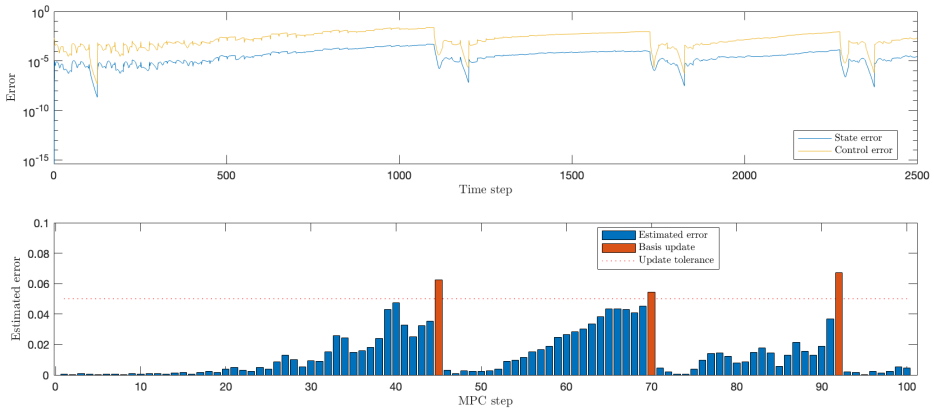


Figure 5.13: Reduced MPC-POD with a-posteriori estimator with $\tau_{\mathsf{upd}} = 5 \times 10^{-2}$. For the respective POD basis we choose rank $\ell = 50$. Reported are the absolute L^2-spatial-difference in state and control over the time (top) and the particular error estimates for every MPC step (bottom).

In Figure 5.13, we can see that after every basis update, the error estimates are reduced, but as expected, they increase again over time. One has to be careful that the update tolerance τ_{upd} is not set too low, as every basis update is computationally costly. If too many basis updates are performed, no computational time will be saved.

	$\mathsf{err}^y_{\mathsf{rel}}$	$\mathsf{err}^u_{\mathsf{rel}}$	CPU time	Speedup
MPC	-	-	380s	-
MPC-POD	7.8×10^{-2}	3.3×10^{-1}	59s (+21s)	6.4 (4.8)
MPC-Gramian	2.7×10^{-2}	1.4×10^{-1}	59s (+59s)	6.4 (3.2)
MPC-POD with error estimator	3.2×10^{-3}	8.9×10^{-3}	79s (+21s)	4.8 (3.8)

Table 5.4: Reduced MPC results with a reduced basis rank $\ell = 50$.

DMD and EDMD

In this section, we conduct several numerical tests to evaluate the performance of linear MPC based on DMD and EDMD. An interesting aspect of DMD is that it transforms the pseudo-forward optimality system into a pure forward dynamical system. The results show the potential and limitation of the methods. Therefore, we consider the standard DMD or EDMD without control, as the external control u is contained in the state x of the dynamical system.
When we consider the small time interval $[0, 4]$, we can see the potential of the DMD methods, as shown in Figure 5.14 with corresponding errors and significant offline speedups in Table 5.5. For EDMD, we use the following lifting function

$$\psi : \mathbb{R}^{2m} \to \mathbb{R}^{2m}, \quad x \mapsto \begin{bmatrix} x_{(1:m)} \\ \frac{1}{\sigma} M^{-1} B(t)^\top x_{(m+1:2m)} \end{bmatrix},$$

where $x_{(1:m)}$ denotes the first m components of x and time t is chosen to correspond to the time point of the respective snapshot x. The purpose of this lifting function is to transforms the adjoint snapshots into the control snapshots to reduce the error in the control. The state error of the DMD approach is already sufficiently small, so it remain unchanged. As we can see, the lifting function fulfills its task by slightly increasing the

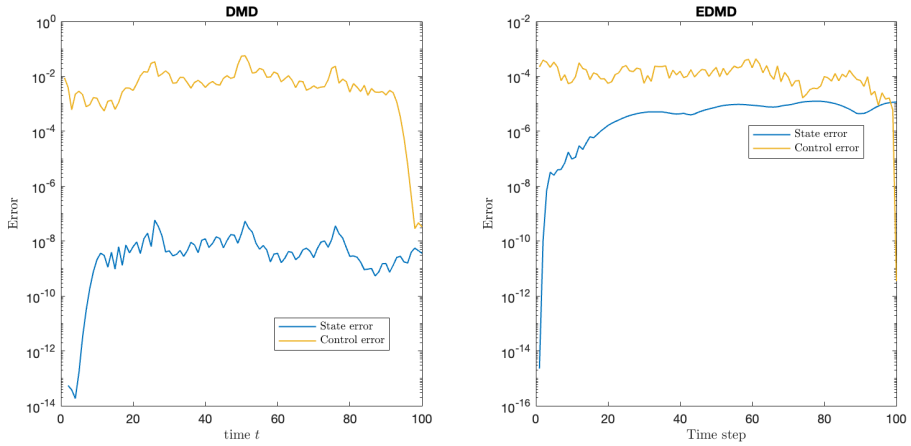


Figure 5.14: Absolute errors over time, i.e. $t \mapsto \|\bar{y}(t) - \bar{y}^{DMD}(t)\|_M$ and $t \mapsto \|\bar{u}(t) - \bar{u}^{DMD}(t)\|_M$, of the (E)DMD method for the dynamical system (5.9) on the time interval $[0, 4]$. The truncation value of the (E)DMD operator is given by $r = 50$.

	$\mathsf{err}^y_{\mathsf{rel}}$	$\mathsf{err}^u_{\mathsf{rel}}$	CPU time	Speedup
FE	-	-	8s	-
DMD	1.1×10^{-7}	1.2×10^{-2}	0.01s	800
EDMD	6.4×10^{-5}	1.6×10^{-3}	0.13s	61.5

Table 5.5: Relative errors and offline CPU times for the (E)DMD method for the dynamical system (5.9) on the time interval $[0, 4]$ with truncation value $r = 50$.

error in the state and naturally increasing the numerical cost due to the computation of the lifted snapshot matrices. In conclusion, the DMD methods work very well on the small time interval.
Next, we want to apply the DMD methods in the MPC setting. In [37], it is pointed out that the DMD framework is designed for time-invariant dynamics. Since we are dealing with a time-varying dynamical system (5.9) we can only expect a satisfactory approximation locally in time. The numerical results confirm this assertion. If we use the same (E)DMD operators as in the previous test, i.e., set $T_{\mathsf{train}} = 4$, the (E)DMD-MPC is not able to reconstruct the full system over the whole time horizon, as shown in Table 5.6. Consequently, we are interested in developing a DMD operator update procedure. We proceed analogously to Algorithm 5. But since at time t we only have access to the control $u(T-t)$, we do not use the a-posteriori estimator Δ_{abs} for the control u given in (5.14). Instead, we use the primal a-posteriori estimator for the state given in Theorem 4.3. As

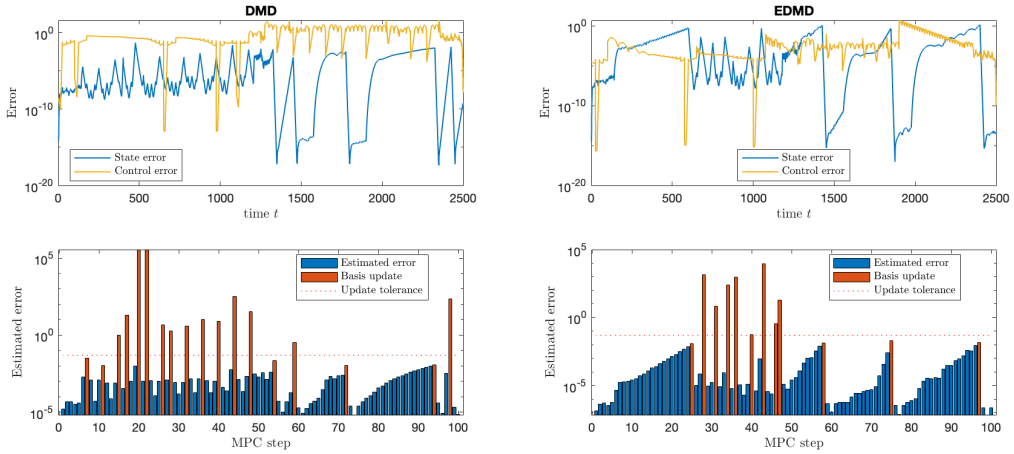


Figure 5.15: Reduced MPC based on DMD/EDMD with a-posteriori estimator with $\tau_{\mathsf{upd}} = 10^{-2}$ and truncation value $r = 50$. Reported are the absolute L^2-spatial-difference in state and control over the time (top) and the particular error estimates for every MPC step (bottom).

one can see in Figure 5.15, the DMD and EDMD operators require many more updates than the POD or gramian approach. This is reflected in the CPU times and speedups. Additionally, the approximation of the DMD model over a large horizon is not as good as in the previous approaches. Therefore, one can conclude that the (E)DMD-MPC can be a valid approach, although the quality of the approximation and the update strategy during the MPC iterations need to be improved. Especially for a short term intervals, (E)DMD can be very useful.

	Basis updates	$\mathsf{err}^y_{\mathsf{rel}}$	$\mathsf{err}^u_{\mathsf{rel}}$	CPU time	Speedup
MPC	-	-	-	380s	-
DMD	0	3.1×10^{26}	1.4×10^{27}	1s	380.0
DMD	18	8.2×10^{-2}	3.8×10^{-1}	260s	1.4
EDMD	0	1.1×10^{14}	1.0×10^{14}	3s	126.7
EDMD	12	5.0×10^{-1}	3.6×10^{-1}	173s	2.2

Table 5.6: Relative errors and offline CPU times for the DMD/EDMD based MPC method for the dynamical system (5.9) on the time interval $[0, 100]$ with truncation value $r = 50$.

6 | Conclusion

In this thesis, we presented theoretical results for a quadratic optimal control problem for linear partial differential equations with time-dependent coefficient functions. The existence and uniqueness of the general linear-quadratic problem were established, along with the derivation of the necessary and sufficient first order optimality conditions. After applying a FE discretization in space the associated first order optimality system can be interpreted as a coupled LTV dynamical system. These findings were applied to design efficient algorithms for solving a numerical test example using model predictive control. The solution to these problems is computationally expensive because in the MPC framework, many optimal control problems must be solved iteratively. To speedup the process, data-driven model-order reduction techniques such as POD, empirical gramians and (E)DMD have been applied to obtain reduced-order models directly for the first order optimality system composed of the coupled state and adjoint equation. Additionally, an a-posteriori estimate for the approximation error of the reduced order model was derived.
A comparison between empirical gramian and POD over a short training time period showed that the gramian method was more effective while replicating the key characteristics of the dynamics. This led to several benefits, including the ability to handle unknown coefficients of the dynamics and the possibility to handle advection. Furthermore, the gramian-based MOR showed a better low-rank approximation. This advantages of the gramian method is due to its ability to incorporate controlability and observability knowledge through empirical gramians. If the correct full-order model snapshots are available, POD has the potential to produce the best reduced-order model for the one specific solution. If the goal is to obtain a reduced-order model for a single solution, the POD method may be the best choice. On the other hand, if the interest lies in capturing the dynamics with varying coefficients, the gramian method is more advantageous.
To efficiently solve the optimal control problem over a long time horizon, only initial snapshots of the full-order model were used to create reduced models. However, it was found that the reduced models based on these snapshots were not accurate for the entire computation of the MPC. Instead they especially brought good results in the beginning. As a result, the a-posteriori error estimates were used to guarantee the accuracy of the POD approximation and to create a procedure for updating the POD basis. Furthermore, the efficiency of the error estimator was demonstrated through a numerical test.
Moreover, the same test were done for a MPC based on DMD, EDMD respectively. Both methods showed the best results locally in time but required many basis updates over a long time horizon, which were computationally expensive. As a result, the speedup suffered from the many basis updates and were not that good, than for the POD method. DMD and EDMD performed similarly, with the main difference being that EDMD pro-

vided more degrees of freedom to influence the output of the underlying dynamic, which is especially useful when approximating a specific output, such as the control in our example.

Regarding future work, we make general remarks on what can be done next. In the Outlook in Chapter 7 we discuss how to solve the resulting optimality systems when additional control constraints come into play (cf. Section 3.2.3). Solving the optimality system through MPC requires even more computational resources than in the linear case because we make use of the semismooth Newton method. Thus, there is interest in a reduced MPC with an update strategy. To achieve this, it is necessary to derive a-priori estimates, which is not a straightforward task, since we have to differ between active and inactive control points.
Another area of future interest is to extend the proposed approach to nonlinear dynamical systems.

7 | Outlook: MPC with additional Control Constraints

In this chapter, we provide an outlook of how to solve optimal control problems with additional control constraints, which are introduced in Section 3.2.3. As previously discussed, solving an optimal control problem with control constraints results in a nonsmooth optimality system. We begin demonstrating how one can solve the optimality system by using the semismooth Newton method. First, we provide a brief overview of the semismooth Newton method. The semismooth Newton method is costly in the MPC setting, but however, the model-order reduction with error estimates for these kinds of problems requires more specialized knowledge and will not be further explored in this thesis.

7.1 Semismooth Newton Method

In many practical applications one is interested to find a zero of a system of nonlinear equations, i.e., we are interested in finding an $\bar{x} \in X$ such that

$$F(\bar{x}) = 0, \quad \text{where } F : X \to Z$$

is a nonlinear function, where X and Z are Banach spaces. For example in Chapter 3 we have derived several optimality systems, which have to be solved. A natural approach to solve this problem is to use a Newton method. For the Newton method one has to require that F is continuously differentiable, but what happens if F is not differentiable everywhere. For instance, it turns out that the resulting optimality system in Section 3.2.3 is nonlinear and even nonsmooth. Thus, dealing with nonsmooth functions is quite natural.
In this chapter we will give an explanation how to deal with a nonsmooth system by introducing the semismooth Newton method. First of all, we will introduce the semismooth Newton method in the finite setting, i.e., we consider a function $F : \mathbb{R}^n \to \mathbb{R}^m$. Afterwards, we also extend the method for nonsmooth Banach space-valued functions. For this chapter we are guided by [44].

7.1.1 Generalized Differentials and semismooth Newton Methods in Finite Dimensions

We start this section with the construction of a generalized derivative of a non-differentiable function. The construction goes back to [7] and is based on Rademacher's theorem.

Therefore we consider a locally Lipschitz continuous map $F : \mathbb{R}^n \to \mathbb{R}^m$. Let $D_F \subset \mathbb{R}^n$ be the set of all points, where F is differentiable.

Theorem 7.1 (Rademacher)
Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be locally Lipschitz continuous, then F is almost everywhere differentiable, i.e. $\lambda(\mathbb{R}^n \backslash D_F) = 0$, where λ denotes the Lebesgue measure.

Proof. A proof is given in [7]. □

Now we introduce several objects from nonsmooth analysis which facilitates us to construct a generalized Newton method.

Definition 7.2 (Clarke's generalized Jacobian)
Let $F : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$, with U open, be locally Lipschitz continuous at $x \in U$. Clarke's generalized Jacobian *is defined as*

$$\partial F(x) = conv(\partial_B F(x)),$$

where conv denotes the convex hull and B-subdifferential is defined as

$$\partial_B F(x) := \{G \in \mathbb{R}^{m \times n} : \exists \{x_k\} \subset D_F \text{ with } x_k \to x, \nabla F(x_k) \to G\}.$$

Next we study some properties of the respective derivative.

Proposition 7.3 (Properties of the generalized derivative)
Let $U \subset \mathbb{R}^n$ be open and $F : U \to \mathbb{R}^m$ be locally Lipschitz continuous. Then for $x \in U$ it holds:

(a) $\partial_B F(x)$ is nonempty and compact.

(b) $\partial F(x)$ is nonempty, compact and convex.

(c) The set-valued mappings ∂F and $\partial_B F$ are locally bounded and upper semicontinuous, i.e. for every $\varepsilon > 0$ there exists a $\delta > 0$ such that, for all $y \in B(x, \delta)$

$$\partial F(y) \subseteq \partial F(x) + B(0, \varepsilon).$$

(d) The following inclusion hold true:

$$\partial_B F(x) \subset \partial F(x).$$

(d) If F is continuously differentiable in a neighborhood of x, then

$$\partial F(x) = \partial_B F(x) = \{F'(x)\}.$$

Proof. A proof is given in [44, Proposition 2.2]. □

For illustration we make the following example.

Example 7.4 (Clarke's Jacobian for a non-differentiable function)
Let $F : \mathbb{R} \to \mathbb{R}, x \mapsto \max(x^2, x)$. *Then we have* $F \in C^1((-\infty, 0) \cup (0, 1) \cup (1, \infty))$ *and* F *is non-differentiable in* $\{0, 1\}$. *In the zero point we get*

$$\lim_{x \nearrow 0} \nabla F(x) = \lim_{x \nearrow 0} 2x = 0 \in \partial_B F(0).$$

Moreover, we obtain

$$\lim_{x \searrow 0} \nabla F(x) = \lim_{x \searrow 0} 1 = 1 \in \partial_B F(0).$$

Analogously we can make the same argument in the point 1. *Consequently, we get*

$$\partial F(x) = \begin{cases} conv\{0, 1\} = [0, 1] & \text{if } x = 0, \\ \{1\} & \text{if } x \in (0, 1), \\ conv\{1, 2\} = [1, 2] & \text{if } x = 1, \\ \{2x\} & \text{otherwise.} \end{cases}$$

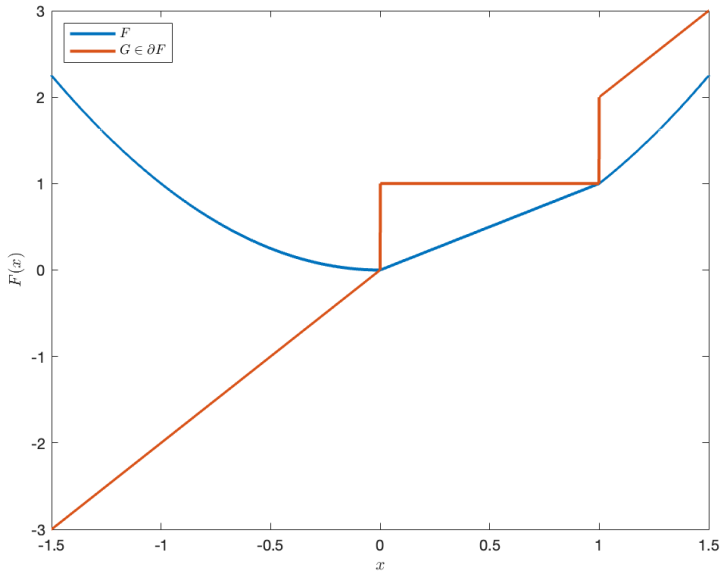


Figure 7.1: Example of a locally Lipschitz continuous, but non-differentiable function F with a generalized derivative $G(x) \in \partial F(x)$.

A first, really intuitive construction of a general Newton method for a locally Lipschitz continuous map $F : \mathbb{R}^n \to \mathbb{R}^m$ could be the following:

Algorithm 6: Newton method for nonsmooth systems

Require: Initial guess $x_0 \in \mathbb{R}^n$.

1: Set $k = 0$;
2: **while** stopping criteria is not satisfied **do**
3: Solve

$$G(x_k)d_k = -F(x_k), \tag{7.1}$$

where $G(x_k) \in \partial F(x_k)$ is arbitrary;
4: Set $x_{k+1} = x_k + d_k$, $k = k + 1$;
5: **end while**

To guarantee the well-posedness of Algorithm 6 we have to ensure that we can solve (7.2), and therefore $G(x_k)$ has to be nonsingular. This is a more involved task than for the standard Newton scheme. Firstly, the generalized Jacobian ∂F can be set-valued, like we saw in Example 7.4. Furthermore, we can not argue that $\|G(y) - G(x)\| \to 0$ as $x \to y$. The idea of the smooth Newton method is to use a local linearization $m(x_k + d)$ of F about x_k. If F is sufficiently smooth this gives us a good approximation of F. If F is only locally Lipschitz this has not to be the case.
In conclusion, without further assumptions on F we can not guarantee the existence of a sufficient small $\varepsilon > 0$, such that $x_{k+1} \in B(\bar{x}, \varepsilon)$ when $x_k \in B(\bar{x}, \varepsilon)$, where $\bar{x}$ satisfies $F(\bar{x}) = 0$.
Therefore, we introduce the class of semismooth functions such that we finally obtain well-definedness and locally superlinear convergence of Algorithm 1.

Definition and Theorem 7.5 (Semismoothness)
Let $U \subset \mathbb{R}^n$ be nonempty and open. The function $F : U \to \mathbb{R}^m$ is called semismooth *at $x \in U$ if F is locally Lipschitz and one of the following equivalent statements hold true:*

(i)

$$\lim_{\substack{G \in \partial F(x + t\tilde{d}) \\ \tilde{d} \to d,\ t \downarrow 0}} G\tilde{d}$$

exists for all $d \in \mathbb{R}^n$.

(ii) F is directionally differentiable at $x \in U$ and

$$\sup_{G \in \partial F(x+d)} \|Gd - F'(x, d)\| = o(\|d\|) \quad \text{as } d \to 0.$$

(iii) F is directionally differentiable at $x \in U$ and

$$\sup_{G \in \partial F(x+d)} \|F(x + d) - F(x) - Gd\| = o(\|d\|) \quad \text{as } d \to 0.$$

If F is semismooth for all $x \in U$, we call F semismooth *(on U).*

Proof. A proof is given in [44, Proposition 2.7]. □

Example 7.6
Every continuous, piecewise C^1-function is semismooth.

Whenever F is semismooth we call Algorithm 1 *semismooth Newton method*. In comparison to the smooth Newton method we can not expect quadratic convergence for the semismooth Newton method, but we are already in the position to proof a local superlinear convergence rate of this method.

Theorem 7.7 (Superlinear local convergence)
Suppose that $\bar{x} \in \mathbb{R}^n$ satisfies $F(\bar{x}) = 0$, and F is locally Lipschitz and semismooth at $\bar{x}$, and $\partial F(\bar{x})$ is nonsingular, i.e., all $G \in \partial F(\bar{x})$ are nonsingular. Then there exists $\varepsilon > 0$ such that for $x_0 \in B(\bar{x}, \varepsilon)$ the sequence $\{x_k\}$ generated by the semismooth Newton method (Algorithm 1) is well-defined, converges to $\bar{x}$ and satisfies

$$\|x_{k+1} - \bar{x}\| = o(\|x_k - \bar{x}\|), \quad \text{as } k \to \infty.$$

Proof. First of all, one can show that there exists an $r > 0$ such that all elements of $\partial F(x)$ are nonsingular for all $x \in B(\bar{x}, r)$. Therefore, let $\beta > 0$ such that for all $G(x) \in \partial F(x)$ it holds $\sup_{x \in B(\bar{x}, r)} \|G(x)^{-1}\| \leq \beta$. Now choose $\varepsilon \leq r$. Similar to the smooth Newton method we can make now an inductive argument. For an arbitrary $G(x_0) \in \partial F(x_0)$ it holds

$$\begin{aligned} x_1 - \bar{x} &= x_0 - \bar{x} - G(x_0)^{-1} F(x_0) \\ &= G(x_0)^{-1}(F(\bar{x}) - F(x_0) - G(x_0)(\bar{x} - x_0)). \end{aligned}$$

Now fix $c \in (0, 1)$, and since F is semismooth at $\bar{x}$, for arbitrary $0 < \eta \leq \frac{c}{\beta}$ there exists $\tilde{r} > 0$ such that for $x_0 \in B(\bar{x}, \tilde{r})$ it holds

$$\|F(\bar{x}) - F(x_0) - G(x_0)(\bar{x} - x_0)\| \leq \eta \|x_0 - \bar{x}\|.$$

Consequently we obtain for $\varepsilon \leq \min\{r, \tilde{r}\}$

$$\|x_1 - \bar{x}\| \leq \beta\eta \|x_0 - \bar{x}\| \leq c\|x_0 - \bar{x}\| < \|x_0 - \bar{x}\|.$$

This implies $x_1 \in B(\bar{x}, \varepsilon)$ if $x_0 \in B(\bar{x}, \varepsilon)$. The induction step is analog. Since $c \in (0, 1)$ we find $x_k \to \bar{x}$ as $k \to \infty$. Furthermore, we get by the semismoothness of F

$$\|x_{k+1} - \bar{x}\| \leq \beta \|F(\bar{x}) - F(x_k) - G(x_k)(\bar{x} - x_k)\| = o(\|x_k - \bar{x}\|)$$

as $k \to \infty$. □

7.1.2 Generalized Newton-type Methods and Semismoothness in Infinite Dimensions

As a motivation, in many practical applications the function F is not given in finite dimensional spaces as considered in the previous section. Instead it turns out from the application that a function $F : U \subset X \to Z$ is given, where X and Z are Banach spaces. To extend the semismooth Newton method also for Banach space-valued functions we introduce in this section a generalized differentiable for these kind of functions. This is straight forward and analog to the finite dimensional case.

Definition 7.8
Let X and Z be Banach spaces, $U \subset X$ open. The mapping $F : U \to Z$ is generalized (or Newton) differentiable on U if there exists a family of mappings $G : U \to \mathcal{L}(X, Z)$ such that

$$\lim_{\varphi \to 0} \frac{1}{\|\varphi\|_X} \|F(u+\varphi) - F(u) - G(u+\varphi)\varphi\|_Z = 0$$

for every $x \in U$.

Now we can also state the Newton method for semismooth operator equations.

Algorithm 7: Newton method for semismooth operator equations

Require: Initial guess $u_0 \in U$, $F : U \subset X \to Z$ generalized differentiable.

1: Set $k = 0$;
2: **while** stopping criteria is not satisfied **do**
3: Solve

$$G(u_k)d_k = -F(u_k), \tag{7.2}$$

where $G(u_k)$ is an arbitrary generalized derivative of F at u_k;
4: Set $x_{k+1} = x_k + d_k$, $k = k + 1$;
5: **end while**

Identical to Theorem 7.7 with adapting the function spaces we get the following superlinear convergence rate theorem.

Theorem 7.9 (Superlinear local convergence)
Suppose that $\bar{u} \in U \subset X$ satisfies $F(\bar{u}) = 0$, and $F : U \to Z$ is Newton differentiable in an open neighborhood $\tilde{U}$ containing $\bar{u}$, $G(u)$ is nonsingular for all $u \in \tilde{U}$ and $\{\|G(u)^{-1}\|_X : u \in \tilde{U}\}$ is bounded. Then Algorithm 2 is well-defined and converges superlinear to $\bar{u}$ provided that $\|u_0 - \bar{u}\|_X$ is sufficiently small.

In conclusion, we have presented a method for solving semismooth optimization problems, such as the optimal control problems with additional control constraints discussed in Section 3.2.3. In a MPC setting, model order reduction is crucial as it is necessary to solve a nonlinear system at every MPC step, which requires multiple Newton steps. This can lead to a significant amount of computation time for a large terminal time T. Developing a-posteriori error estimator for the reduced nonlinear model is a challenging task and is not studied in this master thesis.

Bibliography

[1] A. Antoulas and D.C. Sorensen (2009). *Approximation of Large-Scale Dynamical Systems: An Overview*. International Journal of Applied Mathematics, 709-716, doi:10.1137/1.9780898718713.

[2] S. Banholzer (2017). *POD-Based Bicriterial Optimal Control of Convection-Diffusion Equations*. Master thesis, University of Konstanz, `https://kops.uni-konstanz.de/handle/123456789/39948`.

[3] U. Baur and P. Benner (2008). *Cross-gramian based model reduction for data-sparse systems*. Electronic Transactions on Numerical Analysis, 256-270.

[4] J. T. Betts (2010). *Practical Methods for Optimal Control and Estimation Using Nonlinear Programming*. Advances in Design and Control, SIAM, Philadelphia, second edition, doi:10.1137/1.9780898718577.

[5] S. Casper, D.H. Fuertinger, P. Kotanko, L. Mechelli, J. Rohleff and S. Volkwein (2022). *Data-Driven Modeling and Control of Complex Dynamical Systems Arising in Renal Anemia Therapy*. `https://arxiv.org/abs/2106.11733`.

[6] S. Chaturantabut and D. Sorensen (2012). *A state space estimate for POD-DEIM nonlinear model reduction*. SIAM Journal on Numerical Analysis 50, 46-63, doi:10.1137/110822724.

[7] F.H. Clarke (1983). *Optimization and Nonsmooth Analysis*. Wiley, New York.

[8] R. Dautray, A. Craig, M. Artola, M. Cessenat, J. Lions, and H. Lanchon (1999). *Mathematical Analysis and Numerical Methods for Science and Technology: Volume 5 Evolution Problems I*. Mathematical Analysis and Numerical Methods for Science and Technology, Springer Berlin.

[9] R. Denk and R. Racke (2011). *Kompendium der Analysis, Band 1: Differential- und Integralrechnung, Gewöhnliche Differentialgleichungen*. Vieweg+Teubner Verlag, Springer Fachmedien Wiesbaden GmbH.

[10] R. Denk und R. Racke (2012). *Kompendium der Analysis. Band 2: Maß- und Integrationstheorie, Funktionentheorie, Funktionalanalysis, Partielle Differentialgleichungen*. Vieweg+Teubner Verlag, Springer Fachmedien Wiesbaden GmbH.

[11] L. Evans (1998). *Partial Differential Equations*. Graduate studies in mathematics, American Mathematical Society, Rhode Island.

[12] J. Garcia and J. Basilio (2002). *Computation of reduced-order models of multivariable systems by balanced truncation.* Int. J. Syst. Sci. 33, 847-854, doi:10.1080/00207720210000017308.

[13] L. Grüne and J. Pannek (2016). *Nonlinear Model Predictive Control: Theory and Algorithms.* Communications and Control Engineering, Springer, London, UK, doi:10.1007/978-0-85729-501-9.

[14] M. Gubisch and S. Volkwein (2017). *Proper orthogonal decomposition for linear-quadratic optimal control.* Computational Science and Engineering. Philadelphia: SIAM, chap.1. 3-63, doi:10.1137/1.9781611974829.ch1.

[15] J. Hesthaven, G. Rozza, and B. Stamm (2016). *Certified Reduced Basis Methods for Parametrized Partial Differential Equations.* SpringerBriefs in Mathematics (Cham: Springer), doi:10.1007/978-3-319-22470-1.

[16] C. Himpe (2018). *emgr – The empirical gramian framework. Algorithms.* doi:10.3390/a11070091.

[17] M. Hintermüller (2010). *Semismooth Newton Methods and Applications.* Lecture Notes Department of Mathematics Humboldt-University of Berlin, `https://www.math.uni-hamburg.de/home/hinze/Psfiles/Hintermueller_OWNotes.pdf`.

[18] M. Hintermüller, K. Ito and K. Kunisch (2006). *The primal-dual active set method as a semismooth Newton method.* SIAM.

[19] M. Hinze, R. Pinnau, M. Ulbrich and S. Ulbrich (2009). *Optimization with PDE Constraints.* Mathematical Modeling: Theory and Applications, Springer Science.

[20] M. Hinze and S. Volkwein (2008). *Error estimates for abstract linear-quadratic optimal control problems using proper orthogonal decomposition.* Computational Optimization and Applications, 39:319-345.

[21] T. Kailath (1980). *Linear systems.* (Englewood Cliffs, N.J. : Prentice-Hall).

[22] M. Korda and I. Mezic (2018). *Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control.* Automatica, 149-160, doi:10.1016/j.automatica.2018.03.046.

[23] K. Kunisch and S. Volkwein (2001). *Galerkin proper orthogonal decomposition methods for parabolic problems.* Numer. Math., 90:117-148, doi:10.1007/s002110100282.

[24] J.N. Kutz, S.L. Brunton, B.W. Brunton, and J.L. Proctor (2016). *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems.* SIAM, Philadelphia.

[25] S. Lall, J. E. Marsden and S. Glavaski (1999). *Empirical model reduction of controlled nonlinear systems.* IFAC Proceedings Volumes 32, 2598-2603, Beijing, Chia.

[26] T. Lassila, A. Manzoni, A. Quarteroni and G. Rozza (2014). *Model Order Reduction in Fluid Dynamics: Challenges and Perspectives.* MS&A - Modeling, Simulation and Applications, Springer.

[27] W. Liu and N. Yan (2001). *A posteriori error estimates for distributed convex optimal control problems.* Advances in Computational Mathematics 15,285-309, doi:10.1023/A:1014239012739.

[28] L. Mechelli (2019). *POD-based State-Constrained Economic Model Predictive Control of Convection-Diffusion Phenomena*, PhD thesis, University of Konstanz, `https://kops.uni-konstanz.de/handle/123456789/47538`.

[29] L. Mechelli, J. Rohleff and S. Volkwein (2023). *Model order reduction for optimality sysmtes through empirical Gramians.* Submitted to Frontiers in Applied Mathematics and Statistics.

[30] L. Mechelli and S. Volkwein (2019). *POD-based economic model predictive control for heat-convection phenomena.* In Numerical Mathematics and Advanced Applications ENUMATH 2017, eds. F. A. Radu, K. Kumar, I. Berre, J. M. Nordbotten, and I. S. Pop (Cham: Springer International Publishing), 663-671, doi:10.1007/978-3-319-96415-7 61.

[31] B. Moore (1981). *Principal component analysis in linear systems: controllability, observability, and model reduction.* IEEE Transactions on Automatic Control 26, 17-32, doi:10.1109/TAC.1981.1102568.

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

[33] J. Nocedal and S. Wright (2006). *Numerical Optimization.* Springer, 2nd edition.

[34] A.I. Propoi (1963). *Application of linear programming methods for the synthesis of automatic sampled-data systems.* Avtomat, 912-920.

[35] A. Quarteroni, A. Manzoni and F.Negri (2016). *Reduced Basis Methods for Partial Differential Equations: An Introduction.* UNITEXT - La Matematica per il 3+2 (Cham: Springer), doi:10.1007/442 978-3-319-15431-2.

[36] S. Rogg, S. Trenz and S. Volkwein (2017). *Trust-Region POD using A-Posteriori Error Estimation for Semilinear Parabolic Optimal Control Problems.* Konstanzer Schriften in Mathematik, `http://nbn-resolving.de/urn:nbn:de:bsz:352-0-401106`.

[37] J. Rohleff (2020). *An incremental approach to dynamic mode decompo- sition for time-varying systems with applications to a model for erythropoiesis.* Bachlor thesis, University of Konstanz, `https://kops.uni-konstanz.de/handle/123456789/51127`.

[38] W.E. Schiesser (1991). *The Numerical Method Of Lines. Integration of Partial Differential Equations.* San Diego, Academic Press.

[39] P.J. Schmid (2010). *Dynamic mode decomposition of numerical and experimental data.* Journal of Fluid Mechanics, 656:5–28.

[40] D. Sorensen and A. Antoulas (2002). *The sylvester equation and approximate balanced reduction*. Linear Algebra Appl. 351/352, 671–700, doi:10.1016/S0024-3795(02)00283-5.

[41] S. Trenz (2017). *POD-Based A-posteriori Error Estimation for Control Problems Governed by Nonlinear PDEs*. PhD thesis, University of Konstanz, `https://kops.uni-konstanz.de/handle/123456789/40394`.

[42] F. Tröltzsch (2010). *Optimal Control of Partial Differential Equations: Theory, Methods, and Applications*. Graduate studies in mathematics, American Mathematical Society.

[43] F. Tröltzsch and S. Volkwein (2009). *POD a-posteriori error estimates for linear-quadratic optimal control problems*. Computational Optimization and Applications, 44:83–115, 10, doi:10.1007/s10589-008-9224-3.

[44] M. Ulbrich (2011). *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. Society for Industrial and Applied Mathematics, SIAM.

[45] D. Werner (2007). *Funktionalanalysis*. Springer-Lehrbuch, Springer Berlin Heidelberg.

[46] M. Villanueva, C. Jones, and B. Houska (2021). *Towards global optimal control via koopman lifts*. Automatica doi:10.48550/ARXIV.2003.01265.

[47] S. Volkwein (2021). *Optimization*. Lecture Notes Department of Mathematics and Statistics, Universty of Konstanz, 2021

[48] J. Wloka, C. Thomas, and M. Thomas (1987). *Partial Differential Equations*. Cambridge University Press.

[49] K. Zhou, J. Doyle and K. Glover (1996). *Robust and Optimal Control*. Upper Saddle River: Prentice-Hall.