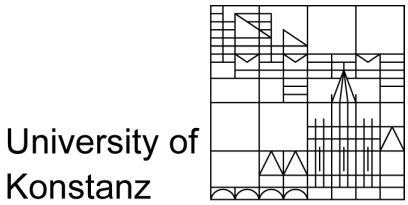
Master Thesis

POD-Based Bicriterial Optimal Control of Convection-Diffusion Equations

submitted by

Stefan Banholzer

at the



Department of Mathematics and Statistics

Konstanz, 10 February 2017

Supervisor and 1st Reviewer:	Prof. Dr. Stefan Volkwein, University of Konstanz
2nd Reviewer:	Prof. Dr. Michael Dellnitz, Paderborn University

Konstanzer Online-Publikations-System (KOPS)
URL: http://nbn-resolving.de/urn:nbn:de:bsz:352-0-421545

Erklärung der Selbstständigkeit

Ich versichere hiermit, dass ich die vorliegende Masterarbeit mit dem Thema:

POD-Based Bicriterial Optimal Control of Convection-Diffusion Equations

selbstständig verfasst und keine anderen Hilfsmittel als die angegebenen benutzt habe. Die Stellen, die anderen Werken dem Wortlaut oder dem Sinne nach entnommen sind, habe ich in jedem einzelnen Falle durch Angaben der Quelle, auch der benutzten Sekundärliteratur, als Entlehnung kenntlich gemacht.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Konstanz, den 10. Februar 2017

Stefan Banholzer

Abstract

In this thesis optimal control problems governed by linear convection-diffusion equations and bilateral control constraints are investigated. The optimal control problem is seen as a multiobjective optimization problem, with the objectives being the deviation of the state variable from a prescribed desired state on the one hand and the costs of the control function on the other hand. Therefore, techniques to handle multiobjective optimization problems are presented. As optimality notion the Pareto optimality is chosen and methods to provide Pareto optimal points are introduced. Analytical and geometrical properties are shown for these methods. The theoretical problem of how to get the set of all Pareto optimal points, the so-called Pareto front, is investigated by looking at two parameter-dependent method classes: the weighted sum method and reference point methods. A continuous dependency of the solution of Euclidean reference point problems on the reference points is proved. Based on that, a numerical algorithm to approximate the Pareto front using the Euclidean reference point method is proposed. The approximation quality is ensured by the algorithm which generates the reference points.
It is shown how the above mentioned optimal control problems can be transformed, such that they fit into the framework of multiobjective optimization. The Euclidean reference point method is applied to the transformed problems and the adjoint equation is introduced to get a numerically evaluateable representation of the derivatives of the cost function.
As the finite element discretization of the controlled partial differential equation (PDE) yields high dimensional equation systems, which have to be solved repeatedly, proper orthogonal decomposition (POD) is used to get a reduced-order approximation of the optimal control problem. A-priori convergence results of the solution of the reduced problem to the solution of the full problem and a-posteriori error estimates are shown.
Lastly, numerical experiments are presented to show the successful functioning of the presented algorithm and to evaluate the quality of the solutions of the model order reduced problem.

Contents

1 Introduction **1**

2 Basic Concepts **5**
2.1 Order on $\mathbb{R}^k$ . . . 5
2.2 Convexity . . . 6
2.3 Differentiability and Optimality . . . 7
2.4 Properties of Hilbert Spaces . . . 8
2.5 Weak Topology . . . 10
2.6 Partial Differential Equations . . . 10

3 Multiobjective Optimization Problems **13**
3.1 Pareto Optimality . . . 13
3.2 Scalarization Methods . . . 15
3.2.1 Analytical Results . . . 15
3.2.2 Geometrical Properties . . . 17
3.2.3 First-Order Optimality Condition . . . 18
3.3 Methods to compute the Pareto Front . . . 19
3.3.1 Weighted Sum Method . . . 19
3.3.2 Reference Point Methods . . . 24
3.3.3 Euclidean Reference Point Method . . . 26
3.4 Algorithm to approximate the Pareto Front in the Bicriterial Case . . . 34
3.4.1 Algorithm using the Euclidean Reference Point Method . . . 35

4 Proper Orthogonal Decomposition for Abstract Evolution Problems **39**
4.1 Model Order Reduction . . . 39
4.2 Proper Orthogonal Decomposition . . . 42
4.2.1 Continuous Version of the POD Method . . . 42
4.2.2 Discrete Version of the POD Method . . . 44

5 Bicriterial Optimal Control of Convection-Diffusion Equations **47**
5.1 Well-Posedness of the Linear Heat Equation with Convection Term . . . 48
5.1.1 Weak Formulation . . . 49
5.1.2 Well-Posedness . . . 50
5.1.3 Properties of the Solution Operator . . . 52
5.2 Reduction of the Optimal Control Problem . . . 53
5.3 The Optimal Control Problem as a Bicriterial Optimization Problem . . . 53
5.4 Adjoint Equation . . . 55
5.5 Applying POD to the Bicriterial Optimal Control Problem . . . 58
5.5.1 POD Approximation for the State Equation . . . 58

5.5.2 POD Approximation for the Cost Function 60
5.5.3 POD Approximation for the Adjoint Equation 62
5.5.4 A-Priori Convergence Analysis 64
5.5.5 A-Posteriori Analysis 66

6 Numerical Results 69
6.1 Description of the Implementation 70
6.1.1 Discretization of the PDEs 70
6.1.2 Solving the Scalar Optimization Problems 70
6.1.3 Details for Algorithm 1 71
6.2 Results for the Full Problem 71
6.2.1 Influence of the Convection Term 72
6.2.2 Controlling the Fineness of the Approximation 76
6.2.3 Quality of the Continuity Estimate on the Controls 78
6.3 Results for the POD Approximated Problem 81
6.3.1 Influence of the Convection Term 81
6.3.2 Quality of the POD A-Posteriori Error Estimate 82
6.3.3 Adaptive POD Basis Extension 86
6.3.4 Comparison of the Computational Time 88

7 Conclusion and Outlook 91

Bibliography 94

Chapter 1

Introduction

Originating from the calculus of variations, optimal control has developed to an own, vivid research field in modern applied mathematics in the last decades. In [36] the authors argue that the corner stone of optimal control was laid in the year 1696, when Johann Bernoulli (1667-1748) published the famous Brachystochrone problem (see [35, pp. 391-399]). However, it was not before the works of Richard Ernest Bellman (1920-1984) in 1957 about dynamic programming [3] and Lev Semyonovich Pontryagin (1908-1988) in 1962 on the maximum principle [29] that the modern optimal control theory was born. In the following years the theory of optimal control experienced a big enhancement, mainly because computers, with which the solving of more complex problems became possible, were spreading (see [5]). As a result optimal control is since being successfully used in many fields, for instance astronautics, medicine, engineering and economy.

Roughly speaking, the idea of optimal control is to influence a process by an input such that it reaches a desired prescribed outcome or is at least as close as possible to it. From this description it can already be seen that many problems can be put in the framework of optimal control.
In mathematical terms an optimal control problem contains a control variable u, i.e. the input, from a set of admissible controls U_{ad}, acting on a state variable y, i.e. the outcome. This 'acting' describes the process and leads mathematically to the so-called state equation, which reads in its general form

$$A(y) = B(u). \tag{1.1}$$

In many applications the process can be modelled by an ordinary or a partial differential equation, implying that the state equation (1.1) is a differential equation. In these cases one speaks of optimal control of ordinary or partial differential equations, see for instance [21, 39, 38] for standard works in this field.
The optimal control consists mathematically of minimizing a cost functional depending on the state y and the control u.
At this point it is already clear that the theory of optimal control of ordinary or partial differential equations is a manifold field of research, as it has to combine aspects from constrained optimization, ordinary or partial differential equations and numerical methods to deal with these problems.

Additionally, in many applications optimal control problems are not just about controlling one but several outcomes, which are all supposed to be close to a corresponding desired outcome. Besides, most of the time it is desirable to use as little control input as possible to reach the desired states. This might be due to resource issues, e.g. if the control stands for some sort of energy, or the wish to not influence the process too much, e.g. in case of a medicament acting on a body. Consequently, this implies several cost functions which have to be minimized in the

course of the optimal control. Therefore, it is natural to consider optimal control problems in the framework of multiobjective optimization.

In multiobjective optimization the functions to be optimized are not scalar but vector-valued. We say that they have multiple objectives. In the case of an optimal control problem the cost functions of the outcomes and the control can thus be seen as the different objectives of a function.
Unfortunately, most of the time there is not one single minimizer of all objective functions, i.e. the objectives are conflicting and an optimal compromise between the different objectives has to be found. This is why the concept of a decision maker is introduced. The purpose is that a decision maker has a deeper insight or knowledge of the problem at hand and can thus, given a set of possible compromises, decide which one is the best decision in the current situation. For this reason the goal of multiobjective optimization is not to present a single optimal compromise to the decision maker, but to provide as many optimal compromises as possible, such that the decision maker can make a decision based on these data.
So far it has not been explained what an optimal compromise actually is. One definition of optimality in this context goes back to Francis Ysidro Edgeworth (1845-1926) and Vilfredo Federico Pareto (1848-1923). However, today it is mostly known by the name of Pareto optimality. Pareto introduced it to describe optimal distribution of individual welfare:

> "We will say that the members of a collectivity enjoy maximum ophelimity in a certain position when it is impossible to find a way of moving from that position very slightly in such a manner that the ophelimity enjoyed by each of the individuals of that collectivity increases or decreases. That is to say, any small displacement in departing from that position necessarily has the effect of increasing the ophelimity which certain individuals enjoy, and decreasing that which others enjoy, of being agreeable to some and disagreeable to others." ([27, p. 261])

In the framework of multiobjective optimization this means that a point is Pareto optimal, or is an optimal compromise, if it is not possible to make one objective better without making at least one other objective worse. This notion will be used in the course of the thesis.
During the last century many solution methods for multiobjective optimization problems, which are based on this or similar solution concepts, have been developed. The most popular approach is to transform the multiobjective function into a scalar function, which can be optimized by using well-known techniques from scalar optimization. Popular scalarizations are for instance the weighted sum method, the ε-constraint method and reference point methods (see [11, 24, 34]).

In optimal control many works implicitly deal with multiobjective optimal control problems by considering a weighted sum of the different cost functions, see e.g. [21, 39, 38]. However, these works focus more on the theory of optimal control problems than on the embedding of those into multiobjective optimization. In particular, the intention is to present techniques to solve a single optimal control problem and not to provide the set of all Pareto optimal controls and points, respectively.
Consequently, the numerical treatment of optimal control problems viewed as multiobjective optimization problems is quite new. An investigation of nonlinear multiobjective optimal control problems was done in [22]. For solving the optimal control problem shooting methods ([4, 32])

were applied and normal boundary methods ([7]) were used to deal with the multiobjective optimization. In [26] the (Euclidean) reference point method (see [24, 31]) was used in the context of space mission design.

More recently model order reduction techniques have been applied to multiobjective optimal control problems governed by differential equations.
As explained above, approximating the set of all Pareto optimal controls implies consecutively solving many scalar optimal control problems. In the course of these optimization processes the arising state equation, which is in this case a differential equation, has to be solved frequently with different control inputs. Using standard methods as for example the finite element method for partial differential equations, this quickly gets too time consuming because the arising equation systems are high-dimensional. The idea of model order reduction methods is to generate low-dimensional spaces, in which the differential equations are solved. These spaces are supposed to contain the most important characteristics of expected solutions of the differential equation, which enables a good approximation quality despite their low dimensions. A thorough introduction into the topic can be found in [33], where the various applications of model order reduction are presented.
There are different approaches on how to construct these reduced-order spaces, for instance the reduced basis method (see [14]). Especially in the framework of nonlinear optimal control the proper orthogonal decomposition (POD) is currently a very popular approach. In the POD approach solutions of the dynamical system at predefined time-instances are taken to construct the so-called snapshot space. The leading eigenfunctions of a singular value decomposition are then chosen as a basis of the reduced-order space, see e.g. [15]. This method has shown to have good properties in the framework of optimal control especially due to available a-posteriori error estimates, see e.g. [37, 20, 1].

In [17] the reduced basis method was applied to a multiobjective optimal control problem governed by a linear elliptic partial differential equation and a POD-greedy algorithm was used in [16] for investigating a multiobjective optimal control problem governed by a semilinear parabolic partial differential equation. To handle the multiobjective optimization, the weighted sum method was used in both publications [17, 16]. In [28] the POD method was used to tackle a multiobjective optimal control problem arising in the field of fluid dynamics. Herein, the (Euclidean) reference point method was utilised for solving the multiobjective optimization problem.
A multiobjective optimal control problem governed by a linear heat equation was considered in [2]. It models an energy efficient heating, ventilation and air conditioning (HVAC) operation of a building with the conflicting objectives of comfort, i.e. reaching a desired temperature distribution, and minimal energy consumption, i.e. minimal heating effort, see e.g. [13]. The multiobjective optimization problems are dealt with by applying the (Euclidean) reference point method.

In this thesis we extend the problem presented in [2]. Instead of considering a linear heat equation, a convection-diffusion equation is used to model an additional airflow in the building. A thorough analysis of convex multiobjective optimization problems is presented to underline the suitability of the Euclidean reference point method in the framework of multiobjective optimal

control. In detail, we extend continuity results for the Euclidean reference point method from [2], which ensure a uniform approximation of the set of all Pareto optimal outcomes. An algorithm to numerically compute a good approximation of the set of Pareto optimal points similar to the ones in [28, 2] is proposed.
In the same manner as in [16, 28, 2] it is shown how to apply the POD method to the optimal control problem and results concerning the convergence of the solutions of the POD approximated problem to the solutions of the full problem as well as efficient a-posteriori error estimates are proved.

The thesis is organized as follows:
In Chapter 2 important concepts and results from various mathematical fields, which will be needed in the further course of the thesis, are presented.
Chapter 3 deals with multiobjective optimization problems. The notion of Pareto optimality is introduced mathematically. Afterwards the scalarization method to solve a multiobjective optimization problem is presented and both analytical and geometrical properties are derived. In the further course we have a closer look at two parameter-dependent method classes for approximating all Pareto optimal solutions, the weighted sum method and the reference point methods, and in particular the Euclidean reference point method. Theoretical results are shown that demonstrate that the Euclidean reference point method is suitable for approximating the Pareto front. The previous results are then used to generate an algorithm to approximate the Pareto front using the Euclidean reference point method.
In Chapter 4 the concept of model order reduction for evolution equations is introduced. We show a result concerning the approximation quality of general model order reduction methods and focus in the second part of the chapter on the method of proper orthogonal decomposition (POD). Theoretical results for the continuous version of POD are shown and the numerical applicability is demonstrated by introducing the discrete version of POD.
Chapter 5 is concerned with the bicriterial optimal control of a linear heat equation with convection term. To be able to show that the problem can be put in the framework of multiobjective optimization from Chapter 3, properties such as the unique solvability of the underlying partial differential equation are proved. In the further course it is shown how to apply the POD method from Chapter 4 to the optimal control problem at hand. A-priori convergence and a-posteriori error estimates are shown in the last part of the chapter.
The results of the numerical implementation of the problem of Chapter 5 using the algorithm proposed in Chapter 3 are presented in Chapter 6. After clarifying implementation details, we present results for the full problem in the first part and discuss the approximation quality of the POD method in the second part. Thereby, the focus is on investigating the influence of the strength of the convection on the solutions of the optimal control problem and testing the quality of the theoretical results and estimates. The quality of the POD approximation is tested in dependence of the strength of the convection and the dimension of the POD space. Furthermore, the efficiency of the a-posteriori error estimate is determined. These insights are then used to propose an adaptive POD basis extension algorithm.
Finally, Chapter 7 will first give a conclusion of the obtained results and then present a brief outlook on interesting subsequent questions.

Chapter 2

Basic Concepts

2.1 Order on $\mathbb{R}^k$

Let $k \in \mathbb{N}$ be arbitrary in this section.

Definition and Remark 2.1. Let $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$.

(i) We write for $x, y \in \bar{\mathbb{R}}^k$

$$x \leq y :\Leftrightarrow \forall i \in \{1, \dots, k\} : x_i \leq y_i,$$
$$x < y :\Leftrightarrow \forall i \in \{1, \dots, k\} : x_i < y_i.$$

The set $(\mathbb{R}^k, \leq)$ is partially but not totally ordered. Note that in contrast to the scalar case, it is not $x \leq y \Rightarrow (x < y \text{ or } x = y)$. Therefore, we additionally define

$$x \lneqq y :\Leftrightarrow x \leq y \text{ and } x \neq y.$$

(ii) For $x \in \mathbb{R}^k$ we define

$$\mathbb{R}^k_{\geq x} := \{y \in \mathbb{R}^k \mid y \geq x\} \subset \mathbb{R}^k,$$
$$\mathbb{R}^k_{> x} := \{y \in \mathbb{R}^k \mid y > x\} \subset \mathbb{R}^k,$$
$$\mathbb{R}^k_{\gneqq x} := \{y \in \mathbb{R}^k \mid y \gneqq x\} \subset \mathbb{R}^k.$$

Analogously, we define $\mathbb{R}^k_{\leq x}$, $\mathbb{R}^k_{< x}$ and $\mathbb{R}^k_{\lneqq x}$. For convenience, we additionally write $\mathbb{R}^k_{\geq} := \mathbb{R}^k_{\geq 0}$, $\mathbb{R}^k_{>} := \mathbb{R}^k_{>0}$ and $\mathbb{R}^k_{\gneqq} := \mathbb{R}^k_{\gneqq 0}$ and in the same manner we define $\mathbb{R}^k_{\leq}$, $\mathbb{R}^k_{<}$ and $\mathbb{R}^k_{\lneqq}$.

(iii) Lastly, we define for $x, y \in \bar{\mathbb{R}}^k$ with $x \leq y$

$$(x, y) := \{z \in \mathbb{R}^k \mid x < z < y\},$$
$$[x, y] := \{z \in \mathbb{R}^k \mid x \leq z \leq y\}.$$

Analogously, $(x, y]$ and $[x, y)$ are defined. Note that it is $(x, y) = (x_1, y_1) \times \cdots \times (x_k, y_k)$ and that $[x, y], (x, y]$ and $[x, y)$ have analogue representations.

Definition 2.2. Let $X, Y \subset \mathbb{R}^k$ be two arbitrary sets. We say that X lies below Y, if

$$\forall x \in X \, \forall y \in Y : \neg(y \lneqq x) \text{ and}$$
$$\forall y \in Y \, \exists x \in X : x \leq y$$

hold.

Definition 2.3. Let $g : \mathbb{R}^k \to \mathbb{R}$ be a function. We call g

(i) monotonically increasing, if

$$\forall x, y \in \mathbb{R}^k : (x \leq y \Rightarrow g(x) \leq g(y)) .$$

(ii) strictly monotonically increasing, if it is monotonically increasing and

$$\forall x, y \in \mathbb{R}^k : (x < y \Rightarrow g(x) < g(y)) .$$

2.2 Convexity

Definition 2.4. Let X be a vector space.

(i) $M \subset X$ is called convex, if

$$\forall x, y \in M \, \forall \lambda \in (0,1) : \lambda x + (1-\lambda) y \in M.$$

(ii) A function $f : M \subset X \to \mathbb{R}$, where M is a convex set, is called convex, if

$$\forall x, y \in M \, \forall \lambda \in (0,1) : f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y).$$

The function f is called strictly convex, if $<$ holds for all $x, y \in M$ with $x \neq y$ and all $\lambda \in (0,1)$.

Lemma 2.5. *Let $(X, \langle \cdot, \cdot \rangle_X)$ be a real vector space with inner product and $\|\cdot\|_X$ the norm on X induced by the inner product as well as $K \subset X$ convex. Then the squared norm restricted on K*

$$\|\cdot\|_X^2 : K \to \mathbb{R}, \; x \mapsto \|x\|_X^2$$

is strictly convex.

Proof. Let $x, y \in K$ with $x \neq y$ and $\lambda \in (0,1)$ be arbitrary. Then it holds

$$\begin{aligned}
&\|\lambda x + (1-\lambda) y\|_X^2 - \lambda \|x\|_X^2 - (1-\lambda) \|y\|_X^2 \\
&= \langle \lambda x + (1-\lambda) y, \lambda x + (1-\lambda) y \rangle_X - \lambda \langle x, x \rangle_X - (1-\lambda) \langle y, y \rangle_X \\
&= \lambda^2 \langle x, x \rangle_X + 2\lambda(1-\lambda) \langle x, y \rangle_X + (1-\lambda)^2 \langle y, y \rangle_X - \lambda \langle x, x \rangle_X - (1-\lambda) \langle y, y \rangle_X \\
&= \lambda(\lambda - 1) \langle x, x \rangle_X + (1-\lambda)(1-\lambda-1) \langle y, y \rangle_X + 2\lambda(1-\lambda) \langle x, y \rangle_X \\
&= \lambda(\lambda - 1) \left(\langle x, x \rangle_X + \langle y, y \rangle_X - 2 \langle x, y \rangle_X \right) \\
&= \lambda(\lambda - 1) \langle x - y, x - y \rangle_X < 0,
\end{aligned}$$

as $\lambda \in (0,1)$ and $x \neq y$. □

Lemma 2.6. *Let X, Y be two vector spaces. If $f : Y \to \mathbb{R}$ is strictly convex, $S : X \to Y$ linear and injective as well as $T : X \to Y$, $x \mapsto Sx + c$ for a $c \in Y$ an affine linear mapping, then $f \circ T : X \to \mathbb{R}$ is also strictly convex.*

Proof. Let $x, y \in X$ with $x \neq y$ and $\lambda \in (0,1)$ be arbitrary. As S is injective, it holds $S(x) \neq S(y)$ and then the strict convexity of f yields

$$\begin{aligned}(f \circ T)(\lambda x + (1-\lambda)y) &= f(S(\lambda x + (1-\lambda)y) + c) \\ &= f(\lambda(S(x)+c) + (1-\lambda)(S(y)+c)) \\ &< \lambda f(S(x)+c) + (1-\lambda) f(S(y)+c) \\ &= \lambda (f \circ T)(x) + (1-\lambda)(f \circ T)(y).\end{aligned}$$

□

Theorem 2.7 (Theorem of Hahn-Banach; Separation Version II)**.** *Let $(X, \|\cdot\|_X)$ be a normed vector space, $U, V \subset X$ disjoint and convex as well as U open. Then there is $x' \in X'$ with*

$$Re\, x'(u) < Re\, x'(v)$$

for all $u \in U$ and $v \in V$.

Proof. A proof can be found in [40, p. 103]. □

A direct consequence of Theorem 2.7 in the case of an Euclidean space is the following hyperplane separation result.

Corollary 2.8. *Let $U, V \subset \mathbb{R}^n$ be two disjoint and convex sets as well as U open. Then there are $\alpha \in \mathbb{R}^n \setminus \{0\}$ and $c \in \mathbb{R}$ such that*

$$\langle v, \alpha \rangle_{\mathbb{R}^n} \geq c \geq \langle u, \alpha \rangle_{\mathbb{R}^n}$$

holds for all $u \in U$ and $v \in V$.

Proof. This follows immediately from Theorem 2.7 by setting $c := \sup_{u \in U} x'(u)$ and the identification $\mathbb{R}^n \cong (\mathbb{R}^n)'$. □

2.3 Differentiability and Optimality

In this section we want to introduce two concepts of differentiability in general spaces – Gâteaux and Fréchet differentiability.

Definition 2.9. Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be two normed vector spaces, $U \subset X$ an open set and $f : U \to Y$ a function.

(i) The function f is called Gâteaux differentiable in $u \in U$, if there is a linear and continuous operator $A : X \to Y$ such that

$$\lim_{t \to 0} \frac{f(u+th) - f(u) - Ah}{t} = 0$$

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

(ii) f is called Fréchet differentiable in $x \in U$, if there is a linear and continuous operator $A : X \to Y$ such that

$$\lim_{\|h\|_X \to 0} \frac{\|f(u+h) - f(u) - Ah\|_Y}{\|h\|_X} = 0.$$

The operator A is called the Fréchet derivative of f in u and we write $f'(u) := A \in L(X,Y)$. If f is Fréchet differentiable for every $u \in U$, we call the function f Fréchet differentiable and the function $f' : U \to L(X,Y)$ the Fréchet derivative of f.

Remark 2.10. (i) The definition of Gâteaux differentiable is a generalization of the usual directional derivative, whereas the definition of Fréchet differentiable generalizes the usual derivative. Thus, it is easy to see that Fréchet differentiable implies Gâteaux differentiable. However, not every Gâteaux differentiable function is Fréchet differentiable.

(ii) The usual derivation rules hold for both Gâteaux and Fréchet derivatives.

The following result deals with the connection between the minimality of a Gâteaux differentiable function and the behaviour of its derivative. It provides a necessary, and in some cases even sufficient, first-order condition on the minimizer of a function.

Lemma 2.11. *Let $(X, \|\cdot\|_X)$ be a real Banach space, $K \subset U$ a convex set and $f : \tilde{K} \supset K \to \mathbb{R}$ for an open superset $\tilde{K}$ of K a Gâteaux-differentiable real-valued functional. If $\bar{u}$ is the minimizer of $f|_K$, then it fulfils*

$$f'(\bar{u})(u - \bar{u}) \geq 0 \text{ for all } u \in K. \tag{2.1}$$

If f is additionally convex and $\bar{u} \in K$ solves (2.1), then $\bar{u}$ is a minimizer of $f|_K$.

Proof. This lemma is proved in [39, Lemma 2.21]. □

In the course of this thesis the notion of a gradient becomes important. The definition of a gradient is clear for a function $f : \mathbb{R}^n \to \mathbb{R}$. By applying Riesz representation theorem, it is possible to generalize the definition of a gradient to functions $f : X \to \mathbb{R}$, where $(X, \langle\cdot,\cdot\rangle_X)$ is a real Hilbert space.

Definition 2.12. Let $(X, \langle\cdot,\cdot\rangle_X)$ be a real Hilbert space, $U \subset X$ open and $f : U \to \mathbb{R}$ a Fréchet differentiable function with Fréchet derivative $f' : U \to X'$.
The gradient of f in $u \in U$ is given by the Riesz representation of $f'(u)$ in X, i.e. by the unique element $v_u \in X$ for which $f'(u)h = \langle v_u, h\rangle_H$ holds for all $h \in X$. We write $\nabla f(u) := v_u$ and call the mapping $\nabla f : U \to X$, $u \mapsto \nabla f(u)$ the gradient of f.

2.4 Properties of Hilbert Spaces

Theorem 2.13 (Approximation Theorem)**.** *Let $(X, \langle\cdot,\cdot\rangle_X)$ be a Hilbert space, $M \subset X$ non-empty, convex and closed, as well as $z_0 \in X$. Then there is exactly one $x \in M$ with $\|x - z_0\|_X = dist(z_0, M) := \inf\{\|y - z_0\|_X \mid y \in M\}$.*

Proof. A proof can be found in [10, Theorem 12.22]. □

Definition and Theorem 2.14. Let $(X, \langle\cdot,\cdot\rangle_X)$ be a Hilbert space and M as in Theorem 2.13. Then the mapping $\mathcal{P}_M : X \to M$, where $\mathcal{P}_M(z) \in M$ is the unique element that fulfils $\|\mathcal{P}_M(z) - z\|_X = \operatorname{dist}(z, M)$, is called the projection on M. For an arbitrary $z \in X$ the projection $\mathcal{P}_M(z)$ is given by the unique solution to the sufficient and necessary condition

$$\langle \mathcal{P}_M(z) - z, y - \mathcal{P}_M(z)\rangle_X \geq 0 \text{ for all } y \in M. \tag{2.2}$$

Furthermore, the mapping $\mathcal{P}_M$ is Lipschitz continuous with Lipschitz constant 1.

Proof. Let $z \in X$ be arbitrary. To prove that (2.2) is a sufficient and necessary condition for the projection $\mathcal{P}_M(z)$ of z onto M, we notice that $\|\mathcal{P}_M(z) - z\|_X = \operatorname{dist}(z, M)$ is equivalent to the statement that the function

$$f_y : [0,1] \to \mathbb{R},\ f_y(t) := \|\mathcal{P}_M(z) - z - t(\mathcal{P}_M(z) - y)\|_X^2$$

has its minimizer at $t = 0$ for all $y \in M$. As $f_y'(t) = 2\langle \mathcal{P}_M(z) - z, y - \mathcal{P}_M(z)\rangle_X + 2t\,\|\mathcal{P}_M(z) - y\|_X^2$, this is equivalent to (2.2) for all $y \in M$ by Lemma 2.11.
To show that the mapping $\mathcal{P}_M$ is Lipschitz continuous with Lipschitz constant 1, let $z_1, z_2 \in X$ be arbitrary. Then it holds

$$\langle \mathcal{P}_M(z_1) - z_1, y - \mathcal{P}_M(z_1)\rangle_X \geq 0 \text{ for all } y \in M, \tag{2.3}$$
$$\langle \mathcal{P}_M(z_2) - z_2, y - \mathcal{P}_M(z_2)\rangle_X \geq 0 \text{ for all } y \in M. \tag{2.4}$$

Plugging $y = \mathcal{P}_M(z_2)$ into (2.3) and $y = \mathcal{P}_M(z_1)$ into (2.4) and adding both inequalities yields

$$\begin{aligned} 0 &\leq \langle \mathcal{P}_M(z_1) - z_1, \mathcal{P}_M(z_2) - \mathcal{P}_M(z_1)\rangle_X + \langle \mathcal{P}_M(z_2) - z_2, \mathcal{P}_M(z_1) - \mathcal{P}_M(z_2)\rangle_X \\ &= \langle \mathcal{P}_M(z_1) - \mathcal{P}_M(z_2) + z_2 - z_1, \mathcal{P}_M(z_2) - \mathcal{P}_M(z_1)\rangle_X \\ &= -\|\mathcal{P}_M(z_1) - \mathcal{P}_M(z_2)\|_X^2 + \langle z_2 - z_1, \mathcal{P}_M(z_2) - \mathcal{P}_M(z_1)\rangle_X. \end{aligned}$$

Now the claim follows by rearranging and using the Cauchy-Schwarz inequality. □

Definition and Theorem 2.15. Let $(X, \langle\cdot,\cdot\rangle_X)$ be a Hilbert space and $M \subset X$ a closed subspace. For each $x \in X$ there is a unique decomposition $x = m + m'$ with $m \in M$ and $m' \in M^\perp$. It holds $\|x - m\|_X = \min_{y \in M} \|x - y\|_X$. The mapping $\mathcal{P} : X \to X$, $x \mapsto m$ is called the orthogonal projection of X on M. It is linear and continuous with $\|\mathcal{P}\|_{L(X)} \leq 1$.

Proof. This result follows from [10, Theorems 12.23 & 18.3]. □

Theorem 2.16. *Let $(X, \langle\cdot,\cdot\rangle_X)$ be a separable Hilbert space and $(\varphi_i)_{i\in\mathbb{N}} \subset X$ an orthonormal basis of X. Then it holds*

$$x = \sum_{i\in\mathbb{N}} \langle x, \varphi_i\rangle_X \varphi_i$$

for all $x \in X$.

Proof. A proof can be found in [10, Theorem 12.36]. □

Corollary 2.17. *Let $(X, \langle\cdot,\cdot\rangle_X)$ be a separable Hilbert space, $(\varphi_i)_{i=1}^n \subset X$ an orthonormal system and $M := \operatorname{span}\{\varphi_1, \ldots, \varphi_n\}$. Then the orthogonal projection of X on M is given by $\mathcal{P}(x) = \sum_{i=1}^n \langle x, \varphi_i\rangle_X \varphi_i$.*
In particular, if $(\varphi_i)_{i\in\mathbb{N}} \subset X$ is an orthonormal basis of X, $M^n := \operatorname{span}\{\varphi_1, \ldots, \varphi_n\}$ and $\mathcal{P}^n$ is the orthogonal projection on M^n for all $n \in \mathbb{N}$, it holds $\|x - \mathcal{P}^n x\|_X \to 0$ for $n \to \infty$.

2.5 Weak Topology

Definition 2.18. Let $(X, \|\cdot\|_X)$ be a normed vector space. A set $U \subset X$ is called weakly closed, if it holds

$$\forall\, (u_n)_{n\in\mathbb{N}} \subset U \,:\, [(\exists u \in X \,:\, u_n \rightharpoonup u \,(n \to \infty)) \Rightarrow u \in U]\,.$$

Lemma 2.19. *Let $(X, \|\cdot\|_X)$ be a Banach space and $U \subset X$ convex. Then U is closed if and only if U is weakly closed.*

Proof. It is clear that weakly closed implies closed. For the other direction see for example [40, Satz III.3.8]. □

Lemma 2.20. *If $(X, \|\cdot\|_X)$ is a reflexive Banach space and $(x_n)_{n\in\mathbb{N}} \subset X$ a bounded sequence, there exists a subsequence $(x_{n_j})_{j\in\mathbb{N}}$ of $(x_n)_{n\in\mathbb{N}}$ and a $x \in X$ with $x_{n_j} \rightharpoonup x$, i.e. $(x_n)_{n\in\mathbb{N}}$ has a weakly convergent subsequence.*

Proof. For a proof see [40, Theorem III.3.7]. □

Definition 2.21. Let $(X, \|\cdot\|_X)$ be a normed vector space. A function $f : X \to \mathbb{R}$ is called lower semi-continuous, if it holds

$$\Big(u_n \to u \text{ in } X \Rightarrow f(u) \leq \liminf_{n\to\infty} f(u_n)\Big)\,.$$

The function f is called weakly lower semi-continuous, if it holds

$$\Big(u_n \rightharpoonup u \text{ in } X \Rightarrow f(u) \leq \liminf_{n\to\infty} f(u_n)\Big)\,.$$

Remark 2.22. It is immediately clear that the continuity of a function $f : X \to \mathbb{R}$ implies its lower semi-continuity. In general, however, continuity of a function does not imply weak lower semi-continuity.

Lemma 2.23. *Let $(X, \|\cdot\|_X)$ be a normed vector space and $f : X \to \mathbb{R}$ a convex function. Then f is lower semi-continuous if and only if f is weakly lower semi-continuous.*

Proof. A proof can be found in [40, Lemma III.5.9]. □

2.6 Partial Differential Equations

Let $(V, \langle\cdot,\cdot\rangle_V)$ and $(H, \langle\cdot,\cdot\rangle_H)$ be two separable Hilbert spaces with $V \subset H$ dense. Furthermore, $V \subset H \subset V'$ are supposed to be a Gelfand triple. For the precise definition of a Gelfand triple, see for example [9, Definition 8.4]. For this thesis it is sufficient to know that by defining $V := H^1(\Omega)$ and $H := L^2(\Omega)$ for a domain $\Omega \subset \mathbb{R}^n$, we get a Gelfand triple $V \subset H \subset V'$.
In the following let additionally $T > 0$.

Definition and Remark 2.24. We define

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

Then $W(0,T)$ endowed with the inner product

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

is a Hilbert space (see [39, pp. 146-148]).

It will be shown that the space $W(0,T)$ is the natural space in which the types of partial differential equations that are studied in this thesis are examined.
An important reason to deal with the notion of a Gelfand triple is the following interpolation theorem.

Theorem 2.25. *If $\varphi \in W(0,T)$, then it holds $\varphi \in C([0,T];H)$ and the embedding*

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

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

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

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

Proof. A proof of this statement can be found in [9, Theorem 8.6]. □

The following result shows that the evaluation of a function on the boundary of its domain can be extended to functions of Sobolev spaces, if the domain fulfils certain smoothness conditions. This leads to the notion of a trace.

Theorem 2.26 (Trace Theorem)**.** *Let $1 \leq p < \infty$ and assume that $\Omega \subset \mathbb{R}^n$ is bounded and $\partial\Omega$ is C^1. 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 can be found in [12, pp. 258-259]. □

The aim is to introduce a solution concept for an abstract evolution equation of the form

$$y_t(t) - A(t)y(t) = f(t) \ (t \in (0,T)), \tag{2.5a}$$

$$y(0) = y_0, \tag{2.5b}$$

where $A : [0,T] \to L(V,V')$ is a time-dependent family of operators, $f : [0,T] \to V'$ is a function and $y_0 \in H$ stands for the initial condition. The equality in (2.5a) is to be understood as equality in V' for almost all $t \in (0,T)$.

Definition 2.27. The bilinear form $a : [0,T] \times V \times V \to \mathbb{R}$ corresponding to the family of operators $(A(t))_{t\in[0,T]}$ is defined by

$$a(t,\varphi,\psi) := -\langle A(t)\varphi,\psi\rangle_{V'\times V} \quad (t \in [0,T],\, \varphi,\psi \in V).$$

The bilinear form and the family of operators are called coercive, if there are positive constants $\alpha > 0$, $\beta \geq 0$ with

$$a(t,\varphi,\varphi) \geq \alpha \|\varphi\|_V^2 - \beta \|\varphi\|_H^2 \quad (t \in [0,T], \varphi \in V). \tag{2.6}$$

In case the coercivity condition (2.6) holds for $\beta = 0$, the bilinear form and the family of operators are called strictly coercive or V-elliptic.

The next theorem guarantees the well-posedness of the abstract evolution equation (2.5) under certain presumptions.

Theorem 2.28. *Let $A \in C([0,T], L(V,V'))$ be coercive. Then for all $f \in L^2(0,T;V')$ and $y_0 \in H$ there is exactly one solution $y \in W(0,T)$ of the abstract evolution equation (2.5). Furthermore, the mapping*

$$W(0,T) \to L^2(0,T;V') \times H,\ y \mapsto (y_t - Ay, y(0))$$

is an isomorphism of Hilbert spaces. In particular, the inverse mapping (i.e. the solution mapping) is linear and continuous. Therefore, there is a constant $C > 0$ such that

$$\|y\|_{W(0,T)} \leq C \left(\|f\|_{L^2(0,T;V')} + \|y_0\|_H \right) \tag{2.7}$$

holds for all $(f, y_0) \in L^2(0,T;V') \times H$, where $y \in W(0,T)$ is the solution of (2.5).

Proof. A proof can be found in [9, Theorem 8.9]. □

Chapter 3

Multiobjective Optimization Problems

The intention of this chapter is to establish the theory of multiobjective optimization problems. First of all, an optimality concept, the so-called Pareto optimality, is introduced in order to develop solution approaches. As briefly mentioned in the introduction, the purpose of multiobjective optimization is to present a set of optimal solutions to the decision maker, who can then decide according to his, her or its knowledge and insight into the problem on one of them. So in the further process of the chapter we intend to analyse methods which enable us to compute the set of optimal solutions or at least an approximation of it.

We first introduce the framework in which we will work for the rest of the chapter.
Let $(U, \langle \cdot, \cdot \rangle_U)$ be a real Hilbert space, $U_{\text{ad}} \subset U$ non-empty, convex and closed as well as $f_1, \ldots, f_k : U_{\text{ad}} \to \mathbb{R}$ real-valued functions for a $k \in \mathbb{N}$ with $k \geq 2$. Define furthermore the function $f : U_{\text{ad}} \to \mathbb{R}^k$, $f(u) := (f_1(u), \ldots, f_k(u))^T$.
In the following we want to deal with the optimization problem

$$\min_{u \in U_{\text{ad}}} \begin{pmatrix} f_1(u) \\ \vdots \\ f_k(u) \end{pmatrix}. \tag{MOP}$$

Definition 3.1. In the situation above we call

(i) the function f a multiobjective function and the optimization problem (MOP) a multiobjective optimization problem.

(ii) $U_{\text{ad}} \subset U$ the feasible set and a vector $u \in U_{\text{ad}}$ feasible. Furthermore, the space U is called the feasible space.

(iii) the space $\mathbb{R}^k$ the objective space, the functions f_i $(i = 1, \ldots, k)$ objective functions and define $Y := f(U_{\text{ad}}) \subset \mathbb{R}^k$ as the image of the feasible set under the function f and call it the objective feasible region. A vector $y \in Y$ is called objective vector.

3.1 Pareto Optimality

In contrast to a one-dimensional optimization problem it is a priori not clear how to define a solution of (MOP). In fact, there are several different concepts of solutions for a multiobjective optimization problem. In this thesis we will work with the notion of the so-called Pareto optimality.
Looking at the scalar case, a feasible vector $\bar{u} \in U_{\text{ad}}$ is a global minimizer of a function

$g : U_{\text{ad}} \to \mathbb{R}$, if $g(\bar{u}) \leq g(u)$ holds for all $u \in U_{\text{ad}}$. A reasonable definition of a minimizer for a multiobjective function cannot demand the same since we cannot expect that it is possible to minimize all functions $f_1, \ldots, f_k$ simultaneously. Instead, we observe that an equivalent definition of a global minimizer $\bar{u} \in U_{\text{ad}}$ of the function g in the scalar case is that there is no $\tilde{u} \in U_{\text{ad}}$ with $g(\tilde{u}) < g(\bar{u})$. A generalization of this definition to the multiobjective case leads to a reasonable minimality concept with possibly several minimizers.

Definition 3.2. (i) An objective vector $y \in Y$ is called Pareto optimal if there is no $\tilde{y} \in Y$ with $\tilde{y} \lneqq y$. The set $P_Y := \{y \in Y \mid y \text{ is Pareto optimal}\} \subset \mathbb{R}^k$ is called the Pareto front.

(ii) A feasible vector $u \in U_{\text{ad}}$ is called Pareto optimal if its corresponding objective vector $f(u) \in Y$ is Pareto optimal. We furthermore define the set $P_{U_{\text{ad}}} := \{u \in U_{\text{ad}} \mid u \text{ is Pareto optimal}\} \subset U_{\text{ad}}$ and call it the Pareto set.

Remark 3.3. We can define the notion of Pareto optimality for arbitrary sets $X \subset \mathbb{R}^k$ in exactly the same way as it is done for the objective feasible region in Definition 3.2 (i).

Definition 3.4. (i) We define the ideal objective vector y^{id} by

$$y^{\text{id}} := \inf_{u \in U_{\text{ad}}} f(u) := \begin{pmatrix} \inf_{u \in U_{\text{ad}}} f_1(u) \\ \vdots \\ \inf_{u \in U_{\text{ad}}} f_k(u) \end{pmatrix}.$$

(ii) The nadir objective vector $y^{\text{nad}} \in Y$ is defined by $y_i^{\text{nad}} := \sup_{y \in P_Y} y_i$ for $i \in \{1, \ldots, k\}$.

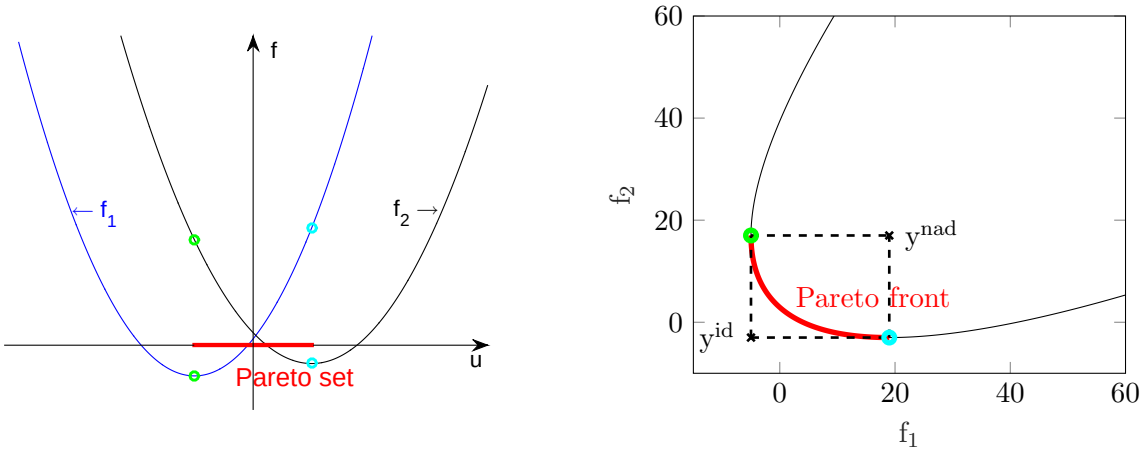
Remark 3.5. It holds $y^{\text{id}} \leq y$ for all $y \in P_Y$ and there is no $\hat{y} \gneqq y^{\text{id}}$ with $\hat{y} \leq y$ for all $y \in P_Y$. So in some sense the ideal objective vector is the infimum of the Pareto front. The nadir objective vector y^{nad} is the supremum of the Pareto front in the sense that $y \leq y^{\text{nad}}$ for all $y \in P_Y$ and there is no $\hat{y} \lneqq y^{\text{nad}}$ with the same property. In particular, we can conclude $P_Y \subset [y^{\text{id}}, y^{\text{nad}}]$.

To make the notion of the Pareto optimality clearer, we provide an easy example in the bicriterial case including two parabolas.

Example 3.6. Let $k = 2$, $U := \mathbb{R}$, $U_{\text{ad}} := [-4, 4] \subset \mathbb{R}$ and define the objective functions f_1 and f_2 by

$$f_1 : U_{\text{ad}} \to \mathbb{R},\ f_1(x) := 6(x+1)^2 - 5$$
$$f_2 : U_{\text{ad}} \to \mathbb{R},\ f_2(x) := 5(x-1)^2 - 3.$$

For this setting we want to illustrate the concept of Pareto optimality.
In Figure 3.1 (a) the graphs of the two objective functions are plotted. By definition a feasible vector $u \in U_{\text{ad}}$ is Pareto optimal, if it is not possible two lower both objective functions at the same time by looking at another feasible vector. Therefore, it is easy to see for this example that the Pareto set $P_{U_{\text{ad}}}$ is given by all $u \in U_{\text{ad}}$ for which $f_1'(u)$ and $f_2'(u)$ have a different sign, i.e. $P_{U_{\text{ad}}} = [-1, 1]$.
The corresponding Pareto front can be seen in Figure 3.1 (b), where the objective feasible region $Y := f(U_{\text{ad}})$ is plotted.



(a) Pareto set

(b) Pareto front with y^{id} and y^{nad}

Figure 3.1: Example for the concept of Pareto optimality using two parabolas

3.2 Scalarization Methods

After introducing a solution concept for multiobjective optimization problems, the obvious next question is how we can get Pareto optimal solutions. In literature there is a vast range of methods with which Pareto optimal points can be obtained. A good overview about these methods can be found in [11, 24, 34, 8].
In this thesis we focus on one specific type, namely the so-called scalarization methods. As the name suggests, the basic idea of this approach is to transform (MOP) into a scalar optimization problem, which we can handle with the well-known techniques for scalar optimization. The scalarization is done by composing an arbitrary function $g : \mathbb{R}^k \to \mathbb{R}$ with the multiobjective function f. In this way, we obtain the scalar optimization problem

$$\min_{u \in U_{\mathrm{ad}}} F^g(u) := (g \circ f)(u). \tag{SOP}$$

The function g can be seen as a cost function, as it assigns a real number – the cost – to each objective vector, enabling us to compare all objective vectors with each other. From a modelling point of view the decision maker's preferences can be incorporated into this cost function g.

3.2.1 Analytical Results

After presenting the idea of transforming a multiobjective optimization problem into a scalar optimization problem, immediately arising questions are:

(i) Are there functions g such that solving (SOP) provides us with Pareto optimal points?

(ii) If so, can all Pareto optimal points be obtained by solving (SOP) for a function g?

In this section we want to deal with the first question. More precisely we want to find properties of the function g that assure on the one hand a unique solvability of (SOP) and on the other

hand that the unique solution is indeed Pareto optimal. The next theorem provides us with sufficient conditions on F^g and g under which this is guaranteed.

Theorem 3.7. *Let $g : \mathbb{R}^k \to \mathbb{R}$ be a function such that $F^g : U_{ad} \to \mathbb{R}$ is strictly convex, lower semi-continuous and bounded from below. Additionally we assume $\lim_{\|u\|_U \to \infty} F^g(u) = \infty$ in the case that U_{ad} is not bounded. Then the optimization problem (SOP) is uniquely solvable. Furthermore, if g is monotonically increasing, the unique solution $\bar{u} \in U_{ad}$ of (SOP) is Pareto optimal.*

Proof. (i) Existence: As U_{ad} is convex and closed, we know by Lemma 2.19 that U_{ad} is weakly closed. Furthermore, the strict convexity and lower semi-continuity of F^g yield by using Lemma 2.23 that F^g is weakly lower semi-continuous.
As F^g is bounded from below, we can define $a := \inf_{u \in U_{\text{ad}}} F^g(u) > -\infty$. By the definition of an infimum there is a sequence $(u_n)_{n\in\mathbb{N}} \subset U_{\text{ad}}$ with $\lim_{n\to\infty} F^g(u_n) = a$. In the case that U_{ad} is bounded, the sequence $(u_n)_{n\in\mathbb{N}} \subset U_{\text{ad}} \subset U$ is automatically a bounded sequence in the Hilbert space U. If U_{ad} is unbounded, we know by assumption that $\lim_{\|u\|_U \to \infty} F^g(u) = \infty$ and thus we can conclude that the sequence $(u_n)_{n\in\mathbb{N}}$ is bounded in U as well. As U is a Hilbert space and hence reflexive, Lemma 2.20 yields that there is a subsequence $(u_{n_j})_{j\in\mathbb{N}}$ of $(u_n)_{n\in\mathbb{N}}$ and $\bar{u} \in U$ with $u_{n_j} \rightharpoonup \bar{u}$ in U. But then we automatically get $\bar{u} \in U_{\text{ad}}$ because U_{ad} is weakly closed. Finally the weak lower semi-continuity of F^g yields

$$F^g(\bar{u}) \leq \liminf_{j\to\infty} F^g(u_{n_j}) = a,$$

and thus $F^g(\bar{u}) = \inf_{u\in U_{\text{ad}}} F^g(u)$, so that $\bar{u}$ is a minimizer of F^g.

(ii) Uniqueness: The uniqueness of the solution follows from the strict convexity of the function F^g.

(iii) Pareto optimality: Now we suppose that g is monotonically increasing and want to show that the unique solution is Pareto optimal. Therefore, let $\bar{u} \in U_{\text{ad}}$ be the unique solution of (SOP) and let $u \in U_{\text{ad}}$ be such that $f(u) \leq f(\bar{u})$. As g is monotonically increasing, we can conclude that $F^g(u) \leq F^g(\bar{u})$ holds. But as $\bar{u}$ is the unique minimizer of F^g, we get $u = \bar{u}$ and of course $f(u) = f(\bar{u})$. Consequently, $\bar{u}$ is Pareto optimal.

□

In a next step we want to find sufficient conditions for the function g to ensure a unique solvability of (SOP). Therefore, we have to make some assumptions on the functions $f_1, \ldots, f_k$.

Assumption 1. *Assume that f_i is strictly convex, continuous and bounded from below for all $i \in \{1, \ldots, k\}$. In the case that U_{ad} is unbounded, suppose additionally that $\lim_{\|u\|_U \to \infty} f_i(u) = \infty$ for all $i \in \{1, \ldots, k\}$.*

Under Assumption 1 it is possible to show the following sufficient condition for the function g.

Lemma 3.8. *Let Assumption 1 be satisfied and let $g : \mathbb{R}^k_{\geq y^{id}} \to \mathbb{R}$ be a convex, lower semi-continuous function that is strictly monotonically increasing and in case the set U_{ad} is unbounded fulfils $\lim_{\|x\|_{\mathbb{R}^k} \to \infty} g(x) = \infty$. Then the optimization problem (SOP) is uniquely solvable and the unique solution $\bar{u} \in U_{ad}$ is Pareto optimal.*

Proof. First, we notice that the function F^g is well-defined as $f(U_{\text{ad}}) \subset \mathbb{R}^k_{\geq y^{\text{id}}}$. As g is strictly monotonically increasing, we can conclude from the proof of Theorem 3.7 that each minimizer of F^g is Pareto optimal. So we only have to check that the function F^g fulfils the properties demanded in Theorem 3.7 under the given assumptions.

(i) Strict Convexity: Let $u, v \in U_{\text{ad}}$ with $u \neq v$ and $\lambda \in (0,1)$ be arbitrary. Then we get

$$\begin{aligned} F^g(\lambda u + (1-\lambda)v) &= g(f(\lambda u + (1-\lambda)v)) \\ &< g(\lambda f(u) + (1-\lambda) f(v)), \end{aligned}$$

where we used the strict convexity of the functions f_i and that g is strictly monotonically increasing. Furthermore, we obtain by using the convexity of g

$$g(\lambda f(u) + (1-\lambda) f(v)) \leq \lambda g(f(u)) + (1-\lambda) g(f(v)) = \lambda F^g(u) + (1-\lambda) F^g(v),$$

and thus the strict convexity of the function F^g.

(ii) Lower semi-continuity: Let $(u_n)_{n\in\mathbb{N}} \subset U_{\text{ad}}$ be a sequence with $u_n \to u$ $(n \to \infty)$ for a $u \in U_{\text{ad}}$. As the functions $f_1, \ldots, f_k$ are continuous, it holds $f(u_n) \to f(u)$ $(n \to \infty)$. Now, the lower semi-continuity of g implies $F^g(u) \leq \liminf_{n\to\infty} F^g(u_n)$, so that F^g is lower semi-continuous.

(iii) Boundedness: The function g is strictly monotonically increasing, so that it holds $g(y^{\text{id}}) \leq g(y)$ for all $y \in \mathbb{R}^k_{\geq y^{\text{id}}}$, i.e. g is bounded from below. Consequently, also F^g is bounded from below.

(iv) Coercivity: In the case that U_{ad} is unbounded, we immediately obtain from the presumptions that $\lim_{\|u\|_U \to \infty} F^g(u) = \infty$.

□

For convenience we summarize the conditions on the function g stated in Lemma 3.8 in a presumption.

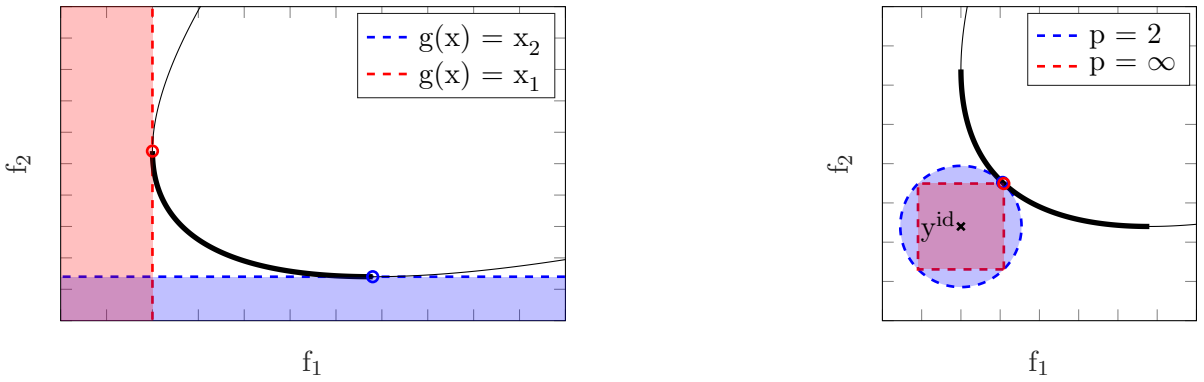
Presumption 1. *A function $g : \mathbb{R}^k_{\geq y^{id}} \to \mathbb{R}$ is said to fulfil Presumption 1, if g is a convex, lower semi-continuous function that is strictly monotonically increasing and in case the set U_{ad} is unbounded fulfils* $\lim_{\|x\|_{\mathbb{R}^k} \to \infty} g(x) = \infty$.

3.2.2 Geometrical Properties

For the further progress of this chapter not only analytical but also geometrical results will be of great importance. They will give us valuable insight into the geometry of the Pareto front, which will be used in constructing an algorithm to approximate the Pareto front.

Remark 3.9. If $\bar{u}$ is the unique minimizer of F^g, it holds by definition

$$g^{-1}((-\infty, F^g(\bar{u})]) \cap Y = \{f(\bar{u})\}.$$



(a) Minimizing a single objective

(b) p-norm

Figure 3.2: The sets $g^{-1}((-\infty, F^g(\bar{u})])$ for different scalarizing functions

By looking at the set $g^{-1}((-\infty, F^g(\bar{u})])$, we can thus gain a region in which there is no Pareto optimal point except for $f(\bar{u})$. Observe that the definition of a Pareto optimal point $f(\bar{u})$ only yields $(\{f(\bar{u})\} + \mathbb{R}^k_{\leq}) \cap Y = \{f(\bar{u})\}$. So knowing that a Pareto optimal point is the unique minimizer of F^g might provide us with some more geometrical insight into the shape of the Pareto front.
If Assumption 1 is satisfied and the function g fulfils Presumption 1, the preimages $g^{-1}((-\infty, a])$ are convex for all $a \in \mathbb{R}$. So due to the convexity of $g^{-1}((-\infty, F^g(\bar{u})])$ the best we can hope for is that $g^{-1}(\{F^g(\bar{u})\})$ is a hyperplane. Later we will see that this is the case for each Pareto optimal point under certain assumptions.

Example 3.10. Let Assumption 1 be satisfied.

(i) Let $g : \mathbb{R}^k_{\geq y^{\mathrm{id}}} \to \mathbb{R}$, $g(x) := x_i$ for any $i \in \{1, \ldots, k\}$. It is easy to check that the function g satisfies Presumption 1, so that the optimization problem (SOP) has a unique, Pareto optimal solution. Minimizing F^g in this case corresponds to minimizing f_i without looking at the other functions. It is interesting to notice that computing the minimizers of all functions f_i and evaluating the remaining functions at these minimizers already provides us with Pareto optimal points.

(ii) Let $g : \mathbb{R}^k_{\geq y^{\mathrm{id}}} \to \mathbb{R}$, $g(x) := \left\|x - y^{\mathrm{id}}\right\|_p$ for any $p \in [1, \infty]$. Then the function g fulfils Presumption 1.

Figure 3.2 (a) and (b) show the sets $g^{-1}((-\infty, F^g(\bar{u})])$ for the respective functions g. By Remark 3.9 we can conclude that these sets do not contain any Pareto optimal points but $f(\bar{u})$. In case of considering the scalarization to one objective by setting $g(x) := x_i$ in Figure 3.2 (a), we get optimal information as described in Remark 3.9: the set $g^{-1}(\{F^g(\bar{u})\})$ is a hyperplane.

3.2.3 First-Order Optimality Condition

For later investigations it is desirable that the function F^g is differentiable. The reason for this is amongst others, e.g. the existence of better optimization algorithms for differentiable functions,

that we can show the following first-order condition for the minimizer of F^g, which will provide us with some important analytical insight.

Assumption 2. *Assume that the functions $f_1, \dots, f_k$ are differentiable.*

Theorem 3.11. *Let Assumptions 1 and 2 be satisfied. Furthermore let the function $g : \mathbb{R}^k_{\geq y^{id}} \to \mathbb{R}$ fulfil Presumption 1 and be additionally differentiable. Then a necessary and sufficient first-order condition for the minimizer of the function F^g is given by*

$$\langle \nabla F^g(\bar{u}), u - \bar{u} \rangle_U = \langle \sum_{i=1}^{k} \partial_i g(f(\bar{u})) \nabla f_i(\bar{u}), u - \bar{u} \rangle_U \geq 0 \text{ for all } u \in U_{ad}. \tag{3.1}$$

Proof. The identity $\nabla F^g = \sum_{i=1}^k (\partial_i g \circ f) \nabla f_i$ can be verified by applying the chain rule. Furthermore, it follows from Lemma 2.11 that (3.1) is a necessary and sufficient condition for a minimizer of F^g. □

Corollary 3.12. *Let Assumptions 1 and 2 be satisfied. Furthermore, let the functions $g, h : \mathbb{R}^k_{\geq y^{id}} \to \mathbb{R}$ fulfil Presumption 1 and be additionally differentiable. Denote by $\bar{u}$ the unique minimizer of F^g. If there is a $\lambda > 0$ such that $\nabla g(f(\bar{u})) = \lambda \nabla h(f(\bar{u}))$ holds, $\bar{u}$ is also the unique minimizer of F^h.*

Proof. This follows directly from Theorem 3.11. □

3.3 Methods to compute the Pareto Front

So far we learned about sufficient conditions on the function g which assure us that (SOP) is uniquely solvable and that the solution is Pareto optimal. However, our goal in the end is to present the whole Pareto front or, in case of the numerical implementation later, at least a good approximation of the Pareto front to the decision maker. For this reason we need a strategy to generate a family of functions $(g_i)_{i \in I}$ for some index set I, such that solving $(\text{SOP})_i$ for all $i \in I$ provides us with all or at least sufficiently many well-distributed Pareto optimal points.
In this section we take a closer look at two different approaches, the first one being the weighted sum method (see e.g. [11, Chapter 3], [24, pp. 78-85]) and the second one being reference point methods, and in particular the Euclidean reference point method (see e.g. [24, pp. 164-170], [31]).

3.3.1 Weighted Sum Method

The idea of the weighted sum method is to provide the different objective functions with positive weights and then to minimize the sum of the weighted objective functions. This is probably the most intuitive way to couple the objective functions as one can directly specify how important it is to lower the i-th objective function f_i in comparison to the other objective functions by providing the ratio of the weights.
Abstractly we consider for some weights $\alpha \in \mathbb{R}^k_{\gneq}$ the weighted sum problem

$$\min_{u \in U_{\text{ad}}} \sum_{i=1}^{k} \alpha_i f_i(u). \tag{WSP}$$

Notation 3.13. For $\alpha \in \mathbb{R}^k_{\gneqq}$ we denote the weighted sum problem with weights α by $(\mathrm{WSP})_\alpha$. Furthermore, we define $F_\alpha : U_{\mathrm{ad}} \to \mathbb{R}$, $F_\alpha(u) := \sum_{i=1}^k \alpha_i f_i(u)$ for all $u \in U_{\mathrm{ad}}$.

Analytical Results

Lemma 3.14. *Let Assumption 1 be satisfied and $\alpha \in \mathbb{R}^k_{\gneqq}$. Then $(WSP)_\alpha$ has a unique Pareto optimal solution.*

Proof. It is easy to check that the function $g_\alpha : \mathbb{R}^k_{\geq y^{\mathrm{id}}} \to \mathbb{R}$, $g_\alpha(x) := \sum_{i=1}^k \alpha_i x_i$ fulfils Presumption 1, if $\alpha \in \mathbb{R}^k_{\gneqq}$ holds. □

Remark 3.15. Having the notation of the introduction to this section in mind, we consider the family of functions $(g_\alpha)_{\alpha \in \mathbb{R}^k_{\gneqq}}$ with $g_\alpha : \mathbb{R}^k_{\geq y^{\mathrm{id}}} \to \mathbb{R}$, $g_\alpha(x) := \sum_{i=1}^k \alpha_i x_i$ for all $\alpha \in \mathbb{R}^k_{\gneqq}$.

Corollary 3.16. *Let Assumptions 1 and 2 be satisfied. Let furthermore $\alpha \in \mathbb{R}^k_{\gneqq}$ be arbitrary. Then the first-order condition stated in Theorem 3.11 for the unique solution $\bar{u}$ of $(WSP)_\alpha$ reads*

$$\langle \sum_{i=1}^k \alpha_i \nabla f_i(\bar{u}), u - \bar{u} \rangle_U \geq 0 \text{ for all } u \in U_{ad}. \tag{3.2}$$

In a next step we want to investigate which points on the Pareto front can be obtained by solving a weighted sum problem. The main result will be that all Pareto optimal points can be computed by a weighted sum problem, if Assumption 1 is satisfied. To be able to show this result, we first have to establish some small statements.

Lemma 3.17. *Let $X \subset \mathbb{R}^k$ be arbitrary. Then $P_X = P_{X+\mathbb{R}^k_\geq}$, i.e. the set of Pareto optimal points does not change when adding $\mathbb{R}^k_\geq$ to the set X.*

Proof. Let $x \in P_X$ be arbitrary. So there is no $\tilde{x} \in X$ with $\tilde{x} \lneqq x$. It follows immediately that there is no $\hat{x} \in X + \mathbb{R}^k_\geq$ with $\hat{x} \lneqq x$.
Now let reversely $z \in P_{X+\mathbb{R}^k_\geq}$ be arbitrary, i.e. $z = x + y$ with $x \in X$ and $y \in \mathbb{R}^k_\geq$. As z is Pareto optimal, it is clear that $y = 0$ has to hold. But then $z \in X$ and because of $X \subset X + \mathbb{R}^k_\geq$ and the Pareto optimality of z in $X + \mathbb{R}^k_\geq$, we obtain that $z \in P_X$. □

Definition 3.18. A set $X \subset \mathbb{R}^k$ is called $\mathbb{R}^k_\geq$-convex, if the set $X + \mathbb{R}^k_\geq$ is convex.

Lemma 3.19. *Let $f_1, \ldots, f_k$ be convex and $Y := f(U_{ad})$. Then Y is $\mathbb{R}^k_\geq$-convex.*

Proof. Let $y_1, y_2 \in Y + \mathbb{R}^k_\geq$ and $\lambda \in (0,1)$ be arbitrary. So there are $u_1, u_2 \in U_{\mathrm{ad}}$ and $x_1, x_2 \in \mathbb{R}^k_\geq$ with $y_i = f(u_i) + x_i$ for $i = 1, 2$. Then we obtain by using the convexity of the functions $f_1, \ldots, f_k$

$$\begin{aligned} \lambda y_1 + (1-\lambda) y_2 &= \lambda f(u_1) + (1-\lambda) f(u_2) + \lambda x_1 + (1-\lambda) x_2 \\ &\geq f(\lambda u_1 + (1-\lambda) u_2) + \lambda x_1 + (1-\lambda) x_2. \end{aligned}$$

With $v := \lambda u_1 + (1-\lambda) u_2 \in U_{\mathrm{ad}}$ we get

$$\lambda y_1 + (1-\lambda) y_2 = f(v) + \tilde{x},$$

for a $\tilde{x} \in \mathbb{R}^k_\geq$, so that $\lambda y_1 + (1-\lambda) y_2 \in Y + \mathbb{R}^k_\geq$. □

The next theorem is the main theorem of this section. It states that all Pareto optimal points can be obtained by solving a weighted sum problem. The statement and the main idea of the proof are taken from [11, Theorem 3.5].

Theorem 3.20. *Let Assumption 1 be satisfied. If $\bar{y} = f(\bar{u}) \in P_Y$, then there is $\alpha \in \mathbb{R}^k_{\gneq}$ such that $\bar{u}$ is the unique solution of (WSP)$_\alpha$.*

Proof. Let $\bar{y} = f(\bar{u}) \in P_Y$ be arbitrary. Then by Lemma 3.17 we get $\bar{y} \in P_{Y+\mathbb{R}^k_\geq}$ and thus

$$\left(Y + \mathbb{R}^k_\geq - \bar{y}\right) \cap \mathbb{R}^k_< = \emptyset. \tag{3.3}$$

According to Lemma 3.19 the set $Y + \mathbb{R}^k_\geq$ is convex and consequently $Y + \mathbb{R}^k_\geq - \bar{y}$ is convex as well. Since $\mathbb{R}^k_<$ is a convex and open set, the hyperplane separation result in Corollary 2.8 yields the existence of an $\alpha \in \mathbb{R}^k \setminus \{0\}$ and a $c \in \mathbb{R}$ with

$$\langle y, \alpha \rangle_{\mathbb{R}^k} \geq c \geq \langle x, \alpha \rangle_{\mathbb{R}^k}$$

for all $y \in Y + \mathbb{R}^k_\geq - \bar{y}$ and all $x \in \mathbb{R}^k_<$. Because of $\langle x, \alpha \rangle_{\mathbb{R}^k} \leq c$ for all $x \in \mathbb{R}^k_<$ it follows $\alpha \geq 0$. But then we immediately obtain $c = 0$. Therefore, for all $y \in Y$ and all $x \in \mathbb{R}^k_\geq$ we get

$$\langle y + x - \bar{y}, \alpha \rangle_{\mathbb{R}^k} \geq 0,$$

and by setting $x = 0$ this yields

$$\langle y, \alpha \rangle_{\mathbb{R}^k} \geq \langle \bar{y}, \alpha \rangle_{\mathbb{R}^k}$$

for all $y \in Y$ and hence

$$\sum_{i=1}^k \alpha_i f_i(u) \geq \sum_{i=1}^k \alpha_i f_i(\bar{u})$$

for all $u \in U_{\text{ad}}$.
Thus, $\bar{u}$ is a solution of (WSP)$_\alpha$ and due to the unique solvability of this optimization problem $\bar{u}$ has to be the unique solution. □

In the following we want to investigate how the solution of a weighted sum problem depends on the weights α.

Lemma 3.21. *Let Assumption 1 be satisfied and assume that the functions $f_1, \ldots, f_k$ are additionally twice differentiable such that $\nabla^2 f_i$ is positive definite for all $i \in \{1, \ldots, k\}$ as well as that there exists $i_p \in \{1, \ldots, k\}$ such that $\nabla^2 f_{i_p}$ is uniformly positive definite with coercivity constant C_{i_p}. If $\alpha \in \mathbb{R}^k_{\gneq}$ with $\alpha_{i_p} > 0$, then $\nabla^2 F_\alpha$ is uniformly positive definite with coercivity constant $C_{i_p} \cdot \alpha_{i_p}$.*

Proof. Let $u \in U_{\text{ad}}$ be arbitrary. Then we get

$$\nabla^2 F_\alpha(u)v = \sum_{i=1}^k \alpha_i \nabla^2 f_i(u)v \text{ for all } v \in U,$$

and therefore

$$\begin{aligned}\langle \nabla^2 F_\alpha(u)v, v\rangle_U &= \sum_{i=1}^{k} \alpha_i \langle \nabla^2 f_i(u)v, v\rangle_U \\ &\geq C_{i_p} \cdot \alpha_{i_p} \|v\|_U^2 \text{ for all } v \in U,\end{aligned}$$

which is the claim. □

Assumption 3. *Assume that the functions $f_1, \dots, f_k$ are twice continuously differentiable such that $\nabla^2 f_i$ is positive definite for all $i \in \{1, \dots, k\}$ as well as that there exists $i_p \in \{1, \dots, k\}$ such that $\nabla^2 f_{i_p}$ is uniformly positive definite with coercivity constant C_{i_p}.*

The following result assures us that the unique solution of the weighted sum problem depends continuously on the weights, if Assumptions 1 and 3 are satisfied.

Theorem 3.22. *Let Assumptions 1 and 3 be satisfied and define $\mathcal{Z} := \{\alpha \in \mathbb{R}^k_{\gneq} \mid \alpha_{i_p} > \kappa\}$ for an arbitrary $\kappa > 0$. Then the mapping $\mathcal{Z} \to U_{ad}$, $\alpha \mapsto \bar{u}^\alpha$, where $\bar{u}^\alpha$ is the unique minimizer of F_α, is locally Lipschitz continuous and thus in particular continuous.*

Proof. Let $\alpha^1 = (\alpha_1^1, \dots, \alpha_k^1)^T, \alpha^2 = (\alpha_1^2, \dots, \alpha_k^2)^T \in \mathcal{Z}$ be arbitrary and denote by $\bar{u}^1$ and $\bar{u}^2$ the unique minimizers of F_{α^1} and F_{α^2}, respectively. If we plug $u = \bar{u}^2$ into the first-order condition (3.2) for F_{α^1} and $u = \bar{u}^1$ into the first-order condition (3.2) for F_{α^2} and add both inequalities, we end up with

$$\begin{aligned}0 &\leq \langle \nabla F_{\alpha^1}(\bar{u}^1), \bar{u}^2 - \bar{u}^1\rangle_U + \langle \nabla F_{\alpha^2}(\bar{u}^2), \bar{u}^1 - \bar{u}^2\rangle_U \\ &= -\langle \nabla F_{\alpha_1}(\bar{u}^1) - \nabla F_{\alpha_1}(\bar{u}^2), \bar{u}^1 - \bar{u}^2\rangle_U + \langle \nabla F_{\alpha^2}(\bar{u}^2) - \nabla F_{\alpha^1}(\bar{u}^2), \bar{u}^1 - \bar{u}^2\rangle_U. \end{aligned} \tag{3.4}$$

Using the mean value theorem for the first term, we can conclude that there exists $\hat{u}$ with

$$-\langle \nabla F_{\alpha^1}(\bar{u}^1) - \nabla F_{\alpha^1}(\bar{u}^2), \bar{u}^1 - \bar{u}^2\rangle_U = -\langle \nabla^2 F_{\alpha^1}(\hat{u})(\bar{u}^1 - \bar{u}^2), \bar{u}^1 - \bar{u}^2\rangle_U,$$

where using the uniform positive definiteness shown in Lemma 3.21 yields

$$-\langle \nabla^2 F_{\alpha^1}(\hat{u})(\bar{u}^1 - \bar{u}^2), \bar{u}^1 - \bar{u}^2\rangle_U \leq -C_{i_p} \cdot \kappa \left\|\bar{u}^1 - \bar{u}^2\right\|_U^2.$$

Plugging this into (3.4) implies

$$C_{i_p} \cdot \kappa \left\|\bar{u}^1 - \bar{u}^2\right\|_U^2 \leq \langle \nabla F_{\alpha^2}(\bar{u}^2) - \nabla F_{\alpha^1}(\bar{u}^2), \bar{u}^1 - \bar{u}^2\rangle_U. \tag{3.5}$$

For the term on the right-hand side we get by using the Cauchy-Schwarz inequality twice

$$\begin{aligned}\langle \nabla F_{\alpha^2}(\bar{u}^2) - \nabla F_{\alpha^1}(\bar{u}^2), \bar{u}^1 - \bar{u}^2\rangle_U &= \sum_{i=1}^{k} (\alpha_i^2 - \alpha_i^1) \langle \nabla f_i(\bar{u}^2), \bar{u}^1 - \bar{u}^2\rangle_U \\ &\leq \left\|\bar{u}^1 - \bar{u}^2\right\|_U \sum_{i=1}^{k} (\alpha_i^2 - \alpha_i^1) \left\|\nabla f_i(\bar{u}^2)\right\|_U \\ &\leq \left(\sum_{i=1}^{k} \left\|\nabla f_i(\bar{u}^2)\right\|_U^2\right)^{\frac{1}{2}} \left\|\alpha^1 - \alpha^2\right\|_{\mathbb{R}^k} \left\|\bar{u}^1 - \bar{u}^2\right\|_U.\end{aligned}$$

Altogether we obtain

$$\left\|\bar{u}^1 - \bar{u}^2\right\|_U \leq C(\alpha^2) \left\|\alpha^1 - \alpha^2\right\|_{\mathbb{R}^k},$$

where $C(\alpha^2) := (C_{i_p} \cdot \kappa)^{-1} \left(\sum_{i=1}^k \left\|\nabla f_i(\bar{u}^2)\right\|_U^2\right)^{\frac{1}{2}}$ is independent of $\bar{u}^1$ and therefore of α^1, which implies the local Lipschitz continuity of the mapping $\mathcal{Z} \to U_{\text{ad}}$, $\alpha \mapsto \bar{u}^\alpha$. □

Geometrical Properties

The following insight will be needed later, when an algorithm is developed which approximates the Pareto front numerically. It is the application of Remark 3.9 to the weighted sum method.

Lemma 3.23. *To $\alpha \in \mathbb{R}^k_{\gneqq}$ let $\bar{u} \in U_{ad}$ be the solution of (WSP)$_\alpha$ and $\bar{y} := f(\bar{u})$. Then the hyperplane $H_\alpha := \{\bar{y} + x \mid \langle x, \alpha\rangle_{\mathbb{R}^k} = 0\}$ is always below the objective feasible region Y (in the sense of Definition 2.2) with only boundary point $\bar{y}$ and hence also below the Pareto front P_Y.*

Proof. We need to show that

$$\forall z \in H_\alpha \, \forall y \in Y : \neg(y \lneqq z) \tag{3.6}$$
$$\forall y \in Y \, \exists z \in H_\alpha : z \leq y \tag{3.7}$$

hold. Equation (3.6) is fulfilled, as $\bar{u}$ is the solution of (WSP)$_\alpha$ and $H_\alpha = g_\alpha^{-1}(\{g_\alpha(\bar{y})\})$, where $g_\alpha : \mathbb{R}^k \to \mathbb{R}$, $g(x) := \sum_{i=1}^k \alpha_i x_i$. This also shows that $\bar{y}$ is the only boundary point.
To show (3.7), let $y \in Y$ be arbitrary and define $\tilde{y} := y - \bar{y}$, i.e. $y = \bar{y} + \tilde{y}$. Furthermore, let $\{\alpha, \varphi_2, \ldots, \varphi_k\}$ be an orthonormal basis of $\mathbb{R}^k$. Then it holds $\tilde{y} = \langle \tilde{y}, \alpha\rangle_{\mathbb{R}^k}\alpha + \sum_{i=2}^k \langle \tilde{y}, \varphi_i\rangle_{\mathbb{R}^k}\varphi_i$. Additionally, we get

$$\langle \tilde{y}, \alpha\rangle_{\mathbb{R}^k} = \langle y - \bar{y}, \alpha\rangle_{\mathbb{R}^k} = \langle y, \alpha\rangle_{\mathbb{R}^k} - \langle \bar{y}, \alpha\rangle_{\mathbb{R}^k} \geq 0,$$

as $\bar{u}$ is the minimizer of F_α, and with $x := \sum_{i=2}^k \langle \tilde{y}, \varphi_i\rangle_{\mathbb{R}^k}\varphi_i$ by definition of an orthonormal basis $\langle x, \alpha\rangle_{\mathbb{R}^k} = 0$.
So in total $y = \bar{y} + x + \langle \tilde{y}, \alpha\rangle_{\mathbb{R}^k}\alpha$ with $\bar{y} + x \in H_\alpha$ and $\langle \tilde{y}, \alpha\rangle_{\mathbb{R}^k}\alpha \geq 0$. Thus, the claim follows. □

Remark 3.24. As mentioned in Remark 3.9, the statement of Lemma 3.23 is the best geometrical information we can hope for. In this sense the weighted sum method yields optimal geometrical information.

By combining the ideas of Corollary 3.12 and Lemma 3.23, we obtain an analogue result for more general scalarization methods.

Lemma 3.25. *Let Assumptions 1 and 2 be satisfied and let the function $g : \mathbb{R}^k_{\gneqq y^{id}} \to \mathbb{R}$ fulfil Presumption 1 and be additionally differentiable. Denote by $\bar{u}$ the unique minimizer of F^g and define $\bar{y} := f(\bar{u})$. If $\nabla g(\bar{y}) \neq 0$, then the hyperplane $H_g := \{\bar{y} + x \mid \langle x, \nabla g(\bar{y})\rangle_{\mathbb{R}^k} = 0\}$ is below the objective feasible region and hence below the Pareto front with only boundary point $\bar{y}$.*

Proof. First of all, by the strict monotonicity and the differentiability of g, we can conclude that it holds $\nabla g(x) \geq 0$ for all $x \in \mathbb{R}^k_{\geq y^{\text{id}}}$. This yields $\nabla g(\bar{y}) \gneqq 0$. By defining $\alpha := \nabla g(\bar{y})$, we can use Corollary 3.12 to conclude that $\bar{u}$ also is the unique solution of (WSP)$_\alpha$. Now, the claim follows from Lemma 3.23. □

3.3.2 Reference Point Methods

Another approach to generate a family of functions $(g_i)_{i\in I}$ to compute Pareto optimal points is given by the so-called reference point methods. For a given reference point $z \in \mathbb{R}^k$ we want to find the point on the Pareto front that has the smallest 'distance' to the point z. Hereby, we consider the expression 'distance' to be a bit more general. In Example 3.10 (ii) we used the distance to the ideal objective vector defined by the p-norms for $p \in [1, \infty]$, but we want to extend this to more general cost functions g, which simply fulfil Presumption 1.
Formally, we define the reference point problem for a function g and a reference point z by

$$\min_{u\in U_{\text{ad}}} g(f(u) - z). \tag{RPP}$$

Analytical Results

Presumption 2. *A function $g : \mathbb{R}^k_{\geq} \to \mathbb{R}$ is said to fulfil Presumption 2, if g fulfils Presumption 1 with $y^{id} := 0$.*

Lemma 3.26. *Let Assumption 1 be satisfied and assume that the function $g : \mathbb{R}^k_{\geq} \to \mathbb{R}$ fulfils Presumption 2. Then the function $g_z : \mathbb{R}^k_{\geq y^{id}} \to \mathbb{R}$, $g_z(x) := g(x - z)$ for a $z \in \mathbb{R}^k_{\leq y^{id}}$ fulfils Presumption 1. Hence, the function F^{g_z} has a unique minimizer that is Pareto optimal.*

Proof. As $z \in \mathbb{R}^k_{\leq y^{\text{id}}}$ holds, this follows immediately. □

Remark 3.27. Again with the notation of the beginning of this section we now consider the family of functions $(g_z)_{z\in\mathbb{R}^k_{\leq y^{\text{id}}}}$ for a given function g.

Notation 3.28. Let us assume that a function g fulfils Presumption 2. In order to not have to define new functions all the time when the reference point z is changing, we simply define $F^g_z(u) := g(f(u) - z)$ for $z \in \mathbb{R}^k_{\leq y^{\text{id}}}$.

Corollary 3.29. *Assume that Assumptions 1 and 2 hold. Additionally, let the function g fulfil Presumption 2 and be differentiable. Furthermore, let $z \in \mathbb{R}^k_{\leq y^{id}}$ be arbitrary. Then the first-order condition stated in Theorem 3.11 for the unique solution $\bar{u}$ of the reference point problem reads*

$$\langle \sum_{i=1}^{k} \partial_i g(f(\bar{u}) - z)\nabla f_i(\bar{u}), u - \bar{u}\rangle_U \geq 0 \text{ for all } u \in U_{ad}. \tag{3.8}$$

As for the weighted sum method in the previous section, we want to know which points on the Pareto front can be computed by using a reference point method. Of course this will depend on the function g.
The next lemma ensures the existence of a non-differentiable function g which can be used to compute all points of the Pareto front via the reference point method.

Lemma 3.30. *Let Assumption 1 be satisfied and let $\bar{y} \in P_Y$ and $\bar{u} \in U_{ad}$ with $\bar{y} = f(\bar{u})$. Define $r := \left\|\bar{y} - y^{id}\right\|_\infty$ and $z \in \mathbb{R}^k_{\leq y^{id}}$ by $z_i := \bar{y}_i - r$ for all $i \in \{1, \dots, k\}$ as well as $g : \mathbb{R}^k_{\geq} \to \mathbb{R}$, $g(x) := \|x\|_\infty$. Then $\bar{u}$ is the unique minimizer of F^g_z.*

Proof. One easily checks that the function g fulfils Presumption 2. Therefore, there is a unique minimizer $\hat{u}$ of the function F_z^g. Assume that $\hat{u} \neq \bar{u}$. Then we obtain

$$\max\{f_1(\hat{u}) - z_1, \dots, f_k(\hat{u}) - z_k\} = F_z^g(\hat{u}) < F_z^g(\bar{u}) = \max\{|f_1(\bar{u}) - z_1|, \dots, |f_k(\bar{u}) - z_k|\} = r.$$

So for each $i \in \{1, \dots, k\}$ we get that $f_i(\hat{u}) - z_i < r$ holds. But this yields $f_i(\hat{u}) < r + z_i = f_i(\bar{u})$ for all $i \in \{1, \dots, k\}$, which is a contradiction to the Pareto optimality of $f(\bar{u})$. Hence, $\bar{u}$ is the unique minimizer of the function F_z^g. □

Remark 3.31. (i) Lemma 3.30 shows that we cannot only compute each point on the Pareto front by solving a reference point problem, but knowing the coordinates of any point $\bar{y}$ on the Pareto front, we can also specify the function g and the reference point z such that the minimizer $\bar{u}$ of F_z^g fulfils $f(\bar{u}) = \bar{y}$. This can also be used to test if an objective vector $y \in Y$ is Pareto optimal or not.

(ii) Having the numerical implementation in mind, the drawback of using this function g is that it is not differentiable. Consequently, we cannot use the standard numerical methods to solve the scalarized optimization problems, as they all need the differentiability of the function that is to be minimized.

If we want to use differentiable functions and still be able to compute (almost) all Pareto optimal points, we have to make some additional assumptions on the function g and its derivatives.

Theorem 3.32. *Assume that Assumptions 1 and 2 are satisfied. Furthermore, the function g is supposed to fulfil Presumption 2 and to be continuously differentiable such that $\partial_i g$ is monotonically increasing for all $i \in \{1, \dots, k\}$, as well as $\lim_{\|x\|_{\mathbb{R}^k} \to \infty} \partial_i g(x) = \infty$ for all $i \in \{1, \dots, k\}$. Additionally, assume that there is $n \in \mathbb{R}^k_>$ such that for all $y \in \mathbb{R}^k_{\geq n}$ the equation system*

$$\nabla g(x) = y \tag{3.9}$$

has a unique solution.
Then it holds

$$\{u \in U_{ad} \mid \exists\, z \in \mathbb{R}^k_{\leq y^{id}} : u \text{ minimizes } F_z^g\} \supset \{u \in U_{ad} \mid \exists\, \alpha \in \mathbb{R}^k_> : u \text{ solves } (WSP)_\alpha\}.$$

Proof. Let $\bar{u} \in U_{\mathrm{ad}}$ be in the second set, i.e. there is $\alpha \in \mathbb{R}^k_>$ such that $\bar{u}$ is the unique solution of $(\mathrm{WSP})_\alpha$. From the presumptions we can conclude that there are $x \in \mathbb{R}^k_{\geq f(\bar{u}) - y^{\mathrm{id}}}$ and $\lambda > 0$ with $\nabla g(x) = \lambda\alpha$. Define $z := f(\bar{u}) - x$. Then it holds $z \leq y^{\mathrm{id}}$ and $\nabla g(f(\bar{u}) - z) = \lambda\alpha$. Now, Corollary 3.12 tells us that $\bar{u}$ is the unique minimizer of F_z^g as well. Hence, $\bar{u}$ is in the first set. □

Remark 3.33. By Theorem 3.20 all points on the Pareto front can be computed by solving a weighted sum problem. Now we know by Theorem 3.32 that the only points we possibly cannot compute using the reference point method are the ones that are the solution to a weighted sum problem with at least one of the weights $\alpha_i = 0$, i.e. the 'boundary' of the Pareto front.

3.3.3 Euclidean Reference Point Method

Probably the most intuitive cost function to use for the reference point method is the squared Euclidean distance, i.e. $g(x) := \|x\|_{\mathbb{R}^k}^2 = \sum_{i=1}^k x_i^2$. Given a reference point $z \in \mathbb{R}^k$, minimizing F_z^g is equivalent to looking for the objective vector y that is closest to z in terms of the natural Euclidean distance. Additionally to the intuitive nature of this approach, it can be shown that it also behaves nice in an analytical point of view. Formally, we define the Euclidean reference point problem for a reference point $z \in \mathbb{R}^k$ by

$$\min_{u \in U_{\text{ad}}} \sum_{i=1}^k \frac{1}{2} \left(f_i(u) - z_i\right)^2, \tag{ERPP}$$

where the factor $\frac{1}{2}$ is only introduced for not having to deal with any factors when considering the derivative.

Notation 3.34. For $z \in \mathbb{R}^k$ we denote the Euclidean reference point problem with reference point z by $(\text{ERPP})_z$. As this will be our standard reference point method in the following, we simply write $F_z(u) := \sum_{i=1}^k \frac{1}{2} \left(f_i(u) - z_i\right)^2$.

Analytical Results

Theorem 3.35. *Let Assumption 1 be satisfied and $z \in P_Y + \mathbb{R}^k_{\leq}$. Then the Euclidean reference point problem $(ERPP)_z$ has a unique, Pareto optimal solution.*

Proof. We want to apply Theorem 2.13 to the set $Y + \mathbb{R}^k_{\geq}$. From Lemma 3.19 we already know that $Y + \mathbb{R}^k_{\geq}$ is convex, wherefore it remains to show that $Y + \mathbb{R}^k_{\geq}$ is closed.
Let $(y_n)_{n\in\mathbb{N}} \subset Y + \mathbb{R}^k_{\geq}$ be arbitrary with $y_n \to y$ $(n \to \infty)$ for a $y \in \mathbb{R}^k$. Let $(u_n)_{n\in\mathbb{N}} \subset U_{\text{ad}}$ and $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}^k_{\geq}$ be two sequences such that $y_n = f(u_n) + x_n$ holds for all $n \in \mathbb{N}$. By Assumption 1 we know that $(f(u_n))_{n\in\mathbb{N}}$ is bounded from below and so is $(x_n)_{n\in\mathbb{N}}$ by definition. As the sum of the two sequences is convergent, they are also bounded from above. By the Bolzano–Weierstrass theorem both sequences have a convergent subsequence, which we again denote by $(f(u_n))_{n\in\mathbb{N}}$ and $(x_n)_{n\in\mathbb{N}}$, with limits $\tilde{x} \in \mathbb{R}^k$ and $\bar{x} \in \mathbb{R}^k_{\geq}$. We can further conclude that $y = \tilde{x} + \bar{x}$ holds.
In the case that U_{ad} is bounded, we immediately get that $(u_n)_{n\in\mathbb{N}}$ is bounded. If U_{ad} is unbounded, we use that $(f(u_n))_{n\in\mathbb{N}}$ is bounded from above and get from Assumption 1 that $(u_n)_{n\in\mathbb{N}}$ is bounded as well. As U is reflexive, according to Lemma 2.20 there exist a subsequence $(u_{n_j})_{j\in\mathbb{N}}$ and $u \in U$ with $u_{n_j} \rightharpoonup u$. We can conclude $u \in U_{\text{ad}}$ since U_{ad} is weakly closed. The weak lower semi-continuity of the functions f_i yields

$$f_i(u) \leq \liminf_{j\to\infty} f_i(u_{n_j}) = \tilde{x}_i.$$

Altogether we get

$$f(u) + \bar{x} \leq \tilde{x} + \bar{x} = y,$$

and therefore $y \in Y + \mathbb{R}^k_{\geq}$, i.e. $Y + \mathbb{R}^k_{\geq}$ is closed.
The approximation theorem tells us now that there is a unique $\bar{y} \in Y + \mathbb{R}^k_{\geq}$ with $\|z - \bar{y}\|_{\mathbb{R}^k} = \inf\{\|y - z\|_{\mathbb{R}^k} \mid y \in Y + \mathbb{R}^k_{\geq}\}$.
Due to the presumption $z \in P_Y + \mathbb{R}^k_{\leq}$, we can also conclude $\bar{y} \in P_Y$. Now it follows directly that $\bar{u} \in U_{\text{ad}}$ with $f(\bar{u}) = \bar{y}$ is the unique solution of $(\text{ERPP})_z$. □

Remark 3.36. (i) Note that the function $g : \mathbb{R}^k_{\leq} \to \mathbb{R}$, $g(x) := \frac{1}{2}\sum_{i=1}^k x_i^2$ fulfils Presumption 2. So for $z \leq y^{\mathrm{id}}$, we could have argued with Lemma 3.26 to show that there is a unique, Pareto optimal solution of $(\mathrm{ERPP})_z$. If $z \in P_Y + \mathbb{R}^k_{\leq}$ but $z \not\leq y^{\mathrm{id}}$, the function F_z is not necessarily convex. Consequently, Theorem 3.7 cannot be applied to show the unique solvability of $(\mathrm{ERPP})_z$ in this case.

(ii) The proof of Theorem 3.35 suggests another way to look at the Euclidean reference point problem. We can see it as a projection of the reference point onto the set $Y + \mathbb{R}^k_{\leq}$ and we will be able to draw further insights from this point of view.

(iii) The relaxed presumption on the reference point in Theorem 3.35 will be beneficial, amongst others, in the construction of an algorithm to approximate the whole Pareto front using the Euclidean reference point method. During the computation it is displeasing to always be forced to assure that the reference point fulfils $z \leq y^{\mathrm{id}}$. Theorem 3.35 yields that we do not need this restriction on z as long as it lies beneath the Pareto front.

Corollary 3.37. *Assume that Assumptions 1 and 2 hold and let $z \in \mathbb{R}^k_{\leq y^{id}}$ be arbitrary. Then the necessary and sufficient first-order condition stated in Theorem 3.11 for the unique solution $\bar{u}$ of $(ERPP)_z$ reads*

$$\langle \sum_{i=1}^k (f_i(\bar{u}) - z)\nabla f_i(\bar{u}), u - \bar{u}\rangle_U \geq 0 \text{ for all } u \in U_{ad}. \tag{3.10}$$

If $z \in P_Y + \mathbb{R}^k_{\leq}$ but $z \not\leq y^{id}$, the above condition is at least necessary.

Proof. The first statement follows directly from Theorem 3.11. For the second statement we use that Lemma 2.11 does not need the convexity of the function F_z to prove that (3.10) is a necessary first-order condition. □

Lemma 3.38. *Assume that Assumptions 1, 2 and 3 are satisfied and let $z \in P_Y + \mathbb{R}^k_{\leq}$ be arbitrary. Then it holds*

$$\langle \nabla^2 F_z(u)v, v\rangle_U \geq \left(f_{i_p}(u) - z_{i_p}\right) C_{i_p} \|v\|_U^2$$

for all $u \in U_{ad}$ with $f(u) \geq z$ and all $v \in U$.
Thus, if $\bar{u} \in U_{ad}$ fulfils (3.10), $f(u) \geq z$ and $f_{i_p}(\bar{u}) > z_{i_p}$, $\bar{u}$ is a local strict minimizer of F_z.

Proof. Let $u \in U_{\mathrm{ad}}$ with $f(u) \geq z$ be arbitrary. By applying the chain rule, we get that

$$\nabla^2 F_z(u)v = \sum_{i=1}^k \left(\langle \nabla f_i(u), v\rangle_U \nabla f_i(u) + (f_i(u) - z_i)\,\nabla^2 f_i(u)v\right) \text{ for all } v \in U,$$

and therefore

$$\begin{aligned}
\langle \nabla^2 F_z(u)v, v\rangle_U &= \langle \sum_{i=1}^k \left(\langle \nabla f_i(u), v\rangle_U \nabla f_i(u) + (f_i(u) - z_i)\,\nabla^2 f_i(u)v\right), v\rangle_U \\
&= \sum_{i=1}^k \langle \nabla f_i(u), v\rangle_U^2 + \sum_{i=1}^k (f_i(u) - z_i)\,\langle \nabla^2 f_i(u)v, v\rangle_U \\
&\geq \sum_{i=1}^k (f_i(u) - z_i)\,\langle \nabla^2 f_i(u)v, v\rangle_U \text{ for all } v \in U.
\end{aligned}$$

Using Assumption 3 yields $\langle \nabla^2 F_z(u)v, v\rangle_U \geq C_{i_p} \cdot \left(f_{i_p}(u) - z_{i_p}\right) \|v\|_U^2$.
Now the second claim follows since F_z is strictly convex in a neighbourhood of $\bar{u}$. □

Remark 3.39. It is possible to show that (3.10) is a sufficient condition for the unique solution $\bar{u}$ of $(\text{ERPP})_z$, if $z \in P_Y + \mathbb{R}^k_{\leq}$ with $z \lneqq f(\bar{u})$ holds. We will prove this later in Theorem 3.45.

Theorem 3.40. *Let Assumption 1 be satisfied and let $z \in P_Y + \mathbb{R}^k_{\leq}$ be arbitrary. Then a necessary and sufficient first-order condition for the image $f(\bar{u})$ of the unique solution $\bar{u}$ of $(ERPP)_z$ reads*

$$\langle f(\bar{u}) - z, y - f(\bar{u})\rangle_{\mathbb{R}^k} \geq 0 \text{ for all } y \in Y + \mathbb{R}^k_{\geq}. \tag{3.11}$$

Proof. As we can see the Euclidean reference point problem as a projection of the reference point z onto the set $Y + \mathbb{R}^k_{\geq}$, this follows from Theorem 2.14. □

Corollary 3.41. *Let Assumption 1 be satisfied and $z \in P_Y + \mathbb{R}^k_{\leq}$ be arbitrary. Moreover, let $\bar{u}$ be the unique minimizer of F_z. Then it holds $z \leq f(\bar{u})$.*

Proof. We know that $f(\bar{u})$ fulfils the first-order condition (3.11) and setting $y = f(\bar{u}) + x$ for an arbitrary $x \in \mathbb{R}^k_{\geq}$ in (3.11) yields

$$\langle f(\bar{u}) - z, x\rangle_{\mathbb{R}^k} \geq 0.$$

As this holds for all $x \in \mathbb{R}^k_{\geq}$, we get $f(\bar{u}) - z \geq 0$. □

Lemma 3.42. *Let Assumption 1 be satisfied and let $z \in P_Y + \mathbb{R}^k_{\leq}$ be arbitrary. Additionally, let $\bar{u}$ be the unique minimizer of F_z. Then $\bar{u}$ is also the unique minimizer of $F_{\tilde{z}}$ for each $\tilde{z} = f(\bar{u}) + \lambda(z - f(\bar{u}))$ with a $\lambda \geq 0$, i.e. for all reference points lying on the ray that is starting in $f(\bar{u})$ and going through z.*

Proof. By Theorem 3.40 the first-order condition (3.11) is fulfilled. Let $\lambda \geq 0$ be arbitrary and $\tilde{z} := f(\bar{u}) + \lambda(z - f(\bar{u}))$. Then it holds $f(\bar{u}) - \tilde{z} = \lambda(f(\bar{u}) - z)$ and therefore

$$\langle f(\bar{u}) - \tilde{z}, \tilde{y} - f(\bar{u})\rangle_{\mathbb{R}^k} \geq 0 \text{ for all } \tilde{y} \in Y + \mathbb{R}^k_{\geq},$$

which is a sufficient first-order condition for the solution of $(\text{ERPP})_z$. □

In Theorem 3.20 we showed that all Pareto optimal points can be obtained by solving a weighted sum problem, if Assumption 1 is satisfied. This result can be used to show that the same holds for the Euclidean reference point method.

Theorem 3.43. *Let Assumption 1 be satisfied. If $\bar{y} = f(\bar{u}) \in P_Y$, then there is $z \in P_Y + \mathbb{R}^k_{\lneqq}$ such that $\bar{u}$ is the unique solution of $(ERPP)_z$.*

Proof. Let $\bar{y} = f(\bar{u}) \in P_Y$. By Theorem 3.20 we know that there is $\alpha \in \mathbb{R}^k_{\gneqq}$ such that $\bar{u}$ is the unique solution of $(\text{WSP})_\alpha$. Define $z := f(\bar{u}) - \alpha$. As $-\alpha \in \mathbb{R}^k_{\lneqq}$, it holds $z \in P_Y + \mathbb{R}^k_{\lneqq}$. We want to show that $\bar{u}$ is the solution of $(\text{ERPP})_z$ by showing that the first-order condition (3.11) holds. Therefore, we make use of Lemma 3.23 again. As $H_\alpha = \{\bar{y} + x \mid \langle x, \alpha\rangle_{\mathbb{R}^k} = 0\}$ is below

the objective feasible region, we can split an arbitrary $\tilde{y} \in P_Y + \mathbb{R}^k_{\leq}$ into $\tilde{y} = \bar{y} + x + a$ with $\langle x, \alpha \rangle_{\mathbb{R}^k} = 0$ and $a \in \mathbb{R}^k_{\leq}$. But then it holds

$$\langle f(\bar{u}) - z, \tilde{y} - f(\bar{u}) \rangle_{\mathbb{R}^k} = \langle \alpha, x + a \rangle_{\mathbb{R}^k} = \langle \alpha, a \rangle_{\mathbb{R}^k} \geq 0,$$

so that the first-order condition (3.11) holds. Hence, $\bar{u}$ is the solution of $(\text{ERPP})_z$. □

Remark 3.44. (i) If we restrict ourselves to the case $z \leq y^{\text{id}}$ and assume Assumption 2, we can also use Theorem 3.32 to prove that we can obtain all Pareto optimal points $\bar{u}$, for which there is $\alpha > 0$ such that $\bar{u}$ is the unique solution of $(\text{WSP})_\alpha$, by solving $(\text{ERPP})_z$.

(ii) Although the last theorem shows that we can obtain all Pareto optimal points by solving a Euclidean reference point problem, we do not know a priori how the set $P_Y + \mathbb{R}^k_{\leq}$ looks like and thus how to choose appropriate reference points. So in practice it is not clear yet how to construct the reference points. This problem will be addressed later when we derive an algorithm to approximate the Pareto front in Section 3.4.1.

Using an idea from the proof of Theorem 3.43, we can show the statement of Remark 3.39.

Theorem 3.45. *Assume that Assumptions 1 and 2 hold and let $z \in P_Y + \mathbb{R}^k_{\leq}$ be arbitrary. If $\bar{u} \in U_{ad}$ satisfies $f(\bar{u}) \gneqq z$ and (3.10), then $\bar{u}$ is the unique solution of $(\text{ERPP})_z$.*

Proof. Let $u \in U_{\text{ad}}$ be arbitrary such that $f(\bar{u}) \gneqq z$ and (3.10) are satisfied. Define $\alpha := f(\bar{u}) - z \in \mathbb{R}^k_{\gneqq}$. By the same argumentation as in Corollary 3.12, $\bar{u}$ is the unique minimizer of F_α, i.e. the weighted sum problem with weights α. Now we can argue in the same way as in Theorem 3.43 to obtain that $\bar{u}$ is the unique minimizer of $(\text{ERPP})_z$. □

In the following we want to show that the unique solution of the Euclidean reference point problem and its image depend continuously on the reference point. A similar result was already derived in [2].

Lemma 3.46. *Let Assumptions 1 and 3 be satisfied and define $\mathcal{Z} := \{z \in P_Y + \mathbb{R}^k_{\leq} \mid z_{i_p} < y^{id}_{i_p} - \kappa\}$ for an arbitrary $\kappa > 0$. Then the mapping $\mathcal{Z} \to U_{ad}$, $z \mapsto \bar{u}^z$, where $\bar{u}^z$ is the unique minimizer of F_z, is Lipschitz continuous with Lipschitz constant $\frac{1}{2}\left(\kappa \cdot C_{i_p}\right)^{-\frac{1}{2}}$.*

Proof. Let $z^1 = (z^1_1, \ldots, z^1_k)^T$, $z^2 = (z^2_1, \ldots, z^2_k)^T \in \mathcal{Z}$ be arbitrary and $\bar{u}^1, \bar{u}^2$ be the unique minimizers of F_{z^1} and F_{z^2}, respectively. By the first-order condition (3.10) it holds

$$\langle \nabla F_{z^1}(\bar{u}^1), u - \bar{u}^1 \rangle_U \geq 0 \text{ for all } u \in U_{\text{ad}}, \tag{3.12}$$

$$\langle \nabla F_{z^2}(\bar{u}^2), u - \bar{u}^2 \rangle_U \geq 0 \text{ for all } u \in U_{\text{ad}}. \tag{3.13}$$

Setting $u = \bar{u}^2$ in (3.12) and $u = \bar{u}^1$ in (3.13) and adding up both inequalities implies

$$\begin{aligned} 0 \leq & \langle \nabla F_{z^1}(\bar{u}^1), \bar{u}^2 - \bar{u}^1 \rangle_U + \langle \nabla F_{z^2}(\bar{u}^2), \bar{u}^1 - \bar{u}^2 \rangle_U \\ = & \sum_{i=1}^{k} \left[\left(f_i(\bar{u}^1) - z^1_i\right) \langle \nabla f_i(\bar{u}^1), \bar{u}^2 - \bar{u}^1 \rangle_U + \left(f_i(\bar{u}^2) - z^2_i\right) \langle \nabla f_i(\bar{u}^2), \bar{u}^1 - \bar{u}^2 \rangle_U \right]. \end{aligned} \tag{3.14}$$

Looking at the i-th summand and assuming first that

$$f_i(\bar{u}^1) - z_i^1 \leq f_i(\bar{u}^2) - z_i^2$$

holds, we get

$$\begin{aligned}
&\left(f_i(\bar{u}^1) - z_i^1\right) \langle \nabla f_i(\bar{u}^1), \bar{u}^2 - \bar{u}^1 \rangle_U + \left(f_i(\bar{u}^2) - z_i^2\right) \langle \nabla f_i(\bar{u}^2), \bar{u}^1 - \bar{u}^2 \rangle_U \\
=& \left(f_i(\bar{u}^1) - z_i^1\right) \langle \nabla f_i(\bar{u}^1) - \nabla f_i(\bar{u}^2), \bar{u}^2 - \bar{u}^1 \rangle_U \\
&+ \left(f_i(\bar{u}^2) - f_i(\bar{u}^1) + z_i^1 - z_i^2\right) \langle \nabla f_i(\bar{u}^2), \bar{u}^1 - \bar{u}^2 \rangle_U. \qquad (3.15)
\end{aligned}$$

For the first summand we get by the mean value theorem that there exists a $\tilde{u}^i \in U_{\text{ad}}$ with

$$\langle \nabla f_i(\bar{u}^1) - \nabla f_i(\bar{u}^2), \bar{u}^2 - \bar{u}^1 \rangle_U = \langle \nabla^2 f_i(\tilde{u}^i)(\bar{u}^1 - \bar{u}^2), \bar{u}^2 - \bar{u}^1 \rangle_U. \qquad (3.16)$$

As f_i is convex and differentiable by assumption, it holds

$$\langle \nabla f_i(\bar{u}^2), \bar{u}^1 - \bar{u}^2 \rangle_U \leq f_i(\bar{u}^1) - f_i(\bar{u}^2), \qquad (3.17)$$

and as we assumed $f_i(\bar{u}^2) - f_i(\bar{u}^1) + z_i^1 - z_i^2 \geq 0$, we can conclude from (3.15), (3.16) and (3.17)

$$\begin{aligned}
&\left(f_i(\bar{u}^1) - z_i^1\right) \langle \nabla f_i(\bar{u}^1), \bar{u}^2 - \bar{u}^1 \rangle_U + \left(f_i(\bar{u}^2) - z_i^2\right) \langle \nabla f_i(\bar{u}^2), \bar{u}^1 - \bar{u}^2 \rangle_U \\
\leq& - \left(f_i(\bar{u}^1) - z_i^1\right) \langle \nabla^2 f_i(\tilde{u}^i)(\bar{u}^2 - \bar{u}^1), \bar{u}^2 - \bar{u}^1 \rangle_U + \left(f_i(\bar{u}^2) - f_i(\bar{u}^1) + z_i^1 - z_i^2\right) \left(f_i(\bar{u}^1) - f_i(\bar{u}^2)\right). \qquad (3.18)
\end{aligned}$$

In the case that

$$f_i(\bar{u}^1) - z_i^1 > f_i(\bar{u}^2) - z_i^2$$

holds, a similar computation yields

$$\begin{aligned}
&\left(f_i(\bar{u}^1) - z_i^1\right) \langle \nabla f_i(\bar{u}^1), \bar{u}^2 - \bar{u}^1 \rangle_U + \left(f_i(\bar{u}^2) - z_i^2\right) \langle \nabla f_i(\bar{u}^2), \bar{u}^1 - \bar{u}^2 \rangle_U \\
\leq& - \left(f_i(\bar{u}^2) - z_i^2\right) \langle \nabla^2 f_i(\tilde{\tilde{u}}^i)(\bar{u}^2 - \bar{u}^1), \bar{u}^2 - \bar{u}^1 \rangle_U + \left(f_i(\bar{u}^2) - f_i(\bar{u}^1) + z_i^1 - z_i^2\right) \left(f_i(\bar{u}^1) - f_i(\bar{u}^2)\right) \qquad (3.19)
\end{aligned}$$

for a $\tilde{\tilde{u}}^i \in U_{\text{ad}}$, so that (3.18) and (3.19) combined imply

$$\begin{aligned}
&\left(f_i(\bar{u}^1) - z_i^1\right) \langle \nabla f_i(\bar{u}^1), \bar{u}^2 - \bar{u}^1 \rangle_U + \left(f_i(\bar{u}^2) - z_i^2\right) \langle \nabla f_i(\bar{u}^2), \bar{u}^1 - \bar{u}^2 \rangle_U \\
\leq& - \min\big[\left(f_i(\bar{u}^1) - z_i^1\right) \langle \nabla^2 f_i(\tilde{u}^i)(\bar{u}^2 - \bar{u}^1), \bar{u}^2 - \bar{u}^1 \rangle_U, \\
&\qquad \left(f_i(\bar{u}^2) - z_i^2\right) \langle \nabla^2 f_i(\tilde{\tilde{u}}^i)(\bar{u}^2 - \bar{u}^1), \bar{u}^2 - \bar{u}^1 \rangle_U \big] \\
&+ \left(f_i(\bar{u}^2) - f_i(\bar{u}^1) + z_i^1 - z_i^2\right) \left(f_i(\bar{u}^1) - f_i(\bar{u}^2)\right). \qquad (3.20)
\end{aligned}$$

As (3.20) holds for all $i \in \{1, \ldots, k\}$, we get from (3.14) by summing up (3.20) for all $i \in \{1, \ldots, k\}$ and using Assumption 3 as well as the Cauchy-Schwarz inequality and Young's inequality

$$\begin{aligned}
0 \leq& - \left(\min(f_{i_p}(\bar{u}^1) - z_{i_p}^1, f_{i_p}(\bar{u}^2) - z_{i_p}^2)\right) C_{i_p} \left\|\bar{u}^2 - \bar{u}^1\right\|_U^2 - \left\|f(\bar{u}^2) - f(\bar{u}^1)\right\|_{\mathbb{R}^k}^2 \\
&+ \left\|f(\bar{u}^2) - f(\bar{u}^1)\right\|_{\mathbb{R}^k} \left\|z^2 - z^1\right\|_{\mathbb{R}^k} \\
\leq& - \left(\min(f_{i_p}(\bar{u}^1) - z_{i_p}^1, f_{i_p}(\bar{u}^2) - z_{i_p}^2)\right) C_{i_p} \left\|\bar{u}^2 - \bar{u}^1\right\|_U^2 + \frac{1}{4} \left\|z^2 - z^1\right\|_{\mathbb{R}^k}^2 \\
\leq& - \kappa \cdot C_{i_p} \left\|\bar{u}^2 - \bar{u}^1\right\|_U^2 + \frac{1}{4} \left\|z^2 - z^1\right\|_{\mathbb{R}^k}^2,
\end{aligned}$$

from which the claim follows by rearranging the terms. □

The next result shows the continuity in the objective space.

Theorem 3.47. *Let Assumption 1 be satisfied. Then the mapping $P_Y + \mathbb{R}^k_{\leq} \to P_Y$, $z \mapsto f(\bar{u}^z)$, where $\bar{u}^z$ is the unique minimizer of F_z, is Lipschitz continuous with Lipschitz constant 1.*

Proof. This follows immediately from Theorem 2.14, as $z \mapsto f(\bar{u}^z)$ is the projection of z onto $Y + \mathbb{R}^k_{\geq}$. □

With a similar proof as the one of Lemma 3.46 we can show a result about the distance between an arbitrary feasible vector $u_p \in U_{\text{ad}}$ and the solution $\bar{u}$ of $(\text{ERPP})_z$. A corresponding result for weighted sum problems can be found in [37, Theorem 3.1], from which the perturbation idea, which was already used before for instance in [23], of the proof is taken.

Lemma 3.48. *Let Assumptions 1 and 3 be satisfied. Define $\mathcal{Z} := \{z \in \mathbb{R}^k_{\leq y^{id}} \mid z_{i_p} \leq y^{id}_{i_p} - \kappa\}$ for an arbitrary $\kappa > 0$ and let $z \in \mathcal{Z}$ be arbitrary. Denote by $\bar{u}$ the solution of (ERPP)$_z$. Then for all $u_p \in U_{ad}$ we have*

$$\begin{aligned} \|\bar{u} - u_p\|_U &\leq \left(C_{i_p}\kappa\right)^{-1} \|\xi\|_U \quad \text{and} \\ \|f(\bar{u}) - f(u_p)\|_{\mathbb{R}^k} &\leq \frac{1}{2}\left(C_{i_p}\kappa\right)^{-\frac{1}{2}} \|\xi\|_U , \end{aligned}$$

where $\xi \in U$ is chosen such that

$$\langle \nabla F_z(u_p) + \xi, u - u_p \rangle_U \geq 0 \text{ for all } u \in U_{ad} \tag{3.21}$$

is fulfilled.

Proof. Let ξ be chosen such that (3.21) is fulfilled. As $\bar{u}$ is the unique solution of $(\text{ERPP})_z$, it fulfils the first-order condition

$$\langle \nabla F_z(\bar{u}), u - \bar{u} \rangle_U \geq 0 \text{ for all } u \in U_{\text{ad}}. \tag{3.22}$$

Choosing $u = \bar{u}$ in (3.21) and $u = u_p$ in (3.22) and adding both inequalities yields

$$\begin{aligned} 0 &\leq \langle \nabla F_z(\bar{u}), u_p - \bar{u} \rangle_U + \langle \nabla F_z(u_p), \bar{u} - u_p \rangle_U + \langle \xi, \bar{u} - u_p \rangle_U \\ &\leq \langle \nabla F_z(\bar{u}), u_p - \bar{u} \rangle_U + \langle \nabla F_z(u_p), \bar{u} - u_p \rangle_U + \|\xi\|_U \|\bar{u} - u_p\|_U . \end{aligned}$$

For the first two summands, we can argue in exactly the same way as in Lemma 3.46 with $z^1 = z^2 = z$, so that we get altogether

$$\begin{aligned} 0 &\leq -\left(\min(f_{i_p}(\bar{u}) - z_{i_p}, f_{i_p}(u_p) - z_{i_p})\right) C_{i_p} \|\bar{u} - u_p\|^2_U - \|f(\bar{u}) - f(u_p)\|^2_{\mathbb{R}^k} + \|\xi\|_U \|\bar{u} - u_p\|_U \\ &\leq -\kappa \cdot C_{i_p} \|\bar{u} - u_p\|^2_U - \|f(\bar{u}) - f(u_p)\|^2_{\mathbb{R}^k} + \|\xi\|_U \|\bar{u} - u_p\|_U , \end{aligned}$$

from which both claims can be obtained easily. □

Under more assumptions on the specific form of the functions $f_1, \ldots, f_k$ it is possible to improve the results of the Lemmata 3.46 and 3.48.

Assumption 4. *Assume that the functions $f_1, \dots, f_k$ are quadratic, i.e. it holds*

$$f_i(u+h) = f_i(u) + \langle \nabla f_i(u), h\rangle_U + \frac{1}{2}\langle \nabla^2 f_i(u)h, h\rangle_U$$

for all $u \in U_{ad}$, $h \in U$ such that $u + h \in U_{ad}$ and for all $i \in \{1, \dots, k\}$.

Theorem 3.49. *Let Assumptions 1, 3 and 4 be satisfied. Furthermore, let $z^1, z^2 \in P_Y + \mathbb{R}^k_{\leq}$ be two reference points and $\bar{u}^1, \bar{u}^2 \in U_{ad}$ the respective minimizers of F_{z^1} and F_{z^2}. Then it holds*

$$\left\|\bar{u}^1 - \bar{u}^2\right\|_U \leq \left[2C_{i_p}\left(f_{i_p}(\bar{u}^1) - z^1_{i_p} + f_{i_p}(\bar{u}^2) - z^2_{i_p}\right)\right]^{-\frac{1}{2}} \left\|z^1 - z^2\right\|_{\mathbb{R}^k}.$$

Proof. Exactly as in the proof of Lemma 3.46 we get

$$0 \leq \sum_{i=1}^{k} \left[\left(f_i(\bar{u}^1) - z^1_i\right)\langle \nabla f_i(\bar{u}^1), \bar{u}^2 - \bar{u}^1\rangle_U + \left(f_i(\bar{u}^2) - z^2_i\right)\langle \nabla f_i(\bar{u}^2), \bar{u}^1 - \bar{u}^2\rangle_U\right]. \tag{3.23}$$

By assumption, the functions $f_1, \dots, f_k$ are quadratic and plugging

$$\langle \nabla f_i(\bar{u}^{j_1}), \bar{u}^{j_2} - \bar{u}^{j_1}\rangle_U = f_i(\bar{u}^{j_2}) - f_i(\bar{u}^{j_1}) - \frac{1}{2}\langle \nabla^2 f_i(\bar{u}^{j_1})(\bar{u}^{j_2} - \bar{u}^{j_1}), \bar{u}^{j_2} - \bar{u}^{j_1}\rangle_U$$

into (3.23) yields by using Assumption 3

$$\begin{aligned}
0 \leq & \sum_{i=1}^{k}\left[\left(f_i(\bar{u}^1) - z^1_i\right)\left(-\frac{1}{2}\langle \nabla^2 f_i(\bar{u}^1)(\bar{u}^2 - \bar{u}^1), \bar{u}^2 - \bar{u}^1\rangle_U + \left(f_i(\bar{u}^2) - f_i(\bar{u}^1)\right)\right)\right. \\
& \left. + \left(f_i(\bar{u}^2) - z^2_i\right)\left(-\frac{1}{2}\langle \nabla^2 f_i(\bar{u}^2)(\bar{u}^1 - \bar{u}^2), \bar{u}^1 - \bar{u}^2\rangle_U + \left(f_i(\bar{u}^1) - f_i(\bar{u}^2)\right)\right)\right] \\
\leq & -\frac{f_{i_p}(\bar{u}^1) - z^1_{i^p} + f_{i_p}(\bar{u}^2) - z^2_{i^p}}{2} C_{i_p}\left\|\bar{u}^2 - \bar{u}^1\right\|^2_U - \left\|f(\bar{u}^2) - f(\bar{u}^1)\right\|^2_{\mathbb{R}^k} \\
& + \sum_{i=1}^{k}(f_i(\bar{u}^2) - f_i(\bar{u}^1))\left(z^2_i - z^1_i\right).
\end{aligned} \tag{3.24}$$

Furthermore, we observe by using the Cauchy-Schwarz inequality and Young's inequality

$$\begin{aligned}
\sum_{i=1}^{k}(f_i(\bar{u}^2) - f_i(\bar{u}^1))\left(z^2_i - z^1_i\right) &\leq \left\|f(\bar{u}^2) - f(\bar{u}^1)\right\|_{\mathbb{R}^k}\left\|z^2 - z^1\right\|_{\mathbb{R}^k} \\
&\leq \left\|f(\bar{u}^2) - f(\bar{u}^1)\right\|^2_{\mathbb{R}^k} + \frac{1}{4}\left\|z^2 - z^1\right\|^2_{\mathbb{R}^k}
\end{aligned}$$

and plugging this into (3.24) implies

$$0 \leq -\frac{f_{i_p}(\bar{u}^1) - z^1_{i^p} + f_{i_p}(\bar{u}^2) - z^2_{i^p}}{2} C_{i_p}\left\|\bar{u}^2 - \bar{u}^1\right\|^2_U + \frac{1}{4}\left\|z^2 - z^1\right\|^2_{\mathbb{R}^k},$$

from which the claim follows. □

Remark 3.50. Let $z^1, z^2 \in P_Y + \mathbb{R}^k_{\leq}$ be two reference points. As $f_{i_p}(\bar{u}^2) - z^2_{i_p} \geq 0$, it holds

$$f_{i_p}(\bar{u}^1) - z^1_{i_p} + f_{i_p}(\bar{u}^2) - z^2_{i_p} \geq f_{i_p}(\bar{u}^1) - z^1_{i_p},$$

which is independent of $f_{i_p}(\bar{u}^2)$ and z^2. So the mapping $\{z \in P_Y + \mathbb{R}^k_{\leq} \mid f_{i_p}(\bar{u}^z) - z_{i_p} > 0\} \to U_{\mathrm{ad}}$, $z \mapsto \bar{u}^z$ is even locally Lipschitz continuous.

Theorem 3.51. *Let Assumptions 1, 3 and 4 be satisfied and $z \in P_Y + \mathbb{R}^k_{\leq}$ be arbitrary as well as $\bar{u} \in U_{ad}$ the unique solution of (ERPP)$_z$. Let furthermore $u_p \in U_{ad}$ be arbitrary. If $f(u_p) \geq z$ and $f_{i_p}(\bar{u}) + f_{i_p}(u_p) > 2z_{i_p}$, then it holds*

$$\|\bar{u} - u_p\|_U \leq \left[C_{i_p}\left(\frac{f_{i_p}(\bar{u}) + f_{i_p}(u_p)}{2} - z_{i_p}\right)\right]^{-1} \|\xi\|_U \quad \text{and}$$

$$\|f(\bar{u}) - f(u_p)\|_{\mathbb{R}^k} \leq \frac{1}{2}\left[C_{i_p}\left(\frac{f_{i_p}(\bar{u}) + f_{i_p}(u_p)}{2} - z_{i_p}\right)\right]^{-\frac{1}{2}} \|\xi\|_U,$$

where $\xi \in U$ is given such that

$$\langle \nabla F_z(u_p) + \xi, u - u_p\rangle_U \geq 0 \text{ for all } u \in U_{ad} \tag{3.25}$$

is fulfilled.

Proof. Let ξ be chosen such that (3.25) is fulfilled. As $\bar{u}$ fulfils the first-order condition (3.10), it holds by using the same arguments as in the proof of Lemma 3.48

$$0 \leq \langle \nabla F_z(\bar{u}), u_p - \bar{u}\rangle_U + \langle \nabla F_z(u_p), \bar{u} - u_p\rangle_U + \langle \xi, \bar{u} - u_p\rangle_U.$$

With the same arguments as in the proof of Theorem 3.49, we conclude now

$$\begin{aligned}
0 \leq & \sum_{i=1}^{k}\left[(f_i(\bar{u}) - z_i)\left(-\frac{1}{2}\langle \nabla^2 f_i(\bar{u})(u_p - \bar{u}), u_p - \bar{u}\rangle_U + (f_i(u_p) - f_i(\bar{u}))\right)\right. \\
& \left. + (f_i(u_p) - z_i)\left(-\frac{1}{2}\langle \nabla^2 f_i(u_p)(\bar{u} - u_p), \bar{u} - u_p\rangle_U + (f_i(\bar{u}) - f_i(u_p))\right)\right] + \langle \xi, \bar{u} - u_p\rangle_U \\
\leq & -\left(\frac{f_{i_p}(\bar{u}) + f_{i_p}(u_p)}{2} - z_{i_p}\right) C_{i_p} \|\bar{u} - u_p\|^2_U - \|f(\bar{u}) - f(u_p)\|^2_{\mathbb{R}^k} + \|\xi\|_U \|\bar{u} - u_p\|_U,
\end{aligned}$$

from which both claims can be obtained easily. □

Geometrical Properties

Combining Theorem 3.47 with Lemma 3.42, we can formulate the following corollary.

Corollary 3.52. *Let Assumption 1 be satisfied and $z^1, z^2 \in P_Y + \mathbb{R}^k_{\leq}$ be arbitrary. Moreover, let $\bar{u}^1, \bar{u}^2 \in U_{ad}$ be the unique minimizers of F_{z^1} and F_{z^2}, respectively. Denote by $\mathcal{R} \subset \mathbb{R}^k$ the ray starting in $f(\bar{u}^1)$ and going through z^1. Then it holds*

$$\left\|f(\bar{u}^1) - f(\bar{u}^2)\right\|_{\mathbb{R}^k} \leq \operatorname{dist}(z^2, \mathcal{R}).$$

Proof. Obviously, $\mathcal{R}$ is convex and closed, so that by Theorem 2.14 there exists a unique $\tilde{z} \in \mathcal{R}$ with $\operatorname{dist}(z^2, \mathcal{R}) = \left\|z^2 - \tilde{z}\right\|_{\mathbb{R}^k}$. As $\tilde{z} \in \mathcal{R}$, we know by Lemma 3.42 that $\bar{u}^1$ is also the minimizer of $F_{\tilde{z}}$. But then it holds by using Theorem 3.47

$$\left\|f(\bar{u}^1) - f(\bar{u}^2)\right\|_{\mathbb{R}^k} \leq \left\|\tilde{z} - z^2\right\|_{\mathbb{R}^k} = \operatorname{dist}(z^2, \mathcal{R}).$$

□

Corollary 3.53. *Let Assumption 1 be satisfied. Furthermore, let $z^1, z^2 \in P_Y + \mathbb{R}^k_{\leq}$ be arbitrary and $\bar{u}^1, \bar{u}^2 \in U_{ad}$ be the unique minimizers of F_{z^1} and F_{z^2}, respectively. Denote by $\mathcal{R} \subset \mathbb{R}^k$ the ray starting in $f(\bar{u}^1)$ and going through z^1 and let $r := \operatorname{dist}(z^2, \mathcal{R})$. Then it holds*

$$f(\bar{u}^2) \in B(f(\bar{u}^1), r) \cap \{f(\bar{u}^1) + x \mid \langle x, f(\bar{u}^1) - z\rangle_{\mathbb{R}^k} \geq 0\} \cap (\{f(\bar{u}^1)\} + \mathbb{R}^k_{\leq})^C.$$

Proof. This follows immediately from Corollary 3.52, Lemma 3.25 and the definition of Pareto optimality. □

3.4 Algorithm to approximate the Pareto Front in the Bicriterial Case

In this section we construct an algorithm in order approximate the Pareto front of a given multiobjective optimization problem with two objective functions numerically. We introduced several methods to compute Pareto optimal points in the last section, but not all of them are suitable for the numerical approximation of the Pareto front. The most important properties that a method has to fulfil are:

(i) Ability to compute each point on the Pareto front.

(ii) Allowing an adaptive method to ensure a good approximation of the Pareto front, i.e. that the approximated points are reasonably distributed.

(iii) The resulting scalar optimization problems are differentiable, if the objective functions $f_1, \ldots, f_k$ are differentiable.

Methods yielding non-differentiable scalar optimization problems are hard to handle numerically as we would have to use algorithms for non-smooth optimization. By this criteria we can exclude the method presented in Lemma 3.30 using the Chebyshev-norm, although it fulfils the other two criteria (i) and (ii).
Both the weighted sum method and reference point methods with suitable functions g, in particular the Euclidean reference point method, fulfil the first two properties (i) and (ii). (Notice that in the case of two objective functions, there are only two points that we cannot compute using reference point methods. Those can be obtained by solving two weighted sum problems instead.)
However, it is hard to find an adaptive method for choosing the weights such that the weighted sum method ensures a good approximation of the Pareto front. By Theorem 3.22 the solution of the weighted sum problem in the feasible space depends continuously on the weights, but we do not know how the corresponding objective vectors behave when changing the weights. So in

general we have no analytical covering that we get a good approximation of the Pareto front (see also e.g. [6]).
For the Euclidean reference point method, Theorem 3.47 guarantees exactly this property, allowing us to construct a good adaptive method for approximating the Pareto front. This will be done in the following.

3.4.1 Algorithm using the Euclidean Reference Point Method

Basic Algorithm

The idea of the algorithm using the Euclidean reference point method is straightforward. An initial point on the Pareto front is computed and an initial reference point z is generated. Until the end of the Pareto front is reached, $(\mathrm{ERPP})_z$ is solved repeatedly and new reference points are generated using the information about previous solutions and reference points. As a pseudocode this looks like the following.

Algorithm 1: Basic Algorithm to compute the Pareto front

Data: $n_{\max}$: Maximal number of points

begin
- Solve $(\mathrm{WSP})_\alpha$ for $\alpha = (1, 0)$ and save the solution $\bar{u}^0$;
- $\bar{y}^0 = f(\bar{u}^0)$;
- Set initial reference point z^1;
- $n := 0$;
- **while** $n < n_{max}$ *& end of Pareto front is not reached yet* **do**
  - $n \leftarrow n + 1$;
  - Set initial guess u_0^n;
  - Solve $(\mathrm{ERPP})_{z^n}$ with initial guess u_0^n and save the solution $\bar{u}^n$;
  - $\bar{y}^n = f(\bar{u}^n)$;
  - Set new reference point z^{n+1};

In the following we will specify the exact approach. Open questions that will be answered are

- How is the initial guess u_0^n chosen?
- What are possible methods to solve $(\mathrm{ERPP})_z$ for a given reference point z?
- How are the reference points constructed?
- How do we know when the end of the Pareto front is reached?

Choice of Initial Guess

Lemma 3.46 – and Theorem 3.49 in the case of quadratic functions – tell us that the solution of the Euclidean reference point problem depends continuously on the reference point. Therefore, it is reasonable to choose the previous solution as a new initial guess, i.e. $u_0^{n+1} = \bar{u}^n$.

Solving an Euclidean Reference Point Problem

There are many numerical methods to tackle the minimization of the scalar function F_z. Depending on the design of the set U_{ad}, we might be confronted with unconstrained optimization or constrained optimization with equality and/or inequality constraints or even more complicated constraints. For a good overview of methods see for example [19, 25].
If Assumptions 1, 3 and 4 are satisfied, we are able to formulate a stopping condition for the solving method by using the statement of Theorem 3.51, given that the computation of ξ fulfilling (3.25) is possible and not too costly in terms of computation time. As we know that $f(\bar{u}) \geq z$ holds for the minimizer $\bar{u}$ of F_z, we know by Theorem 3.51 that it holds for the current iterative u_p

$$\|\bar{u} - u_p\|_U \leq \left[C_{i_p}\left(\frac{f_{i_p}(u_p) - z_{i_p}}{2}\right)\right]^{-1} \|\xi\|_U \quad \text{and}$$
$$\|f(\bar{u}) - f(u_p)\|_{\mathbb{R}^k} \leq \frac{1}{2}\left[C_{i_p}\left(\frac{f_{i_p}(u_p) - z_{i_p}}{2}\right)\right]^{-\frac{1}{2}} \|\xi\|_U ,$$

if $f(u_p) \geq z$ is fulfilled. Using this as a stopping condition, we can directly control how large the error between the true and the numerical solution can get in both the feasible and the objective space.

Construction of Reference Points

The most important part of an algorithm approximating the Pareto front using the Euclidean reference point method is the choice of the reference points. Results from the previous sections will help us to establish an algorithm to choose the new reference point based on the previous reference and Pareto optimal point.

We start with the most basic property that a reference point z has to fulfil: Theorem 3.35 requires that $z \in P_Y + \mathbb{R}^k_{\leq}$ holds, i.e. z has to lie below the Pareto front to ensure the unique solvability and the Pareto optimality of the solution of $(\text{ERPP})_z$.
Let us assume that we have a recent reference point z^n and that we have already computed the solution $\bar{u}^n$ of $(\text{ERPP})_{z^n}$. Lemma 3.25 indicates that $H_{z^n} := \{f(\bar{u}^n) + x \mid \langle x, f(\bar{u}^n) - z^n)\rangle_{\mathbb{R}^k} \leq 0\}$ lies below the set P_Y. So it is reasonable to choose our next reference point z^{n+1} such that $z^{n+1} \in H_{z^n}$. We can ensure this by setting

$$z^{n+1} := f(\bar{u}^n) + h_x \left(\frac{f(\bar{u}^n) - z^n}{\|f(\bar{u}^n) - z^n\|_{\mathbb{R}^k}}\right)^{\perp} - h_p \frac{f(\bar{u}^n) - z^n}{\|f(\bar{u}^n) - z^n\|_{\mathbb{R}^k}} \tag{3.26}$$

for some $h_x \in \mathbb{R}$ and $h_p \geq 0$ (see Figure 3.3). If we choose $h_x = 0$, we get by Lemma 3.42 that the solution to $(\text{ERPP})_{z^{n+1}}$ is again $\bar{u}^n$. So $h_x \neq 0$ is necessary for getting a different Pareto optimal point.
Now the question arises on how to choose the parameters h_x and h_p such that we can analytically ensure a good approximation of the Pareto front. For this purpose we can use Corollary 3.52, which states that

$$\left\|f(\bar{u}^{n+1}) - f(\bar{u}^n)\right\|_{\mathbb{R}^k} \leq |h_x| . \tag{3.27}$$

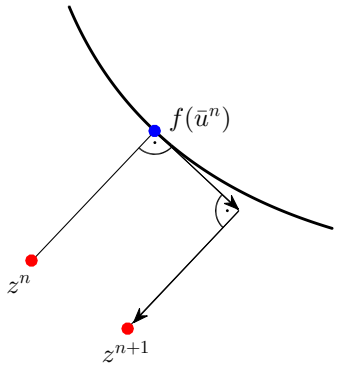


Figure 3.3: Generation of a new reference point

Consequently, h_x can be chosen according to our preferred approximation quality and h_p to control how large the distance between the reference points and the Pareto front is.
This was the case in which we already have a recent reference point z^n and a corresponding Pareto optimal point $f(\bar{u}^n)$. We still have to specify how to choose the first reference point after solving the weighted sum method for $\alpha = (1,0)$. Let $\bar{u}^\alpha$ denote the solution to $(\text{WSP})_\alpha$. Then we apply the formula $z^1 := f(\bar{u}^\alpha) - h_x(0,1)^T - h_p(1,0)^T$. By Lemma 3.23 it holds again $z^1 \in P_Y + \mathbb{R}^k_{\leq}$. Overall, we propose the algorithm

$$\begin{aligned} z^1 &:= f(\bar{u}^\alpha) - h_x(0,1)^T - h_p(1,0)^T, \\ z^{n+1} &:= f(\bar{u}^n) + h_x \left(\frac{f(\bar{u}^n) - z^n}{\|f(\bar{u}^n) - z^n\|_{\mathbb{R}^k}} \right)^\perp - h_p \frac{f(\bar{u}^n) - z^n}{\|f(\bar{u}^n) - z^n\|_{\mathbb{R}^k}} \quad \text{for all } n \geq 1 \end{aligned}$$

for suitable parameters h_x, h_p to compute the reference points.

Remark 3.54. This choice of reference points together with Corollary 3.53 implies

$$f_2(\bar{u}^{n+1}) - z_2^{n+1} \geq \frac{h_p}{\|f(\bar{u}^n) - z^n\|_{\mathbb{R}^k}} (f_2(\bar{u}^n) - z_2^n).$$

In a situation, where $\bar{u}^n$, z^n and z^{n+1} are known, this can be used in the Theorems 3.49 and 3.51 to get better estimates.

Stopping Condition

By using the weighted sum method with weights $\alpha = (0,1)$, we can easily compute the 'end' of the Pareto front. So we can stop the iteration if the new reference point z^{n+1} fulfils $z_1^{n+1} \geq f_1(\bar{u}^\alpha)$.

Chapter 4

Proper Orthogonal Decomposition for Abstract Evolution Problems

The goal of this chapter is to introduce the concept of model order reduction for abstract evolution equations. In particular, we present a specific method of model order reduction, namely the proper orthogonal decomposition, which is widely used in the context of optimal control or parameter-dependent problems. As we only give a brief introduction into the topic, we refer to [33] for a good overview and the possible applications of model order reduction in various fields.

4.1 Model Order Reduction

In the following we consider similarly to Section 2.6 an abstract evolution equation of the form

$$y_t(t) + a(t, y(t), \cdot) = f(t, u) \text{ in } V' \quad (t \in (0, T)), \tag{4.1a}$$

$$y(0) = y_0, \tag{4.1b}$$

where $V \subset H \subset V'$ is a Gelfand Triple and $a : [0, T] \times V \times V \to \mathbb{R}$ is a bilinear form, such that $a \in C([0, T]; L(V, V'))$ with $\|a\|_{C([0,T];L(V,V'))} =: \gamma$ is additionally coercive with coercivity constants α, β. Furthermore, U is a vector space and $f : [0, T] \times U \to V'$ is supposed to fulfil $f(\cdot, u) \in L^2(0, T; V')$ for all $u \in U$.

Corollary 4.1. *For each $u \in U$, $y_0 \in H$ (4.1) has a unique solution $y = y(u) \in W(0, T)$, which fulfils the a-priori estimate*

$$\|y\|_{W(0,T)} \leq C \left(\|f(\cdot, u)\|_{L^2(0,T;V')} + \|y_0\|_H \right).$$

Proof. This follows immediately from Theorem 2.28. □

In applications this situation appears in parameter-dependent or optimal control problems. The space U can thus be seen as parameter or control space. It is characteristic for these problems that (4.1) has to be solved iteratively for different inputs $u \in U$. If we direct our attention to the numerical implementation, the first idea is to apply a finite element discretization to (4.1). By doing this, we get a high- but finite-dimensional subspace $V_{FE}^\ell \subset V$ and a reduced-order evolution equation

$$y_t^\ell(t) + a(t, y^\ell(t), \cdot) = f^\ell(t, u) \text{ in } (V_{FE}^\ell)' \quad (t \in (0, T)), \tag{4.2a}$$

$$y^\ell(0) = y_0^\ell, \tag{4.2b}$$

where in this case $y_0^\ell \in V_{FE}^\ell$ has to hold and f^ℓ is the finite dimensional approximation of f.

Of course, one is not restricted to finite element spaces V_{FE}^{ℓ}, but it is possible to consider (4.2) for an arbitrary subspace $V^{\ell} \subset V$. For these we get the following result.

Theorem 4.2. *Let $V^{\ell} \subset V$ be finite-dimensional, $y_0^{\ell} \in V^{\ell}$ and $u \in U$ such that $f^{\ell}(\cdot, u) \in L^2(0,T;(V^{\ell})')$ holds. Then there is a unique solution $y^{\ell} \in H^1(0,T;V)$ of (4.2), which fulfils the a-priori estimate*

$$\left\|y^{\ell}\right\|_{H^1(0,T;V)} \leq C \left(\left\|f^{\ell}(\cdot,u)\right\|_{L^2(0,T;(V^{\ell})')} + \left\|y_0^{\ell}\right\|_{V^{\ell}} \right).$$

with a constant C independent of the finite-dimensional space V^{ℓ}.

Proof. The proof of a similar version of this theorem can be found in [12, pp. 354-356, Theorems 1 and 2]. □

Equation (4.2) can be solved numerically by using a time-discretization method as for example the backward Euler method or the Crank-Nicolson method. However, it is easy to see that applying the finite element method to parameter-dependent or optimal control problems gets quickly too costly because finite element spaces are usually high-dimensional and computing solutions of (4.2) for different inputs u hence requires the repeated solving of very large equation systems.
The idea of model order reduction is to find a low-dimensional subspace $V^{\ell} \subset V$ – in applications $V^{\ell} \subset V_{FE}^{\ell}$ most of the times – for which the solution of the reduced-order evolution equation (4.2) is still a good approximation to the solution of (4.1). The difference to the finite element method is the following: Whereas the finite element subspace V_{FE}^{ℓ} is supposed to be a good approximation for the whole space V, the reduced-order subspace V^{ℓ} shall only approximate a space that contains in some sense the characteristics of the solutions of (4.1). For many equations, in particular linear equations, we can hope that solving the reduced-order equation for these low-dimensional subspaces V^{ℓ} is already approximating the solution of the full equation quite well. Of course, there are different approaches and methods to construct such low-dimensional subspaces. In the next section we will focus on one of them, namely proper orthogonal decomposition.
However, before dealing with details about how to construct the low-dimensional subspace V^{ℓ}, we will show a result about the quality of the solution of the reduced-order equation (4.2) in comparison to the solution of the full equation (4.1).
Assume therefore that $V^{\ell} \subset V$ is a finite-dimensional subspace.

Definition 4.3. Let the projection of V onto the subspace V^{ℓ} introduced in Definition and Theorem 2.15 be denoted by $\mathcal{P}^{\ell}$.

The following theorem is a slightly modified version of [15, Theorem 1.29], from which the structure of the proof is taken.

Theorem 4.4. *Let $u \in U$, $y_0 \in H$ and $y_0^{\ell} \in V^{\ell}$ be arbitrary and denote by y, y^{ℓ} the solution to (4.1) and (4.2), respectively. Then there is a constant $C = C(\alpha, \beta, \gamma, T)$ such that*

$$\left\|y^{\ell} - y\right\|^2_{L^2(0,T;V)} \leq C \left(\left\|y - \mathcal{P}^{\ell} y\right\|^2_{W(0,T)} + \left\|f^{\ell}(\cdot,u) - f(\cdot,u)\right\|^2_{L^2(0,T;V')} + \left\|y_0^{\ell} - \mathcal{P}^{\ell} y_0\right\|^2_{H} \right)$$

holds.

Proof. We can write for almost all $t \in [0,T]$

$$y^\ell(t) - y(t) = y^\ell(t) - \mathcal{P}^\ell(y(t)) + \mathcal{P}^\ell(y(t)) - y(t) =: \theta^\ell(t) + \rho^\ell(t),$$

where $\theta^\ell(t) := y^\ell(t) - \mathcal{P}^\ell(y(t)) \in V^\ell$ and $\rho^\ell(t) := \mathcal{P}^\ell(y(t)) - y(t) \in V$. For ρ^ℓ we get

$$\left\|\rho^\ell\right\|^2_{L^2(0,T;V)} = \left\|\mathcal{P}^\ell y - y\right\|^2_{L^2(0,T;V)}. \tag{4.3}$$

Furthermore, we get for all $\psi \in V^\ell$ and for almost all $t \in [0,T]$ by using (4.1a), (4.2a), the continuity of a and $(\mathcal{P}^\ell(y(t)))_t = \mathcal{P}^\ell(y_t(t))$

$$\begin{aligned}
&\frac{d}{dt}\langle\theta^\ell(t),\psi\rangle_H + a(t,\theta^\ell_t(t),\psi)\\
=&\langle y^\ell_t(t),\psi\rangle_{V',V} + a(t,y^\ell(t),\psi) - \langle\mathcal{P}^\ell(y_t(t)),\psi\rangle_{V',V} - a(t,\mathcal{P}^\ell(y(t)),\psi)\\
=&\langle f^\ell(t,u),\psi\rangle_{V',V} - \langle f(t,u),\psi\rangle_{V',V} + \langle y_t(t),\psi\rangle_{V',V} + a(t,y(t),\psi)\\
&- \langle\mathcal{P}^\ell(y_t(t)),\psi\rangle_{V',V} - a(t,\mathcal{P}^\ell(y(t)),\psi)\\
=&\langle f^\ell(t,u) - f(t,u),\psi\rangle_{V',V} + \langle y_t(t) - \mathcal{P}^\ell(y_t(t)),\psi\rangle_{V',V} + a(t,y(t) - \mathcal{P}^\ell(y(t)),\psi)\\
\le&\left\|f^\ell(t,u) - f(t,u)\right\|_{V'}\|\psi\|_V + \left\|y_t(t) - \mathcal{P}^\ell(y_t(t))\right\|_{V'}\|\psi\|_V + \gamma\left\|y(t) - \mathcal{P}^\ell(y(t))\right\|_V\|\psi\|_V.
\end{aligned} \tag{4.4}$$

Choosing $\psi = \theta^\ell(t) \in V^\ell$ in (4.4), using the coercivity of a and Young's inequality yields

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\left\|\theta^\ell(t)\right\|^2_H + \alpha\left\|\theta^\ell(t)\right\|^2_V - \beta\left\|\theta^\ell(t)\right\|^2_H\\
\le&\left\|f^\ell(t,u) - f(t,u)\right\|_{V'}\left\|\theta^\ell(t)\right\|_V + \left\|y_t(t) - \mathcal{P}^\ell(y_t(t))\right\|_{V'}\left\|\theta^\ell(t)\right\|_V\\
&+ \gamma\left\|y(t) - \mathcal{P}^\ell(y(t))\right\|_V\left\|\theta^\ell(t)\right\|_V\\
\le&\frac{3}{2\alpha}\left\|f^\ell(t,u) - f(t,u)\right\|^2_{V'} + \frac{\alpha}{6}\left\|\theta^\ell(t)\right\|^2_V + \frac{3}{2\alpha}\left\|y_t(t) - \mathcal{P}^\ell(y_t(t))\right\|^2_{V'} + \frac{\alpha}{6}\left\|\theta^\ell(t)\right\|^2_V\\
&+ \frac{3\gamma^2}{2\alpha}\left\|y(t) - \mathcal{P}^\ell(y(t))\right\|^2_V + \frac{\alpha}{6}\left\|\theta^\ell(t)\right\|^2_V,
\end{aligned}$$

and so

$$\begin{aligned}
&\frac{d}{dt}\left\|\theta^\ell(t)\right\|^2_H + \alpha\left\|\theta^\ell(t)\right\|^2_V - 2\beta\left\|\theta^\ell(t)\right\|^2_H\\
\le&\frac{3}{\alpha}\left\|f^\ell(t,u) - f(t,u)\right\|^2_{V'} + \frac{3}{\alpha}\left\|y_t(t) - \mathcal{P}^\ell(y_t(t))\right\|^2_{V'} + \frac{3\gamma^2}{\alpha}\left\|y(t) - \mathcal{P}^\ell(y(t))\right\|^2_V =: \chi(t).
\end{aligned} \tag{4.5}$$

Gronwall's Lemma yields

$$\begin{aligned}
\left\|\theta^\ell(t)\right\|^2_H \le&\exp\left(\int_0^t 2\beta\,ds\right)\left(\left\|\theta^\ell(0)\right\|^2_H + \int_0^t \chi(s)\,ds\right)\\
\le&\exp(2\beta T)\left(\left\|\mathcal{P}^\ell y_0 - y^\ell_0\right\|^2_H + \int_0^T \chi(s)\,ds\right)\\
\le&c_1\left(\left\|\mathcal{P}^\ell y_0 - y^\ell_0\right\|^2_H + \left\|f^\ell(\cdot,u) - f(\cdot,u)\right\|^2_{L^2(0,T;V')} + \left\|y_t - \mathcal{P}^\ell y_t\right\|^2_{L^2(0,T;V')}\right.\\
&\left.+ \left\|y - \mathcal{P}^\ell y\right\|^2_{L^2(0,T;V)}\right),
\end{aligned} \tag{4.6}$$

with $c_1 := \exp(2\beta T)\max(1, \frac{3}{\alpha}, \frac{3\gamma^2}{\alpha})$. Using (4.5) again, we get by using (4.6) and integrating over $(0,T)$

$$\begin{aligned}\left\|\theta^\ell\right\|^2_{L^2(0,T;V)} \leq & \frac{1}{\alpha}\left(2\beta\left\|\theta^\ell\right\|^2_{L^2(0,T;H)} + \frac{3}{\alpha}\left\|f^\ell(\cdot,u) - f(\cdot,u)\right\|^2_{L^2(0,T;V')} + \frac{3}{\alpha}\left\|y_t - \mathcal{P}^\ell y_t\right\|^2_{L^2(0,T;V')}\right.\\ & \left. + \frac{2\gamma^2}{\alpha}\left\|y - \mathcal{P}^\ell y\right\|^2_{L^2(0,T;V)}\right)\\ \leq & c_2\left(\left\|\mathcal{P}^\ell y_0 - y_0^\ell\right\|^2_H + \left\|f^\ell(\cdot,u) - f(\cdot,u)\right\|^2_{L^2(0,T;V')} + \left\|y_t - \mathcal{P}^\ell y_t\right\|^2_{L^2(0,T;V')}\right.\\ & \left. + \left\|y - \mathcal{P}^\ell y\right\|^2_{L^2(0,T;V)}\right), \end{aligned} \tag{4.7}$$

with $c_2 = c_2(\alpha,\beta,\gamma,T)$. Altogether we can conclude by using (4.3) and (4.7) that there is a constant $C = C(\alpha,\beta,\gamma,T)$ such that

$$\left\|y^\ell - y\right\|^2_{L^2(0,T;V)} \leq C\left(\left\|y - \mathcal{P}^\ell y\right\|^2_{W(0,T)} + \left\|f^\ell(\cdot,u) - f(\cdot,u)\right\|^2_{L^2(0,T;V')} + \left\|y_0^\ell - \mathcal{P}^\ell y_0\right\|^2_H\right)$$

holds. □

4.2 Proper Orthogonal Decomposition

As already mentioned in the previous section, proper orthogonal decomposition (POD) is one method to construct a reduced-order subspace $V^\ell \subset V$. The idea of POD is to create a low dimensional subspace which is still capturing the most dominant dynamics of the investigated system. Therefore, it needs data obtained by previous experiments – in our case previous solutions of the considered PDE computed by the finite element method.
For a theoretical introduction of the POD method, we start with the continuous case, i.e. the case in which the provided data are functions mapping from $[0,T]$ to V. Later, we will focus on the numerical implementation of the POD method, for which only discrete data will be available. The whole section follows strongly [15, Chapter 1.2].

4.2.1 Continuous Version of the POD Method

Given p trajectories $y^1,\dots,y^p \in L^2(0,T;V)$, we consider the linear subspace $\mathcal{V} := \operatorname{span}\{y^i(t) \mid 1 \leq i \leq p, t \in [0,T]\}$. For this space we want to find a low-dimensional subspace V^ℓ, such that the difference between y and its projection onto V^ℓ is as small as possible for all $y \in \mathcal{V}$.

Definition 4.5. Let $y^1,\dots,y^p \in L^2(0,T;V)$ be given trajectories, possibly solutions of (4.1) for different inputs $u \in U$ or their time-derivatives. In the context of POD we call $y^1,\dots,y^p$ snapshots and define the snapshot subspace by $\mathcal{V} := \operatorname{span}\{y^i(t) \mid 1 \leq i \leq p, t \in [0,T]\}$. Furthermore, let $d := \dim(\mathcal{V}) \in \mathbb{N} \cup \{\infty\}$ be the dimension of the snapshot subspace.

Given this setting, we define the notion of a POD basis and a POD space, respectively.

Definition 4.6. Let $y^1, \dots, y^p \in L^2(0,T;V)$ be snapshots. Then for any $\ell \leq d$ the solution $\{\bar{\psi}_1, \dots, \bar{\psi}_\ell\}$ to the minimization problem

$$\min_{\psi_1,\dots,\psi_l \in V} \sum_{k=1}^{p} \int_0^T \left\| y^k(t) - \sum_{i=1}^{\ell} \langle y^k(t), \psi_i \rangle_V \psi_i \right\|_V^2 dt \qquad \text{(ContPOD)}$$
$$\text{subject to } \langle \psi_i, \psi_j \rangle_V = \delta_{ij} \text{ for } 1 \leq i,j \leq \ell,$$

is called a POD basis of rank ℓ. The corresponding linear space $V^\ell := \operatorname{span}\{\bar{\psi}_1, \dots, \bar{\psi}_\ell\}$ is called a POD space of rank ℓ.

Remark 4.7. (i) A POD basis $\{\bar{\psi}_1, \dots, \bar{\psi}_\ell\}$ is chosen such that the sum of the mean square errors between the snapshots $y^1, \dots, y^p$ and their projections onto $\operatorname{span}\{\bar{\psi}_1, \dots, \bar{\psi}_\ell\}$ is minimized.

(ii) This procedure is motivated by the idea that the POD spaces V^ℓ contain the main characteristics of the dynamics of solutions to (4.1) for different values of u. For this reason we can hope for a good approximation, if the inputs u which generate the snapshots are chosen appropriately.

There are different approaches to tackle the optimization problem (ContPOD), whose solution provides us with POD bases. In this thesis we want to present a method using an eigenvalue problem.

Definition and Theorem 4.8. Let $y^1, \dots, y^p \in L^2(0,T;V)$ be snapshots. Then the operator

$$\mathcal{R} : V \to \mathcal{V},\ \psi \mapsto \sum_{k=1}^{p} \int_0^T \langle y^k(t), \psi \rangle_V y^k(t)\, dt$$

is linear, continuous, compact, non-negative and self-adjoint.

Proof. A proof can be found in [15, Lemma 1.13]. □

In the next theorem we will show that the operator $\mathcal{R}$ can indeed be used to solve (ContPOD). It is proved in [15, Theorem 1.15].

Theorem 4.9. *Let $y^1, \dots, y^p \in L^2(0,T;V)$ be snapshots and $\mathcal{R}$ be given as in Definition and Theorem 4.8. Then there exist non-negative eigenvalues $\{\bar{\lambda}_i\}_{i\in\mathbb{N}}$ and associated orthonormal eigenfunctions $\{\bar{\psi}_i\}_{i\in\mathbb{N}}$ satisfying*

$$\mathcal{R}\bar{\psi}_i = \bar{\lambda}_i \bar{\psi}_i, \quad \bar{\lambda}_1 \geq \dots \geq \bar{\lambda}_d > \bar{\lambda}_{d+1} = \dots = 0.$$

For every $\ell \in \{1, \dots, d\}$, the first ℓ eigenfunctions $\{\bar{\psi}_i\}_{i=1}^{\ell}$ solve (ContPOD). Moreover, it holds

$$\sum_{k=1}^{p} \int_0^T \left\| y^k(t) - \sum_{i=1}^{\ell} \langle y^k(t), \bar{\psi}_i \rangle_V \bar{\psi}_i \right\|_V^2 dt = \sum_{i=\ell+1}^{d} \bar{\lambda}_i. \qquad (4.8)$$

Remark 4.10. Although we only defined a POD basis for $\ell \in \{1, \dots, d\}$, it is beneficial for later considerations, if we set $V^\ell := \operatorname{span}\{\bar{\psi}_1, \dots, \bar{\psi}_\ell\}$ for all $\ell \in \mathbb{N}$. Note that the system $\{\bar{\psi}_i\}_{i\in\mathbb{N}}$ is an orthonormal basis of the separable Hilbert space V, so that V^ℓ is an orthonormal system for all $\ell \in \mathbb{N}$.

Analytical Results

Theorem 4.11. *Let $u \in U$ be arbitrary such that the solution $y = y(u) \in W(0,T)$ of (4.1) fulfils $y \neq 0$.*

(i) If the ℓ-dimensional POD basis $\{\bar{\psi}_i\}_{i=1}^{\ell}$ is generated by $y^1 = y$ and $y^\ell = y^\ell(u)$ is the solution of (4.2), then it holds

$$\left\|y^\ell - y\right\|^2_{L^2(0,T;V)} \leq C \left(\sum_{i=\ell+1}^{\infty} \bar{\lambda}_i + \left\|y_t - \mathcal{P}^\ell y_t\right\|^2_{L^2(0,T;V')} + \left\|f^\ell(\cdot,u) - f(\cdot,u)\right\|^2_{L^2(0,T;V')} + \left\|y_0^\ell - \mathcal{P}^\ell y_0\right\|^2_H \right).$$

(ii) If $y(u) \in H^1(0,T;V)$ holds and the ℓ-dimensional POD basis $\{\bar{\psi}_i\}_{i=1}^{\ell}$ is generated by $y^1 = y$, $y^2 = y_t$ and $y^\ell = y^\ell(u)$ is the solution of (4.2), then it holds

$$\left\|y^\ell - y\right\|^2_{L^2(0,T;V)} \leq C \left(\sum_{i=\ell+1}^{\infty} \bar{\lambda}_i + \left\|f^\ell(\cdot,u) - f(\cdot,u)\right\|^2_{L^2(0,T;V')} + \left\|y_0^\ell - \mathcal{P}^\ell y_0\right\|^2_H \right)$$

(iii) If $(V^\ell)_{\ell\in\mathbb{N}}$ are the POD spaces generated by an arbitrary family of snapshots, it holds

$$\left\|y^\ell(u) - y(u)\right\|_{L^2(0,T;V)} \to 0 \quad (\ell \to \infty)$$

for an arbitrary $u \in U$, if $y(u) \in H^1(0,T;V)$, $\left\|f^\ell(\cdot,u) - f(\cdot,u)\right\|_{L^2(0,T;V')} \to 0$ and $\left\|y_0^\ell - \mathcal{P}^\ell y_0\right\|_H \to 0$ $(\ell \to \infty)$ hold.

Proof. The parts (i) and (ii) follow directly from Theorem 4.4 and (4.8). For part (iii) we also use Theorem 4.4 and that $\left\|v - \mathcal{P}^\ell v\right\|_V \to 0$ for $\ell \to \infty$ for all $v \in V$ (see Corollary 2.17), from which we can conclude by using dominated convergence $\left\|y - \mathcal{P}^\ell y\right\|_{L^2(0,T;V)} \to 0$ as $\ell \to \infty$ for all $y \in L^2(0,T;V)$. □

4.2.2 Discrete Version of the POD Method

In this section we want to give a brief introduction into how to use the POD method in numerical applications. Imagine that we are in a situation in which we already computed finite-dimensional discrete solutions $y_1, \ldots, y_n : \{t_1, \ldots, t_n\} \to \mathbb{R}^m$ of (4.1) for different inputs, for instance by using the finite element method. Here, m is the degree of freedom of the discrete solutions. In an analogous way to the continuous case we define the space $\mathcal{V}^n := \operatorname{span}\{y_j^i \mid 1 \leq i \leq p, 1 \leq j \leq n\}$ and want to find a low-dimensional subspace $V^\ell \subset \mathcal{V}^n$, such that the difference between y and its projection onto V^ℓ is as small as possible for all $y \in \mathcal{V}^n$ in some sense, which shall be specified later. The general procedure will be exactly the same as in the continuous case.

Definition 4.12. Let $y_1^i, \ldots, y_n^i \in \mathbb{R}^m$ for $1 \leq i \leq p$ be given vectors. In the context of discrete POD, we call the vectors $y_1^i, \ldots, y_n^i \in \mathbb{R}^m$ for $1 \leq i \leq n$ discrete snapshots and define the discrete snapshot subspace by $\mathcal{V}^n := \operatorname{span}\{y_j^i \mid 1 \leq i \leq p, 1 \leq j \leq n\}$. Furthermore, let $d^n := \dim(\mathcal{V}^n) \in \{1, \ldots, np\}$ be the dimension of the discrete snapshot subspace.

Given this setting, we can introduce the notion of a discrete POD basis and a discrete POD space, respectively.

Definition 4.13. Let $y_1^i, \dots, y_n^i \in \mathbb{R}^m$ for $1 \le i \le p$ be discrete snapshots and $\alpha_1^n, \dots, \alpha_n^n > 0$ positive weighting parameters, as well as $W \in \mathbb{R}^{m\times m}$ symmetric and positive definite, such that $\langle \cdot, \cdot \rangle_W := \langle W\cdot, \cdot \rangle_{\mathbb{R}^m}$ is an inner product on $\mathbb{R}^m$. Then for any $\ell \le d^n$ the solution $\{\bar{\psi}_1, \dots, \bar{\psi}_\ell\}$ to the minimization problem

$$\min_{\psi_1,\dots,\psi_\ell \in \mathbb{R}^m} \sum_{k=1}^{p}\sum_{j=1}^{n} \alpha_j^n \left\| y_j^k - \sum_{i=1}^{\ell} \langle y_j^k, \psi_i\rangle_W \psi_i \right\|_W^2 \quad \text{subject to } \langle \psi_i, \psi_j\rangle_W = \delta_{ij} \text{ for } 1 \le i,j \le \ell, \tag{DiscPOD}$$

is called a discrete POD basis of rank ℓ. The corresponding linear space $V^\ell := \operatorname{span}\{\bar{\psi}_1, \dots, \bar{\psi}_\ell\}$ is called discrete POD space of rank ℓ.

Remark 4.14. (i) The discrete POD basis $\{\bar{\psi}_1, \dots, \bar{\psi}_\ell\}$ is chosen such that the sum of the weighted mean square errors between the snapshots $y_1^i, \dots, y_n^i \in \mathbb{R}^m$ $(1 \le i \le p)$ and their projections onto $\operatorname{span}\{\bar{\psi}_1, \dots, \bar{\psi}_\ell\}$ is minimized. If we compare this to the continuous case, it is clear that the weights $\alpha_1^n, \dots, \alpha_n^n$ are supposed to model the numerical integration and should thus be chosen accordingly.

(ii) A weighted inner product $\langle \cdot, \cdot \rangle_W$ can be used to model the inner product on the finite element space.

Exactly as in the continuous case we want to tackle the optimization problem (DiscPOD) by using an eigenvalue problem. Therefore, we make the next definition.

Definition and Theorem 4.15. Let $y_1^i, \dots, y_n^i \in \mathbb{R}^m$ for $1 \le i \le p$ be discrete snapshots. Then the operator

$$\mathcal{R}^n : \mathbb{R}^m \to \mathcal{V}^n,\ \psi \mapsto \sum_{k=1}^{p}\sum_{j=1}^{n} \alpha_j^n \langle y_j^k, \psi\rangle_W y_j^k$$

is linear, continuous, compact, non-negative and self-adjoint.

Proof. A proof can be found in [15, Lemma 1.3]. □

As in the continuous case it can be shown that the solution of (DiscPOD) is given by eigenvectors of the operator defined in Definition and Theorem 4.15.

Theorem 4.16. *Let $y_1^i, \dots, y_n^i \in \mathbb{R}^m$ for $1 \le i \le p$ be discrete snapshots and $\mathcal{R}^n$ be given as in Definition and Theorem 4.15. Then there exist non-negative eigenvalues $\{\bar{\lambda}_i^n\}_{i=1}^m$ and associated orthonormal eigenvectors $\{\bar{\psi}_i^n\}_{i=1}^m$ satisfying*

$$\mathcal{R}^n \bar{\psi}_i^n = \bar{\lambda}_i^n \bar{\psi}_i^n, \quad \bar{\lambda}_1^n \ge \dots \ge \bar{\lambda}_m^n. \tag{4.9}$$

For every $\ell \in \{1, \dots, \min(d^n, m)\}$, the first ℓ eigenfunctions $\{\bar{\psi}_i^n\}_{i=1}^\ell$ solve (DiscPOD). Moreover, it holds

$$\sum_{k=1}^{p}\sum_{j=1}^{n} \alpha_j^n \left\| y_j^k - \sum_{i=1}^{\ell} \langle y_j^k, \bar{\psi}_i^n\rangle_W \bar{\psi}_i^n \right\|_W^2 = \sum_{i=\ell+1}^{d} \bar{\lambda}_i^n. \tag{4.10}$$

Proof. For a proof see [15, Theorem 1.8]. □

Remark 4.17. (i) As already mentioned in Remark 4.14, it is essential to choose the weights $\alpha_1^n, \ldots, \alpha_n^n$ appropriately. It can be shown that using trapezoidal weights

$$\alpha_1^n := \frac{T}{2(n-1)}, \; \alpha_j^n := \frac{T}{n-1} \text{ for } 2 \leq j \leq n-1, \; \alpha_n^n := \frac{T}{2(n-1)} \tag{4.11}$$

are a reasonable choice (for a convergence result see [15, Theorem 1.19]).

(ii) There are different numerical methods to solve the eigenvalue problem (4.9). For a nice description of a method using singular value decomposition (SVD), we refer to [15, pp. 10-13]. This method is also used later in the numerical experiments.

Chapter 5

Bicriterial Optimal Control of Convection-Diffusion Equations

In this chapter an optimal control problem modelling an energy efficient heating, ventilation and air conditioning (HVAC) operation of a room is introduced. The control input models the heating and the controlled outcome is the temperature in the room. Seeing the deviation of the temperature from a prescribed temperature and the heating costs as the two objectives of the cost functions, we can view the problem as a bicriterial optimal control problem.
It is given by

$$\min_{(u,y)} \quad J(u,y) := \begin{pmatrix} \frac{1}{2}\int_0^T \int_\Omega (y(t,x) - y_d(t,x))^2 \, dx \, dt \\ \frac{1}{2}\int_0^T \|u(t)\|_{\mathbb{R}^m}^2 \, dt \end{pmatrix} \tag{BOCP}$$

subject to (s.t.)

$$\begin{aligned} y_t(t,x) - \kappa \Delta y(t,x) + b(x) \cdot \nabla y(t,x) &= \textstyle\sum_{i=1}^m u_i(t)\chi_i(x) && \text{for } (t,x) \in Q := (0,T) \times \Omega \\ \tfrac{\partial y}{\partial \eta}(t,x) + \alpha_i y(t,x) &= \alpha_i y_a(t) && \text{for } (t,x) \in \Sigma_i := (0,T) \times \Gamma_i \\ y(0,x) &= y_0(x) && \text{for } x \in \Omega. \end{aligned} \tag{PPDE}$$

and

$$u_a(t) \le u(t) \le u_b(t) \text{ for almost all } t \in [0,T]. \tag{BC}$$

Equation (PPDE) is supposed to model the temperature distribution in the room, which is given by the set $\Omega \subset \mathbb{R}^n$ with $n = 2$ or $n = 3$. Essentially it is a modified heat equation containing a convection term with convection function b in addition to the diffusion term with diffusion parameter $\kappa > 0$. This models the presumed flow of air in the room, caused for example by an open window or an air conditioner. The right-hand side of the differential equation describes the heating process itself. There are m heaters in the room, whose positions are given by the functions $\chi_1, \dots, \chi_m$, i.e. χ_i is an indicator function of a certain set for all $i \in \{1, \dots, m\}$. The space-independent control functions $u_1, \dots, u_m$ display how much the respective heater is heating in each moment. In order to limit the heating, these functions underlie the bilateral box constraints $u_a(t) \le u(t) \le u_b(t)$ for almost all $t \in [0,T]$, where $u_a(t) \le u_b(t)$ holds for almost all $t \in [0,T]$. As we consider the heating up and not the cooling down of the room, it is reasonable to demand $u_a(t) \ge 0$ for almost all $t \in [0,T]$.
The boundary condition of the PDE is of Robin type and describes the heat exchange with the outside world. Therefore, the boundary of the room $\Gamma := \partial\Omega$ is divided into several disjoint

parts $\Gamma_1, \ldots, \Gamma_r$. For each of these parts we assume a constant isolation coefficient α_i. It displays how good the room is isolated in the respective part or how strong the heat exchange of the room with the outside world is in this part of the wall, respectively. If $\alpha_i = 0$ holds, there is perfect isolation, i.e. no heat exchange with the outside world takes place. By these coefficients it is possible to model windows, outer and inner walls, all of which influence the temperature development in the room differently.
The initial condition y_0 is the initial temperature distribution in the room.

As mentioned before, the goal of our optimal control problem is to find a heating input, such that the resulting temperature distribution is close to a given temperature distribution while the heating costs are supposed to be low as well. This is modelled by the cost function J. The first component measures the difference of the real and the desired temperature distributions y and $y_d \in L^2(Q;\mathbb{R})$, whereas the second component measures the heating costs.
In the first part of this chapter we want to show that this problem fits into the framework of multiobjective optimization presented in Chapter 3, whereas it is illustrated that the POD method can be applied to this problem in the second part.

5.1 Well-Posedness of the Linear Heat Equation with Convection Term

The goal of this section is to show that the heat equation with Robin boundary condition (PPDE) is well-posed. For this purpose we need to specify which conditions the functions and parameters appearing in (PPDE) shall fulfil.

Assumption 5. *We make the following assumptions:*

- $\Omega \subset \mathbb{R}^n$ *is a bounded domain with* C^1*-boundary* $\Gamma := \partial\Omega$*. In our application it is reasonable to assume* $n = 2$ *or* $n = 3$*, but the theoretical considerations in this section also hold for an arbitrary* $n \in \mathbb{N}$ *with* $n \geq 2$*.*
- *The diffusion parameter fulfils* $\kappa > 0$*.*
- *The convection function* b *is supposed to be time-independent and to fulfil* $b \in L^\infty(\Omega;\mathbb{R}^n)$*.*
- *The functions* $\chi_1, \ldots, \chi_m$ *are supposed to be indicator functions. We assume the indicator sets to be disjoint and measurable with non-zero measure so that* $\chi_1, \ldots, \chi_m \in L^2(\Omega)$ *holds.*
- *The isolation coefficients shall fulfil* $\alpha_1, \ldots, \alpha_r \geq 0$*.*
- *The outer temperature* $y_a \in L^2(0,T)$ *is assumed to be space-independent.*
- *The initial temperature distribution shall satisfy* $y_0 \in L^2(\Omega)$*.*
- *For the control function* u *we have* $u \in U := L^2(0,T;\mathbb{R}^m)$*. Note that* U *is a real Hilbert space.*

For the rest of this chapter we assume that Assumption 5 is satisfied.

5.1.1 Weak Formulation

We first need to introduce the weak formulation of the parabolic PDE

$$y_t(t,x) - \kappa \Delta y(t,x) + b(x) \cdot \nabla y(t,x) = \sum_{i=1}^{m} u_i(t)\chi_i(x) \tag{5.1a}$$

$$\frac{\partial y}{\partial \eta}(t,x) + \alpha_i y(t,x) = \alpha_i y_a(t) \tag{5.1b}$$

$$y(0,x) = y_0(x). \tag{5.1c}$$

Multiplying (5.1a) with an arbitrary test function $\varphi \in C^\infty(\Omega)$ and integrating over Ω yield

$$\int_\Omega y_t(t,x)\varphi(x)\,dx - \kappa \int_\Omega \Delta y(t,x)\varphi(x)\,dx + \int_\Omega b(x)\cdot\nabla y(t,x)\varphi(x)\,dx = \int_\Omega \sum_{i=1}^{m} u_i(t)\chi_i(x)\varphi(x)\,dx. \tag{5.2}$$

For the second summand we get by applying Gauss' Theorem and plugging in the Robin boundary condition (5.1b)

$$\begin{aligned}\int_\Omega \Delta y(t,x)\varphi(x)\,dx &= -\int_\Omega \nabla y(t,x)\cdot\nabla\varphi(x)\,dx + \int_{\partial\Omega} \varphi(x)\frac{\partial y}{\partial\eta}(t,x)\,dA(x) \\ &= -\int_\Omega \nabla y(t,x)\cdot\nabla\varphi(x)\,dx + \sum_{i=1}^{r}\alpha_i\int_{\Gamma_i}\varphi(x)(y_a(t)-y(t,x))\,dA(x).\end{aligned} \tag{5.3}$$

Inserting (5.3) into (5.2) yields

$$\begin{aligned}&\int_\Omega y_t(t,x)\varphi(t,x)\,dx + \kappa\int_\Omega \nabla y(t,x)\cdot\nabla\varphi(x)\,dx + \int_\Omega b(x)\cdot\nabla y(t,x)\varphi(t,x)\,dx \\ &+\kappa\sum_{i=1}^{r}\alpha_i\int_{\Gamma_i}\varphi(x)y(t,x)\,dA(x) = \sum_{i=1}^{m}\int_\Omega u_i(t)\chi_i(x)\varphi(x)\,dx + \kappa\sum_{i=1}^{r}\alpha_i\int_{\Gamma_i}\varphi(x)y_a(t)\,dA(x)\end{aligned} \tag{5.4}$$

for an arbitrary $\varphi \in C^\infty(\Omega)$. As $C^\infty(\Omega) \subset H^1(\Omega)$ is dense, this equality also holds for all $\varphi \in H^1(\Omega)$. Hence, by defining $V := H^1(\Omega)$, $H := L^2(\Omega)$ and $W(0,T) := L^2(0,T;V) \cap H^1(0,T;V')$ and introducing the mappings $a : V \times V \to \mathbb{R}$ and $f : (0,T)\times V \to \mathbb{R}$ defined by

$$a(\varphi,\psi) := \kappa\int_\Omega \nabla\varphi(x)\cdot\nabla\psi(x)\,dx + \int_\Omega b(x)\cdot\nabla\varphi(x)\psi(x)\,dx + \kappa\sum_{i=1}^{r}\alpha_i\int_{\Gamma_i}\varphi(x)\psi(x)\,dA(x),$$

and

$$f(t,\varphi) := \sum_{i=1}^{m}\int_\Omega u_i(t)\chi_i(x)\varphi(x)\,dx + \kappa\sum_{i=1}^{r}\alpha_i\int_{\Gamma_i}\varphi(x)y_a(t)\,dA(x),$$

we obtain

$$\begin{aligned} y_t(t) + a(y(t),\cdot) &= f(t,\cdot) \quad \text{in } V' \text{ a.e.} \\ y(0) &= y_0 \quad \text{in } H. \end{aligned} \tag{5.5}$$

with the unknown function $y \in W(0,T)$. We call this the weak formulation of (5.1).

5.1.2 Well-Posedness

The goal of this section is to show that under Assumption 5 the weak formulation (5.5) possesses a unique solution that depends continuously on the data. Therefore, we need to show some results for the operators a and f.

Lemma 5.1. *(i) The mapping $\tilde{a} : V \to V'$, $v \mapsto a(v,\cdot)$ is well-defined and it holds $\tilde{a} \in L(V,V')$.*

(ii) a is coercive with coercivity constants $\alpha = \frac{\kappa}{2}$ and $\beta = \frac{\|b\|^2_{L^\infty(\Omega,\mathbb{R}^n)}}{2\kappa} + \frac{\kappa}{2}$.

Proof. (i) As a is bilinear, the operator $\tilde{a}(\varphi)$ for an arbitrary $\varphi \in V$ is linear as well. To show the well-definedness and $\tilde{a} \in L(V,V')$, we observe that for arbitrary $\varphi, \psi \in V$ we get by using Theorem 2.26

$$\begin{aligned}
|a(\varphi,\psi)| &= \left| \kappa \int_\Omega \nabla\varphi(x)\cdot\nabla\psi(x)\,dx + \int_\Omega b(x)\cdot\nabla\varphi(x)\psi(x)\,dx + \kappa\sum_{i=1}^r \alpha_i \int_{\Gamma_i} \varphi(x)\psi(x)\,dA(x) \right| \\
&\le \kappa \|\nabla\varphi\|_{L^2(\Omega;\mathbb{R}^n)} \|\nabla\psi\|_{L^2(\Omega;\mathbb{R}^n)} + \|b\cdot\nabla\varphi\|_{L^2(\Omega)} \|\psi\|_{L^2(\Omega)} \\
&\quad + \kappa\sum_{i=1}^r \alpha_i \|\varphi\|_{L^2(\partial\Omega)} \|\psi\|_{L^2(\partial\Omega)} \\
&\le \kappa \|\varphi\|_V \|\psi\|_V + \|b\|_{L^\infty(\Omega;\mathbb{R}^n)} \|\nabla\varphi\|_{L^2(\Omega;\mathbb{R}^n)} \|\psi\|_V + \kappa C_1 \sum_{i=1}^r \alpha_i \|\varphi\|_V \|\psi\|_V \\
&\le \kappa \|\varphi\|_V \|\psi\|_V + \|b\|_{L^\infty(\Omega;\mathbb{R}^n)} \|\varphi\|_V \|\psi\|_V + \kappa C_1 \sum_{i=1}^r \alpha_i \|\varphi\|_V \|\psi\|_V \\
&= C \|\varphi\|_V \|\psi\|_V .
\end{aligned}$$

This shows $\tilde{a} \in L(V,V')$ with $\|\tilde{a}\|_{L(V,V')} \le C := \kappa + \|b\|_{L^\infty(\Omega;\mathbb{R}^n)} + \kappa C_1 \sum_{i=1}^r \alpha_i$.

(ii) Let $\varphi \in V$ be arbitrary and define $C_b := \|b\|_{L^\infty(\Omega;\mathbb{R}^n)}$. Then it holds

$$\begin{aligned}
a(\varphi,\varphi) &= \kappa \int_\Omega \nabla\varphi(x)\cdot\nabla\varphi(x)\,dx + \kappa\sum_{i=1}^r \alpha_i \int_{\Gamma_i} \varphi(x)^2\,dA(x) + \int_\Omega b(x)\cdot\nabla\varphi(x)\varphi(x)\,dx \\
&\ge \kappa \|\nabla\varphi\|^2_{L^2(\Omega;\mathbb{R}^n)} - \|b\cdot\nabla\varphi\|_{L^2(\Omega)} \|\varphi\|_{L^2(\Omega)} \\
&\ge \kappa \|\nabla\varphi\|^2_{L^2(\Omega;\mathbb{R}^n)} - C_b \|\nabla\varphi\|_{L^2(\Omega;\mathbb{R}^n)} \|\varphi\|_{L^2(\Omega)} \\
&\ge \kappa \|\nabla\varphi\|^2_{L^2(\Omega;\mathbb{R}^n)} - C_b \left(\frac{\kappa}{2C_b} \|\nabla\varphi\|^2_{L^2(\Omega;\mathbb{R}^n)} + \frac{C_b}{2\kappa} \|\varphi\|^2_{L^2(\Omega)} \right) \\
&= \frac{\kappa}{2} \|\nabla\varphi\|^2_{L^2(\Omega;\mathbb{R}^n)} - \frac{C_b^2}{2\kappa} \|\varphi\|^2_{L^2(\Omega)} \\
&= \frac{\kappa}{2} \|\varphi\|^2_V - \left(\frac{C_b^2}{2\kappa} + \frac{\kappa}{2} \right) \|\varphi\|^2_H .
\end{aligned}$$

Consequently, the claim follows.

□

Lemma 5.2. *The function $\tilde{f} : (0,T) \to V'$, $t \mapsto f(t,\cdot)$ is well-defined and it holds $\tilde{f} \in L^2(0,T;V')$ for all $u \in U$.*

Proof. Let $u \in U$ be arbitrary. It is clear that $\tilde{f}(t) : V \to \mathbb{R}$ is a linear operator for all $t \in (0,T)$. For almost all $t \in (0,T)$ and arbitrary $\varphi \in V$ we get

$$\begin{aligned}
\left|\tilde{f}(t)\varphi\right| &= \left|\sum_{i=1}^{m} \int_\Omega u_i(t)\chi_i(x)\varphi(x)\,dx + \kappa \sum_{i=1}^{r} \alpha_i \int_{\Gamma_i} \varphi(x) y_a(t)\,dA(x)\right| \\
&\le \sum_{i=1}^{m} |u_i(t)| \int_\Omega |\chi_i(x)\varphi(x)|\,dx + \kappa \sum_{i=1}^{r} \alpha_i |y_a(t)| \int_{\partial\Omega} |\varphi(x)|\,dA(x) \\
&\le \sum_{i=1}^{m} |u_i(t)| \,\|\chi_i\|_{L^2(\Omega)} \,\|\varphi\|_{L^2(\Omega)} + \kappa \sum_{i=1}^{r} \alpha_i |y_a(t)| \,\|\varphi\|_{L^2(\partial\Omega)} \sqrt{\lambda(\partial\Omega)} \\
&\le \sum_{i=1}^{m} |u_i(t)| \,\|\chi_i\|_{L^2(\Omega)} \,\|\varphi\|_V + \kappa C_1 \sum_{i=1}^{r} \alpha_i |y_a(t)| \sqrt{\lambda(\partial\Omega)} \,\|\varphi\|_V \\
&\le C(t) \,\|\varphi\|_V .
\end{aligned}$$

By this we get the well-definedness of the mapping $\tilde{f}$ and $\left\|\tilde{f}(t)\right\|_{V'} \le C(t)$ for almost all $t \in (0,T)$. Because of

$$\begin{aligned}
\left\|\tilde{f}\right\|^2_{L^2(0,T;V')} &= \int_0^T \left\|\tilde{f}(t)\right\|^2_{V'} dt \le \int_0^T C(t)^2\,dt \\
&= \int_0^T \left(\sum_{i=1}^{m} |u_i(t)| \,\|\chi_i\|_{L^2(\Omega)} + \kappa C_1 \sum_{i=1}^{r} \alpha_i |y_a(t)| \sqrt{\lambda(\partial\Omega)}\right)^2 dt \\
&\le \int_0^T 2\left(\sum_{i=1}^{m} |u_i(t)| \,\|\chi_i\|_{L^2(\Omega)}\right)^2 + 2\left(\kappa C_1 \sum_{i=1}^{r} \alpha_i |y_a(t)| \sqrt{\lambda(\partial\Omega)}\right)^2 dt \\
&\le 2\,\|\chi\|^2_{L^2(\Omega;\mathbb{R}^m)} \int_0^T \|u(t)\|^2_{\mathbb{R}^m}\,dt + 4\kappa^2 C_1^2 \lambda(\partial\Omega) \sum_{i=1}^{r} \alpha_i^2 \int_0^T |y_a(t)|^2\,dt \\
&= 2\,\|\chi\|^2_{L^2(\Omega;\mathbb{R}^m)} \,\|u\|^2_U + 4\kappa^2 C_1^2 \lambda(\partial\Omega) \sum_{i=1}^{r} \alpha_i^2 \,\|y_a\|^2_{L^2(0,T)} \\
&< \infty
\end{aligned}$$

it follows $\tilde{f} \in L^2(0,T;V')$. □

With these results we can conclude the well-posedness of the weak formulation (5.5).

Theorem 5.3. *Under the given assumptions there is a unique $y \in W(0,T)$ that satisfies (5.5). Furthermore, the estimate*

$$\|y\|_{W(0,T)} \le C\left(\left\|\tilde{f}\right\|_{L^2(0,T;V')} + \|y_0\|_H\right). \tag{5.6}$$

holds.

Proof. This follows directly from Theorem 2.28. □

Definition 5.4. Having the last theorem in mind, we can define the continuous and linear solution operator $T : L^2(0,T;V') \times H \to W(0,T)$ which maps a right-hand side $\tilde{f}$ and an initial value y_0 to the unique solution y of (5.5).

5.1.3 Properties of the Solution Operator

In the following we want to analyse how the solution of (5.5) depends on the control $u \in U$. Therefore, we split the right-hand side function $\tilde{f}$ into $\tilde{f}(t) = r(t) + (\mathcal{B}u)(t)$, where

$$r : (0,T) \to V'$$
$$r(t)\varphi := \kappa \sum_{i=1}^{r} \alpha_i \int_{\Gamma_i} \varphi(x) y_a(t)\, dA(x) \quad (t \in (0,T),\ \varphi \in V)$$

and

$$\mathcal{B} : U \to L^2(0,T;V')$$
$$(\mathcal{B}u)(t)\varphi := \sum_{i=1}^{m} \int_{\Omega} u_i(t)\chi_i(x)\varphi(x)\, dx \quad (t \in (0,T),\ u \in U,\ \varphi \in V).$$

With a similar computation as the one in the proof of Lemma 5.2 it is possible to show that

$$\|\mathcal{B}u\|_{L^2(0,T;V')} \le \|\chi\|_{L^2(\Omega;\mathbb{R}^m)} \|u\|_U \tag{5.7}$$

holds.

Definition and Remark 5.5. We define the inhomogeneous part of the solution $\hat{y} := T(r, y_0)$ and the linear operator $S : U \to W(0,T) \hookrightarrow L^2(0,T;H)$ by $Su := T(\mathcal{B}u, 0)$. By linearity we can conclude that $T(\tilde{f}, y_0) = Su + \hat{y}$ holds for all $u \in U$.

Lemma 5.6. *The operator S is continuous with*

$$\|S\|_{L(U,L^2(0,T;H))} \le C \|\chi\|_{L^2(\Omega;\mathbb{R}^m)},$$

where C is the constant from estimate (5.6).

Proof. Let $u \in U$ be arbitrary. Using the estimates (5.6) and (5.7) we conclude

$$\begin{aligned}\|Su\|_{L^2(0,T;H)} &\le \|Su\|_{W(0,T)} = \|T(\mathcal{B}u, 0)\|_{W(0,T)} \le C \|\mathcal{B}u\|_{L^2(0,T;V')} \\ &\le C \|\chi\|_{L^2(\Omega;\mathbb{R}^m)} \|u\|_U,\end{aligned}$$

and hence the claim follows. □

Lemma 5.7. *The operator S is injective.*

Proof. Let $u \in U \setminus \{0\}$ with $y := Su = 0$ be arbitrary. We need to show that $u = 0$ holds. Using the y is the solution of the PDE (5.5) yields

$$0 = \langle y_t(t), \varphi \rangle_H + a(y(t), \varphi) = \sum_{i=1}^{m} \int_\Omega u_i(t)\chi_i(x)\varphi(x)\,dx \text{ for all } \varphi \in V \text{ for almost all } t \in (0,T).$$

As the indicator functions have non-zero measure by Assumption 5, we can conclude $u = 0$. □

5.2 Reduction of the Optimal Control Problem

Having shown the well-posedness of the PDE in the side condition, we can reduce (BOCP) to a problem that is only depending on the control u by simply defining a new reduced cost function $\hat{J}$.

Definition 5.8. Let $U_{\text{ad}} := \{u \in U \mid u_a(t) \le u(t) \le u_b(t) \text{ for almost all } t \in [0,T]\}$. We define the reduced cost function $\hat{J} : U_{\text{ad}} \to \mathbb{R}^2$ by

$$\hat{J}(u) := J(u, Su + \hat{y}) \quad (u \in U_{\text{ad}}).$$

With this definition the reduced bicriterial optimal control problem reads

$$\min_{u \in U_{\text{ad}}} \hat{J}(u) := \begin{pmatrix} \frac{1}{2}\int_0^T \int_\Omega [(Su)(t,x) + \hat{y}(t,x) - y_d(t,x)]^2 \, dx\, dt \\ \frac{1}{2}\int_0^T \|u(t)\|^2_{\mathbb{R}^m} \, dt \end{pmatrix}. \tag{RBOCP}$$

It is clear that solving (RBOCP) is equivalent to solving the original problem (BOCP).

5.3 The Optimal Control Problem as a Bicriterial Optimization Problem

First we note that the problem formulation (RBOCP) has the form of a general multiobjective optimization problem from Chapter 3 with the feasible set U_{ad} and the multiobjective function

$$f := \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} := \begin{pmatrix} \hat{J}_1 \\ \hat{J}_2 \end{pmatrix}.$$

In this section we show that this problem fulfils Assumptions 1, 2, 3 and 4 from Chapter 3, allowing us to apply the results shown in Chapter 3 to (RBOCP).

Lemma 5.9. *The set U_{ad} is non-empty, convex and closed.*

Proof. As $u_a(t) \le u_b(t)$ holds for almost all $t \in (0,T)$, we get $u_a, u_b \in U_{\text{ad}}$ and therefore U_{ad} is non-empty.

To show that U_{ad} is convex, let $u, v \in U_{\text{ad}}$ and $\lambda \in (0,1)$ be arbitrary. Then it holds for almost all $t \in (0,T)$

$$u_a(t) = \lambda u_a(t) + (1-\lambda)u_a(t) \leq \lambda u(t) + (1-\lambda)v(t) \leq \lambda u_b(t) + (1-\lambda)u_b(t) = u_b(t).$$

Thus, U_{ad} is convex.
To show that U_{ad} is closed, let $(u_n)_{n\in\mathbb{N}} \subset U_{\text{ad}}$ with $u_n \to u$ for $n \to \infty$ in U for a $u \in U$. Then there is a subsequence $(u_{n_j})_{j\in\mathbb{N}}$ such that $u_{n_j} \to u$ pointwise a.e. for $j \to \infty$. As all the functions u_{n_j} fulfil the box constraints (BC) almost everywhere, so does the pointwise limit u and hence $u \in U_{\text{ad}}$, so that U_{ad} is closed. □

Lemma 5.10. *The functions $\hat{J}_1$ and $\hat{J}_2$ are strictly convex.*

Proof. For $\hat{J}_2$ this follows directly from Lemma 2.5. To show the strict convexity of $\hat{J}_1$, we can apply the Lemmata 2.5 and 2.6. □

Lemma 5.11. *The functions $\hat{J}_1$ and $\hat{J}_2$ are twice continuously differentiable. The gradients are given by*

$$\nabla \hat{J}_1(u) = S^*(Su + \hat{y} - y_d) \text{ and}$$
$$\nabla \hat{J}_2(u) = u,$$

where S^* is the Hilbert space adjoint of S. The second derivatives are given by

$$\nabla^2 \hat{J}_1(u) = S^*S \text{ and}$$
$$\nabla^2 \hat{J}_2(u) = id_U.$$

Proof. The claim follows directly by applying the chain rule and the definition of the adjoint operator. □

Corollary 5.12. *The functions $(\hat{J}_1, \hat{J}_2)$ satisfy Assumptions 1, 2, 3 and 4 stated in Chapter 3.*

Proof. It is clear that $\hat{J}_1$ and $\hat{J}_2$ are quadratic. So the only property that has not yet been shown is the positive definiteness of $\nabla^2\hat{J}_1$ and $\nabla^2\hat{J}_2$. Therefore, we observe that

$$\langle \nabla^2 \hat{J}_1(u)h, h\rangle_U = \langle S^*Sh, h\rangle_U = \langle Sh, Sh\rangle_{L^2(0,T;H)} > 0$$

holds for all $u \in U_{\text{ad}}$ and all $h \in U$ since S is injective. Moreover, we obtain

$$\langle \nabla^2 \hat{J}_2(u)h, h\rangle_U = \langle h, h\rangle_U = \|h\|_U^2$$

for all $u \in U_{\text{ad}}$ and all $h \in U$. So $\nabla^2\hat{J}_2$ is even uniformly positive definite with coercivity constant $C_2 = 1$.
Hence, $(\hat{J}_1, \hat{J}_2)$ fulfil Assumptions 1, 2, 3 and 4. □

With Corollary 5.12 we finished the necessary analysis of the problem (RBOCP). We have shown that all the results of Chapter 3 can be applied to the bicriterial cost function $\hat{J}$. The Euclidean reference point problem for a reference point $z \in \mathbb{R}^2$ for this function thus reads

$$\min_{u\in U_{\text{ad}}} F_z(u) := \sum_{i=1}^{2} \frac{1}{2}\left(\hat{J}_i(u) - z_i\right)^2.$$

Exactly as in Chapter 3 we refer to this problem by $(\mathrm{ERPP})_z$.
However, some more work has to be done before being able to implement Algorithm 1. Most approaches to solve scalar optimization problems need evaluations of the gradient and the second derivative of the cost function. The current representations of the gradient and the second derivative of the function $\hat{J}_1$ contain the Hilbert adjoint of the operator S, for which it is not clear yet how it can be evaluated. This problem is addressed in the next section.

5.4 Adjoint Equation

As already mentioned, we deal with the representations of the gradient and the second derivative of the function $\hat{J}_1$ in this section. As seen in Lemma 5.11, the gradient of $\hat{J}_1$ can be written as

$$\nabla \hat{J}_1(u) = S^*(Su + \hat{y} - y_d),$$

where $S^* : L^2(0,T;H) \to U$ is the adjoint operator of S. Thinking of the numerical treatment of this problem for which we finally want to use the methods in Chapter 3, it will be essential to compute $\nabla \hat{J}_1(u)$ and $\nabla^2 \hat{J}_1(u)h$ very quickly for a given control $u \in U_{\mathrm{ad}}$ and direction $h \in U$. However, it is not clear how to evaluate the adjoint S^*. So we are looking for a new representation of the operator S^*S and the vector $S^*(\hat{y} - y_d)$, which only contains numerically evaluateable operators. It will be shown that such a new representation can be found by introducing the so-called adjoint equation.

Definition 5.13. For a PDE of the form (5.5), the function $\hat{J}_1$ and a given control $u \in U_{\mathrm{ad}}$, we call the end value problem

$$\begin{aligned} -p_t(t) + a(\cdot, p(t)) &= y_d(t) - Su(t) - \hat{y}(t) && \text{in } V' \text{ for almost all } t \in (0,T) \\ p(T) &= 0 && \text{in } H. \end{aligned} \tag{5.8}$$

the adjoint equation.

Lemma 5.14. *For each $u \in U_{ad}$ the adjoint equation (5.8) has a unique solution $p \in W(0,T)$ that fulfils*

$$\|p\|_{W(0,T)} \le C \, \|y_d - Su - \hat{y}\|_{L^2(0,T;V')} .$$

Proof. We begin by transforming (5.8) into an initial value problem. Define therefore the function $\tilde{p}(t) := p(T-t)$ for all $t \in [0,T]$. Then p solves (5.8) if and only if $\tilde{p}$ solves

$$\begin{aligned} \tilde{p}_t(t) + a(\cdot, \tilde{p}(t)) &= y_d(T-t) - Su(T-t) - \hat{y}(T-t) && \text{in } V' \text{ for almost all } t \in (0,T) \\ \tilde{p}(0) &= 0 && \text{in } H. \end{aligned} \tag{5.9}$$

We already know that the bilinear form a is coercive. To apply Theorem 2.28, we only need to show, that $y_d(T-\cdot) - Su(T-\cdot) - \hat{y}(T-\cdot) \in L^2(0,T;V')$. But this follows immediatly from $y_d(T-\cdot) - Su(T-\cdot) - \hat{y}(T-\cdot) \in L^2(0,T;V)$. So there is a unique solution $\tilde{p}$ to (5.9) for each control $u \in U_{\mathrm{ad}}$, which fulfils

$$\|\tilde{p}\|_{W(0,T)} \le C \, \|y_d - Su - \hat{y}\|_{L^2(0,T;V')} . \tag{5.10}$$

But then there is a unique solution p of (5.8) which fulfils the same inequality (5.10). □

Definition 5.15. Let the operator $\mathcal{T} : L^2(0,T;V') \to W(0,T)$ denote the linear and continuous solution operator which maps an arbitrary right-hand side f to the solution of (5.8).

Definition and Remark 5.16. Analogously to Definition and Remark 5.5, we split the right-hand side of the adjoint equation in two parts and define $\hat{p} := \mathcal{T}(y_d - \hat{y})$ as well as $\mathcal{A} : U \to W(0,T) \hookrightarrow L^2(0,T;H)$, $\mathcal{A}u := \mathcal{T}(-Su)$. Then $\mathcal{A}$ is linear and continuous as a composition of two continuous functions and it holds by linearity $\mathcal{T}(y_d - Su - \hat{y}) = \mathcal{A}u + \hat{p}$.

The following statements are the reason why we consider the adjoint equation. In [15] they were proved in the case of a symmetric bilinear form a.

Lemma 5.17. *Let $u, v \in U$ be arbitrary. Define $y := Su \in W(0,T)$, $w := Sv \in W(0,T)$ and $p := \mathcal{A}v \in W(0,T)$. Then it holds*

$$\int_0^T \langle (\mathcal{B}u)(t), p(t)\rangle_{V',V}\, dt \;=\; -\int_0^T \langle w(t), y(t)\rangle_H\, dt.$$

Proof. Using that $y = Su$ is the solution of (5.5) with right-hand side $\mathcal{B}u$ and $p(t) \in V$ holds, yields

$$\int_0^T \langle (\mathcal{B}u)(t), p(t)\rangle_{V',V}\, dt = \int_0^T \langle y_t(t), p(t)\rangle_{V',V} + a(y(t), p(t))\, dt.$$

With integration by parts and $p(T) = y(0) = 0$ we get

$$\begin{aligned}\int_0^T \langle y_t(t), p(t)\rangle_{V',V} + a(y(t), p(t))\, dt &= \int_0^T -\langle p_t(t), y(t)\rangle_{V',V} + a(y(t), p(t))\, dt \\ &\quad + \langle p(T), y(T)\rangle_H - \langle p(0), y(0)\rangle_H \\ &= \int_0^T -\langle p_t(t), y(t)\rangle_{V',V} + a(y(t), p(t))\, dt.\end{aligned}$$

As we know that $y(t) \in V$ holds and p solves (5.8) with right-hand side Su, we obtain

$$\int_0^T -\langle p_t(t), y(t)\rangle_{V',V} + a(y(t), p(t))\, dt = -\int_0^T \langle (Sv)(t), y(t)\rangle_H\, dt,$$

which is the claim as $w = Sv$. □

Lemma 5.18. *It holds $\mathcal{B}^*\mathcal{A} = -S^*S\,(\in L(U))$ as well as $\mathcal{B}^*\hat{p} = S^*(y_d - \hat{y})$.*

Proof. Let $u, v \in U$ be arbitrary. Define $y := Su \in W(0,T) \subset L^2(0,T;H)$ and $w := Sv \in W(0,T) \subset L^2(0,T;H)$. Using Lemma 5.17 yields

$$\begin{aligned}\langle S^*Sv, u\rangle_U = \langle Sv, Su\rangle_{L^2(0,T;H)} &= \int_0^T \langle w(t), y(t)\rangle_H\, dt = \int_0^T \langle -(\mathcal{B}u)(t), p(t)\rangle_{V',V}\, dt \\ &= -\langle \mathcal{B}u, p\rangle_{L^2(0,T;V'),L^2(0,T;V)} = -\langle u, \mathcal{B}^*p\rangle_U \\ &= -\langle \mathcal{B}^*\mathcal{A}v, u\rangle_U.\end{aligned}$$

As this holds for all $u, v \in U$, we get the first claim.
Furthermore, it holds for all $u \in U$

$$\begin{aligned}\langle S^*(y_d - \hat{y}), u\rangle_U = \langle y_d - \hat{y}, Su\rangle_{L^2(0,T;H)} &= \int_0^T \langle y_d - \hat{y}, y(t)\rangle_H \, dt \\ &= \int_0^T \langle -\hat{p}_t(t), y(t)\rangle_H + a(y(t), \hat{p}(t)) \, dt \\ &= \int_0^T \langle y_t(t), \hat{p}(t)\rangle_H + a(y(t), \hat{p}(t)) \, dt \\ &= \int_0^T \langle (\mathcal{B}u)(t), \hat{p}(t)\rangle_{V',V} \, dt = \langle \mathcal{B}^*\hat{p}, u\rangle_U\end{aligned}$$

and thus the second claim follows. □

Corollary 5.19. *For the gradient of the function $\hat{J}_1$ we get*

$$\nabla \hat{J}_1(u) = -\mathcal{B}^*(\mathcal{A}u + \hat{p}), \tag{5.11}$$

and for the second derivative

$$\nabla^2 \hat{J}_1(u) = -\mathcal{B}^*\mathcal{A}. \tag{5.12}$$

As mentioned at the beginning of this section, we want to introduce a representation of $\nabla \hat{J}_1$ and $\nabla^2 \hat{J}_1$ that is numerically evaluateable. By introducing the adjoint equation, we ended up with (5.11) and (5.12). It is clear how to compute $\mathcal{A}u$ for a given control u, but we have to deal again with an adjoint operator $\mathcal{B}^*$. The next lemma will show us how to evaluate $\mathcal{B}^*$ in the case that we consider.

Lemma 5.20. *Let the operator $\mathcal{B}$ be given as before. Then the adjoint operator is given by*

$$\begin{aligned}\mathcal{B}^* &: L^2(0,T;V) \to U \\ \mathcal{B}^* v(t) &= \begin{pmatrix} \int_\Omega \chi_1(x) v(t,x) \, dx \\ \vdots \\ \int_\Omega \chi_m(x) v(t,x) \, dx \end{pmatrix}.\end{aligned}$$

Proof. Let $u \in U$ and $v \in L^2(0,T;V)$ be arbitrary. Then it holds

$$\begin{aligned}\int_0^T \langle \mathcal{B}u(t), v(t)\rangle_{V',V} \, dt &= \int_0^T \sum_{i=1}^m \int_\Omega u_i(t)\chi_i(x) v(t,x) \, dx \, dt \\ &= \int_0^T \sum_{i=1}^m u_i(t) \int_\Omega \chi_i(x) v(t,x) \, dx \, dt \\ &= \langle u, \mathcal{B}^* v\rangle_U,\end{aligned}$$

with

$$\mathcal{B}^* v(t) = \begin{pmatrix} \int_\Omega \chi_1(x) v(t,x)\,dx \\ \vdots \\ \int_\Omega \chi_m(x) v(t,x)\,dx \end{pmatrix}.$$

□

5.5 Applying POD to the Bicriterial Optimal Control Problem

In the first part of this chapter we illustrated how the bicriterial optimal control problem (BOCP) governed by the PDE (PPDE) and bilateral constraints on the controls (BC) can be transformed into the problem (RBOCP), which fits into the framework of multiobjective optimization from Chapter 3. However, when applying Algorithm 1 to (RBOCP), many scalar optimization problems $(\text{ERPP})_z$ for different reference points z have to be solved. In the course of this ∇F_z and $\nabla^2 F_z$ have to be evaluated numerous times. In our case evaluating each of the two implies solving both the state and the adjoint equation. To make the computational effort feasible, we apply the POD method from Chapter 4 to both equations in this section. By doing so, we do not only get reduced-order equations, but also the cost function $\hat{J}$ will change to a POD approximated cost function $\hat{J}^\ell$ depending on the POD basis.

5.5.1 POD Approximation for the State Equation

Assume that a POD space $V^\ell \subset V$ of rank ℓ is given and that we have already computed the inhomogeneous part of the solution to the state equation $\hat{y}$. With the same notations as in Section 5.1 we introduce the weak formulation of the homogeneous part of the reduced-order state equation

$$\begin{aligned} y_t^\ell(t) + a(y^\ell(t), \cdot) &= \mathcal{B}u(t) && \text{in } (V^\ell)' \text{ a.e.} \\ \langle y^\ell(0), \varphi \rangle_H &= 0 && \text{for all } \varphi \in (V^\ell)'. \end{aligned} \tag{5.13}$$

Lemma 5.21. *Let Assumption 5 be satisfied. Then there is a unique solution $y^\ell \in H^1(0,T;V)$ of (5.13). Furthermore, the estimate*

$$\left\| y^\ell \right\|_{H^1(0,T;V)} \leq C \left\| \mathcal{B}u \right\|_{L^2(0,T;(V^\ell)'} \tag{5.14}$$

holds, with a constant C independent of ℓ.

Proof. This follows directly from Theorem 4.2. □

Definition and Remark 5.22. As in Definition and Remark 5.5 we define the linear operator $S^\ell : U \to H^1(0,T;V) \hookrightarrow L^2(0,T;H)$, where $S^\ell u$ is the unique solution to (5.13) for any $u \in U$. By linearity we can conclude that $\tilde{y}^\ell := S^\ell u + \hat{y}$ satisfies

$$\begin{aligned} \tilde{y}_t^\ell(t) + a(\tilde{y}^\ell, \cdot) &= r(t) + \mathcal{B}u(t) && \text{in } (V^\ell)' \text{ a.e.} \\ \langle \tilde{y}^\ell(0), \varphi \rangle_H &= \langle y_0, \varphi \rangle_H && \text{for all } \varphi \in (V^\ell)'. \end{aligned} \tag{5.15}$$

Note that $\tilde{y}^\ell$ is not the reduced-order solution of (5.15), as in general $\tilde{y}^\ell(t) \notin V^\ell$. The reason why we still split up the solution like this is that we avoid an approximation error in the inhomogeneous part of the solution.

Lemma 5.23. *The operator S^ℓ is continuous with*

$$\left\|S^\ell\right\|_{L(U,L^2(0,T;H))} \leq C \left\|\chi\right\|_{L^2(\Omega;\mathbb{R}^m)},$$

where C is the constant from estimate (5.14). In particular, C does not depend on ℓ.

Proof. This follows immediately from Lemma 5.21 together with (5.7). □

If we look back at Section 5.2, showing the injectivity of the operator S was crucial to show the strict convexity of the reduced cost function $\hat{J}_1$. Unfortunately, using an arbitrary POD space of rank ℓ does not guarantee the injectivity of the operator S^ℓ without further assumptions.

Assumption 6. *Assume that the POD space V^ℓ is chosen in a way such that $\mathcal{B}$ seen as $\mathcal{B} \in L(U; L^2(0,T;(V^\ell)'))$ is injective.*

For the rest of this chapter let Assumption 6 be satisfied.

Lemma 5.24. *The operator S^ℓ is injective.*

Proof. Let $u \in U$ with $y^\ell := S^\ell u = 0$ be arbitrary. We need to show that $u = 0$ holds. By plugging y into the weak formulation of the reduced-order PDE (5.13), we get that

$$0 = \langle y_t^\ell(t), \varphi\rangle_{(V^\ell)',V^\ell} + a(y^\ell(t),\varphi) = \langle \mathcal{B}u(t), \varphi\rangle_{(V^\ell)',V^\ell} \text{ for all } \varphi \in V^\ell \text{ a.e.},$$

so that $\mathcal{B}u = 0$ in $L^2(0,T;(V^\ell)')$. As we assumed that $\mathcal{B}$ is injective, this implies $u = 0$, so that the operator S^ℓ is injective. □

Theorem 5.25. *Let $u \in U$ be arbitrary and $y := Su$.*

(i) If V^ℓ is an arbitrary POD space and $y^\ell := S^\ell u$ is the solution of (5.13), the estimate

$$\left\|y^\ell - y\right\|^2_{L^2(0,T;V)} \leq C_1 \left\|y - \mathcal{P}^\ell y\right\|^2_{W(0,T)} \tag{5.16}$$

holds.

(ii) If we compute the POD space V^ℓ using the snapshot $y^1 = Su$ and $y^\ell := S^\ell u$ is the solution of (5.13), the estimate

$$\left\|y^\ell - y\right\|^2_{L^2(0,T;V)} \leq C_2 \left(\sum_{i=\ell+1}^{\infty} \bar{\lambda}_i + \left\|y_t - \mathcal{P}^\ell y_t\right\|^2_{L^2(0,T;V')}\right)$$

holds, where $(\bar{\lambda}_i)_{i\in\mathbb{N}}$ are the eigenvalues of the operator $\mathcal{R}$ (see Theorem 4.9).

(iii) If $Su \in H^1(0,T;V)$ holds and the POD space V^ℓ is computed using the snapshots $y^1 = Su$ and $y^2 = (Su)_t$, then it holds for $y^\ell := S^\ell u$

$$\left\|y^\ell - y\right\|^2_{L^2(0,T;V)} \leq C_3 \sum_{i=\ell+1}^{\infty} \bar{\lambda}_i,$$

where $(\bar{\lambda}_i)_{i\in\mathbb{N}}$ are the eigenvalues of the operator $\mathcal{R}$.

(iv) Let $(V^\ell)_{\ell\in\mathbb{N}}$ be the POD spaces generated by an arbitrary family of snapshots. If $Su \in H^1(0,T;V)$ holds and $y^\ell := S^\ell u$ for all $\ell \in \mathbb{N}$, then it holds

$$\left\| y^\ell - y \right\|_{L^2(0,T;V)} \to 0 \quad (\ell \to \infty).$$

Proof. This follows directly from the Theorems 4.4 and 4.11. □

Remark 5.26. As $V \hookrightarrow H$, all the statements also hold, if we consider the error in the $L^2(0,T;H)$ norm instead of the $L^2(0,T;V)$ norm.

5.5.2 POD Approximation for the Cost Function

With the introduction of the solution operator S^ℓ of the reduced-order equation in the last section, it is possible to define a POD approximated cost function and the corresponding Euclidean reference point problem.

Definition 5.27. Assume that $V^\ell \subset V$ is a POD space of rank ℓ. Define the POD approximated cost function by

$$\hat{J}^\ell(u) := \begin{pmatrix} \hat{J}_1^\ell(u) \\ \hat{J}_2(u) \end{pmatrix} := \begin{pmatrix} \frac{1}{2}\left\| S^\ell u + \hat{y} - y_d \right\|^2_{L^2(Q)} \\ \frac{1}{2}\left\| u \right\|^2_U \end{pmatrix}.$$

Furthermore, define the objective feasible region of the function $\hat{J}^\ell$ by $Y^\ell := \hat{J}^\ell(U_{\mathrm{ad}})$.

Notation 5.28. By using the POD approximated cost function, the bicriterial optimal control problem reads

$$\min_{u\in U_{\mathrm{ad}}} \begin{pmatrix} \hat{J}_1^\ell(u) \\ \hat{J}_2(u) \end{pmatrix}. \tag{PODRBOCP}$$

In the following we will refer to (PODRBOCP) by calling it the POD approximated (bicriterial optimal control) problem.
The original bicriterial optimal control problem (RBOCP) is called the full (bicriterial optimal control) problem.

Notation 5.29. The Euclidean reference point problem for a reference point $z \in \mathbb{R}^2$ for the POD approximated cost function is given by

$$\min_{u\in U_{\mathrm{ad}}} F_z^\ell(u) := \frac{1}{2}\left(\hat{J}_1^\ell(u) - z_1\right)^2 + \frac{1}{2}\left(\hat{J}_2(u) - z_2\right)^2.$$

In the following we denote this problem by $(\mathrm{ERPP})_z^\ell$.

Lemma 5.30. *Let Assumption 6 be satisfied. Then $(\hat{J}_1^\ell, \hat{J}_2)$ fulfil the Assumptions 1, 2, 3 and 4, where the derivatives of $\hat{J}_1^\ell$ are given by*

$$\nabla \hat{J}_1^\ell(u) = (S^\ell)^*(S^\ell u + \hat{y} - y_d) \text{ and}$$
$$\nabla^2 \hat{J}_1^\ell(u) = (S^\ell)^* S^\ell.$$

Proof. The proofs are exactly the same as for $(\hat{J}_1, \hat{J}_2)$ by just replacing S with S^ℓ. □

Using the convergence result for the state equation shown in Theorem 5.25, we are able to prove some convergence results for $\hat{J}_1^\ell$.

Theorem 5.31. *Let $(V^\ell)_{\ell\in\mathbb{N}}$ be the POD spaces generated by an arbitrary family of snapshots.*

(i) If $u \in U_{ad}$ with $Su \in H^1(0,T;V)$, then it holds $\hat{J}_1^\ell(u) \to \hat{J}_1(u)$ as $\ell \to \infty$.

(ii) If $(u^\ell)_{\ell\in\mathbb{N}} \subset U_{ad}$ is a sequence with $u^\ell \to u$ $(\ell \to \infty)$ for a $u \in U_{ad}$ with $Su \in H^1(0,T;V)$, then it holds $\left|\hat{J}_1^\ell(u^\ell) - \hat{J}_1(u^\ell)\right| \to 0$ and $\hat{J}_1^\ell(u^\ell) \to \hat{J}_1(u)$ as $\ell \to \infty$.

Proof. Let $u \in U_{\text{ad}}$ be arbitrary with $Su \in H^1(0,T;V)$. Then it holds

$$\begin{aligned}
2\left|\hat{J}_1(u) - \hat{J}_1^\ell(u)\right| &= \left|\langle Su + \hat{y} - y_d, Su + \hat{y} - y_d\rangle_{L^2(0,T;H)} - \langle S^\ell u + \hat{y} - y_d, S^\ell u + \hat{y} - y_d\rangle_{L^2(0,T;H)}\right| \\
&= \left|\langle Su, Su\rangle_{L^2(0,T;H)} - \langle S^\ell u, S^\ell u\rangle_{L^2(0,T;H)} + 2\langle (S - S^\ell)u, \hat{y} - y_d\rangle_{L^2(0,T;H)}\right| \\
&\le \left|\|Su\|^2_{L^2(0,T;H)} - \left\|S^\ell u\right\|^2_{L^2(0,T;H)}\right| + 2\left\|(S - S^\ell)u\right\|_{L^2(0,T;H)} \|\hat{y} - y_d\|_{L^2(0,T;H)}.
\end{aligned}$$

By Theorem 5.25 it holds $2\left\|(S - S^\ell)u\right\|_{L^2(0,T;H)} \|\hat{y} - y_d\|_{L^2(0,T;H)} \to 0$ as $\ell \to \infty$. For the first summand we get that

$$\begin{aligned}
&\left|\|Su\|^2_{L^2(0,T;H)} - \left\|S^\ell u\right\|^2_{L^2(0,T;H)}\right| \\
=&\left|\|Su\|_{L^2(0,T;H)} + \left\|S^\ell u\right\|_{L^2(0,T;H)}\right| \left|\|Su\|_{L^2(0,T;H)} - \left\|S^\ell u\right\|_{L^2(0,T;H)}\right| \\
\le&\left|\|Su\|_{L^2(0,T;H)} + \left\|S^\ell u\right\|_{L^2(0,T;H)}\right| \left\|(S - S^\ell)u\right\|_{L^2(0,T;H)} \\
\to&0 \text{ as } \ell \to \infty
\end{aligned}$$

because $\left\|(S - S^\ell)u\right\|_{L^2(0,T;H)} \to 0$ as $\ell \to \infty$ by Theorem 5.25, which implies additionally that $\left\|S^\ell u\right\|_{L^2(0,T;H)}$ is bounded from above. This shows the first claim.
For the second claim we assume $(u^\ell)_{\ell\in\mathbb{N}} \subset U_{\text{ad}}$ to be a sequence with $u^\ell \to u$ for a $u \in U_{\text{ad}}$ with $Su \in H^1(0,T;V)$. It holds

$$\left|\hat{J}_1(u) - \hat{J}_1^\ell(u^\ell)\right| \le \left|\hat{J}_1(u) - \hat{J}_1(u^\ell)\right| + \left|\hat{J}_1(u^\ell) - \hat{J}_1^\ell(u^\ell)\right|.$$

For the first summand we get due to the continuity of $\hat{J}_1$ that $\left|\hat{J}_1(u) - \hat{J}_1(u^\ell)\right| \to 0$ as $\ell \to \infty$. For the second summand we argue as above and get

$$\begin{aligned}
2\left|\hat{J}_1(u^\ell) - \hat{J}_1^\ell(u^\ell)\right| \le &\left(\left\|Su^\ell\right\|_{L^2(0,T;H)} + \left\|S^\ell u^\ell\right\|_{L^2(0,T;H)}\right) \left\|(S - S^\ell)u^\ell\right\|_{L^2(0,T;H)} \\
&+ 2\left\|(S - S^\ell)u^\ell\right\|_{L^2(0,T;H)} \|\hat{y} - y_d\|_{L^2(0,T;H)}.
\end{aligned}$$

As $(u^\ell)_{\ell\in\mathbb{N}}$ is convergent and hence bounded, we get that $\left\|Su^\ell\right\| + \left\|S^\ell u^\ell\right\|$ is bounded. Additionally it holds

$$\begin{aligned}\left\|(S-S^\ell)u^\ell\right\|_{L^2(0,T;H)} &\leq \left\|S(u^\ell-u)\right\|_{L^2(0,T;H)} + \left\|(S-S^\ell)u\right\|_{L^2(0,T;H)} + \left\|S^\ell(u-u^\ell)\right\|_{L^2(0,T;H)} \\ &\to 0 \text{ as } \ell\to\infty,\end{aligned}$$

where we used again Lemma 5.23 and Theorem 5.25. Now, the second claim follows easily. □

5.5.3 POD Approximation for the Adjoint Equation

As for the full problem, the adjoint operator $(S^\ell)^*$ has to be evaluated to compute the gradient of the POD approximated cost function $\hat{J}_1^\ell$. In this section we show that the reduced-order equation of the adjoint equation (5.8) can be used to get an analogous representation of the gradient as for the full problem.
Assume that a POD space $V^\ell \subset V$ of rank ℓ is given and let $\hat{p}$ be the solution of the inhomogeneous part of the adjoint equation (see Definition and Remark 5.16). Then we introduce the reduced-order adjoint equation by

$$\begin{aligned} -p_t^\ell(t) + a(\cdot, p^\ell(t)) &= -S^\ell u(t) && \text{in } (V^\ell)' \text{ a.e.} \\ \langle p(T), \varphi\rangle_H &= 0 && \text{for all } \varphi \in V^\ell. \end{aligned} \tag{5.17}$$

Lemma 5.32. *Let Assumption 5 be satisfied. Then there is a unique solution $p^\ell \in H^1(0,T;V)$ of (5.17). Furthermore, the estimate*

$$\left\|p^\ell\right\|_{H^1(0,T;V)} \leq C \left\|S^\ell u\right\|_{L^2(0,T;(V^\ell)')} \tag{5.18}$$

holds, with a constant C independent of ℓ.

Proof. This follows from Theorem 4.2 in an analogous way as Lemma 5.14 follows from Theorem 2.28. □

Definition and Remark 5.33. Analogously to Definition and Remark 5.16 we define the operator $\mathcal{A}^\ell : U \to H^1(0,T;V) \hookrightarrow L^2(0,T;H)$, where $\mathcal{A}u$ is the unique solution of (5.17) for any $u \in U$. By linearity we can conclude that $\tilde{p}^\ell := \mathcal{A}^\ell u + \hat{p}$ satisfies

$$\begin{aligned} -\tilde{p}_t^\ell(t) + a(\cdot, \tilde{p}^\ell(t)) &= y_d - S^\ell u(t) - \hat{y} && \text{in } (V^\ell)' \text{ a.e.} \\ \langle p^\ell(T), \varphi\rangle_H &= 0 && \text{for all } \varphi \in V^\ell. \end{aligned} \tag{5.19}$$

As for the state equation, $\tilde{p}^\ell$ is not the reduced-order solution of (5.19) since in general $\tilde{p}^\ell \notin V^\ell$, but this splitting is made to avoid the approximation error for $\hat{p}$.

Lemma 5.34. *The operator $\mathcal{A}^\ell$ is continuous with*

$$\left\|\mathcal{A}^\ell\right\|_{L(U,L^2(0,T;H))} \leq C \left\|\chi\right\|_{L^2(\Omega;\mathbb{R}^m)},$$

with a constant C independent of ℓ.

Proof. This follows immediately from Lemma 5.32 together with Lemma 5.23. □

Using the same arguments as in Section 5.4, we can prove the following statement.

Lemma 5.35. *It holds* $\mathcal{B}^*\mathcal{A}^\ell = -(S^\ell)^* S^\ell \, (\in L(U))$.

Corollary 5.36. *For the gradient of the function* $\hat{J}_1^\ell$ *we get*

$$\nabla \hat{J}_1^\ell(u) = -\mathcal{B}^*(\mathcal{A}^\ell u + \hat{p}), \tag{5.20}$$

and for the second derivative

$$\nabla^2 \hat{J}_1^\ell(u) = -\mathcal{B}^*\mathcal{A}^\ell. \tag{5.21}$$

Theorem 5.37. *Let* $u \in U$ *be arbitrary and* $y := Su$ *as well as* $p := \mathcal{A}u$.

1. *If* V^ℓ *is an arbitrary POD space,* $y^\ell := S^\ell u$ *and* $p^\ell := \mathcal{A}^\ell u$, *the estimate*

$$\left\| p^\ell - p \right\|^2_{L^2(0,T;V)} \leq C_1 \left(\left\| p - \mathcal{P}^\ell p \right\|^2_{W(0,T)} + \left\| y - \mathcal{P}^\ell y \right\|^2_{W(0,T)} \right) \tag{5.22}$$

   holds.

2. *If* $y \in H^1(0,T;V)$ *and we compute the POD space* V^ℓ *using the snapshots* $y^1 = y$ *and* $y^2 = y_t$, *then it holds for* $y^\ell := S^\ell u$ *and* $p^\ell := \mathcal{A}^\ell u$

$$\left\| p^\ell - p \right\|^2_{L^2(0,T;V)} \leq C_2 \left(\left\| p - \mathcal{P}^\ell p \right\|^2_{W(0,T)} + \sum_{i=\ell+1}^{\infty} \bar{\lambda}_i \right),$$

   where $(\bar{\lambda}_i)_{i\in\mathbb{N}}$ *are the eigenvalues of the operator* $\mathcal{R}$ *(see Theorem 4.9).*

3. *If* $y, p \in H^1(0,T;V)$ *and the POD space* V^ℓ *is computed using the snapshots* $y^1 = y$, $y^2 = y_t, y^3 = p$, $y^4 = p_t$, *then it holds for* $p^\ell := \mathcal{A}^\ell u$

$$\left\| p^\ell - p \right\|^2_{L^2(0,T;V)} \leq C_3 \sum_{i=\ell+1}^{\infty} \bar{\lambda}_i,$$

   where $(\bar{\lambda}_i)_{i\in\mathbb{N}}$ *are the eigenvalues of the operator* $\mathcal{R}$.

4. *Let* $(V^\ell)_{\ell\in\mathbb{N}}$ *be the POD spaces generated by an arbitrary family of snapshots. If* $y, p \in H^1(0,T;V)$ *and* $p^\ell := \mathcal{A}^\ell u$, *then it holds that*

$$\left\| p^\ell - p \right\|_{L^2(0,T;V)} \to 0 \quad (\ell \to \infty).$$

Proof. This follows directly from the Theorems 4.4, 4.11 5.25. □

Remark 5.38. Since $V \hookrightarrow H$, all the statements also hold if we consider the error in the $L^2(0,T;H)$ norm instead of the $L^2(0,T;V)$ norm.

With these convergence results we can extend Theorem 5.31.

Theorem 5.39. *Let $(V^\ell)_{\ell\in\mathbb{N}}$ be the POD spaces generated by an arbitrary family of snapshots.*

(i) If $u \in U_{ad}$ with $Su, \mathcal{A}u \in H^1(0,T;V)$, then it holds that $\nabla \hat{J}_1^\ell(u) \to \nabla \hat{J}_1(u)$ as $\ell \to \infty$.

(ii) If $(u^\ell)_{\ell\in\mathbb{N}} \subset U_{ad}$ is a sequence with $u^\ell \to u$ for an $u \in U_{ad}$ with $Su, \mathcal{A}u \in H^1(0,T;V)$, then it holds that $\left\|\nabla \hat{J}_1^\ell(u^\ell) - \nabla \hat{J}_1(u^\ell)\right\|_U \to 0$ and $\nabla \hat{J}_1^\ell(u^\ell) \to \nabla \hat{J}_1(u)$ as $\ell \to \infty$.

Proof. (i) Let $u \in U_{\text{ad}}$ with $Su, \mathcal{A}u \in H^1(0,T;V)$. Then it holds by Theorem 5.37

$$\begin{aligned}\left\|\nabla \hat{J}_1^\ell(u) - \nabla \hat{J}_1(u)\right\|_U &= \left\|\mathcal{B}^*(\mathcal{A}^\ell - \mathcal{A})u\right\|_U \\ &\le \|\mathcal{B}^*\|_{L(L^2(0,T;H),U)} \left\|(\mathcal{A}^\ell - \mathcal{A})u\right\|_{L^2(0,T;H)} \to 0. \end{aligned} \tag{5.23}$$

(ii) Let now $(u^\ell)_{\ell\in\mathbb{N}} \subset U_{\text{ad}}$ be a sequence with $u^\ell \to u$ for an $u \in U_{\text{ad}}$ with $Su, \mathcal{A}u \in H^1(0,T;V)$. Then it holds

$$\begin{aligned}\left\|\nabla \hat{J}_1^\ell(u^\ell) - \nabla \hat{J}_1(u)\right\|_U &= \left\|\nabla \hat{J}_1^\ell(u^\ell) - \nabla \hat{J}_1^\ell(u) + \nabla \hat{J}_1^\ell(u) - \nabla \hat{J}_1(u)\right\|_U \\ &\le \left\|\mathcal{B}^*\mathcal{A}^\ell(u^\ell - u)\right\|_U + \left\|\mathcal{B}^*(\mathcal{A}^\ell - \mathcal{A})u\right\|_U \\ &\le \|\mathcal{B}^*\| \left(\left\|\mathcal{A}^\ell\right\| \left\|u^\ell - u\right\| + \left\|(\mathcal{A}^\ell - \mathcal{A})u\right\|\right) \to 0, \end{aligned} \tag{5.24}$$

where we used (5.23), Theorem 5.37 and the uniform boundedness of $\mathcal{A}^\ell$.
As $\nabla \hat{J}_1$ is continuous, we get $\nabla \hat{J}_1(u^\ell) \to \nabla \hat{J}_1(u)$ and thus

$$\left\|\nabla \hat{J}_1^\ell(u^\ell) - \nabla \hat{J}_1(u^\ell)\right\|_U \le \left\|\nabla \hat{J}_1^\ell(u^\ell) - \nabla \hat{J}_1(u)\right\|_U + \left\|\nabla \hat{J}_1(u^\ell) - \nabla \hat{J}_1(u)\right\|_U \to 0$$

using (5.24).

□

5.5.4 A-Priori Convergence Analysis

For the rest of this chapter let $(V^\ell)_{\ell\in\mathbb{N}}$ be the POD spaces generated by an arbitrary family of snapshots. With the previous results we are able to turn towards the main goal of the POD analysis, namely the convergence of the solution of $(\text{ERPP})_z^\ell$ to the one of $(\text{ERPP})_z$ as $\ell \to \infty$ for a suitable reference point z. In [15, 37] this is already demonstrated for an optimal control problem using the weighted sum method instead of the Euclidean reference point method. In this thesis we adapt the results shown in these two publications to the reference point method.

Lemma 5.40. *Let $z \in \bigcap_{\ell\in\mathbb{N}} \left(P_{Y^\ell} + \mathbb{R}^2_{\le}\right)$ be a reference point and denote by $\bar{u}^\ell$ the minimizer of F_z^ℓ for all $\ell \in \mathbb{N}$. Then $(\bar{u}^\ell)_{\ell\in\mathbb{N}} \subset U_{ad}$ is bounded.*

Proof. Let $u \in U_{\text{ad}}$ be arbitrary. Then the sequence $(F_z^\ell(u))_{\ell\in\mathbb{N}}$ is bounded because of Lemma 5.23. Of course, it holds $F_z^\ell(\bar{u}^\ell) \le F_z^\ell(u)$ for all $\ell \in \mathbb{N}$, from which the claim follows easily. □

Theorem 5.41. *Let $z \in \bigcap_{\ell\in\mathbb{N}} \left(P_{Y^\ell} + \mathbb{R}^2_{\le}\right)$ be a reference point and denote by $\bar{u}^\ell$ the minimizer of F_z^ℓ for all $\ell \in \mathbb{N}$ as well as by $\bar{u}$ the minimizer of F_z. If $S\bar{u}, \mathcal{A}\bar{u} \in H^1(0,T;V)$, it holds*

$$\lim_{\ell\to\infty} \left\|\bar{u}^\ell - \bar{u}\right\|_U = 0.$$

Proof. The first-order conditions

$$\langle\left(\hat{J}_1(\bar{u}) - z_1\right)\nabla\hat{J}_1(\bar{u}) + \left(\hat{J}_2(\bar{u}) - z_2\right)\nabla\hat{J}_2(\bar{u}), u - \bar{u}\rangle_U \geq 0, \tag{5.25}$$

$$\langle\left(\hat{J}_1^\ell(\bar{u}^\ell) - z_1\right)\nabla\hat{J}_1^\ell(\bar{u}^\ell) + \left(\hat{J}_2(\bar{u}^\ell) - z_2\right)\nabla\hat{J}_2(\bar{u}^\ell), u - \bar{u}^\ell\rangle_U \geq 0 \tag{5.26}$$

are fulfilled for all $u \in U_{\text{ad}}$ and all $\ell \in \mathbb{N}$. Setting $u = \bar{u}^\ell$ for an arbitrary $\ell \in \mathbb{N}$ in (5.25) and $u = \bar{u}$ in (5.26) and summing up both inequalities yields

$$\begin{aligned}0 \leq &\left(\hat{J}_1(\bar{u}) - z_1\right)\langle\nabla\hat{J}_1(\bar{u}), \bar{u}^\ell - \bar{u}\rangle_U + \left(\hat{J}_2(\bar{u}) - z_2\right)\langle\nabla\hat{J}_2(\bar{u}), \bar{u}^\ell - \bar{u}\rangle_U\\ &+ \left(\hat{J}_1^\ell(\bar{u}^\ell) - z_1\right)\langle\nabla\hat{J}_1^\ell(\bar{u}^\ell), \bar{u} - \bar{u}^\ell\rangle_U + \left(\hat{J}_2(\bar{u}^\ell) - z_2\right)\langle\nabla\hat{J}_2(\bar{u}^\ell), \bar{u} - \bar{u}^\ell\rangle_U.\end{aligned}$$

Using that the cost functions are quadratic, we get

$$\begin{aligned}0 \leq &\left(\hat{J}_1(\bar{u}) - z_1\right)\left(\hat{J}_1(\bar{u}^\ell) - \hat{J}_1(\bar{u}) - \frac{1}{2}\langle\nabla^2\hat{J}_1(\bar{u})(\bar{u}^\ell - \bar{u}), \bar{u}^\ell - \bar{u}\rangle_U\right)\\ &+ \left(\hat{J}_2(\bar{u}) - z_2\right)\left(\hat{J}_2(\bar{u}^\ell) - \hat{J}_2(\bar{u}) - \frac{1}{2}\langle\nabla^2\hat{J}_2(\bar{u})(\bar{u}^\ell - \bar{u}), \bar{u}^\ell - \bar{u}\rangle_U\right)\\ &+ \left(\hat{J}_1^\ell(\bar{u}^\ell) - z_1\right)\left(\hat{J}_1^\ell(\bar{u}) - \hat{J}_1^\ell(\bar{u}^\ell) - \frac{1}{2}\langle\nabla^2\hat{J}_1^\ell(\bar{u}^\ell)(\bar{u} - \bar{u}^\ell), \bar{u} - \bar{u}^\ell\rangle_U\right)\\ &+ \left(\hat{J}_2(\bar{u}^\ell) - z_2\right)\left(\hat{J}_2(\bar{u}) - \hat{J}_2(\bar{u}^\ell) - \frac{1}{2}\langle\nabla^2\hat{J}_2(\bar{u}^\ell)(\bar{u} - \bar{u}^\ell), \bar{u} - \bar{u}^\ell\rangle_U\right)\\ =&-\frac{1}{2}\left(\hat{J}_1(\bar{u}) - z_1\right)\left\|S(\bar{u}^\ell - \bar{u})\right\|^2_{L^2(0,T;H)} - \frac{1}{2}\left(\hat{J}_1^\ell(\bar{u}^\ell) - z_1\right)\left\|S^\ell(\bar{u}^\ell - \bar{u})\right\|^2_{L^2(0,T;H)}\\ &+ \left(\hat{J}_1(\bar{u}) - z_1\right)\left(\hat{J}_1(\bar{u}^\ell) - \hat{J}_1(\bar{u})\right) + \left(\hat{J}_1^\ell(\bar{u}^\ell) - z_1\right)\left(\hat{J}_1^\ell(\bar{u}) - \hat{J}_1^\ell(\bar{u}^\ell)\right)\\ &- \left(\frac{\hat{J}_2(\bar{u}) + \hat{J}_2(\bar{u}^\ell)}{2} - z_2\right)\left\|\bar{u}^\ell - \bar{u}\right\|^2_U - \left(\hat{J}_2(\bar{u}) - \hat{J}_2(\bar{u}^\ell)\right)^2.\end{aligned} \tag{5.27}$$

Furthermore, a calculation shows

$$\begin{aligned}&\left(\hat{J}_1(\bar{u}) - z_1\right)\left(\hat{J}_1(\bar{u}^\ell) - \hat{J}_1(\bar{u})\right) + \left(\hat{J}_1^\ell(\bar{u}^\ell) - z_1\right)\left(\hat{J}_1^\ell(\bar{u}) - \hat{J}_1^\ell(\bar{u}^\ell)\right)\\ =&-\left(\hat{J}_1^\ell(\bar{u}^\ell) - \hat{J}_1(\bar{u})\right)^2 + \left(\hat{J}_1(\bar{u}) - z_1\right)\left(\hat{J}_1(\bar{u}^\ell) - \hat{J}_1^\ell(\bar{u}^\ell)\right) + \left(\hat{J}_1^\ell(\bar{u}^\ell) - z_1\right)\left(\hat{J}_1^\ell(\bar{u}) - \hat{J}_1(\bar{u})\right).\end{aligned} \tag{5.28}$$

Moreover, using that the cost functions are quadratic again yields

$$\begin{aligned}\hat{J}_1(\bar{u}^\ell) - \hat{J}_1^\ell(\bar{u}^\ell) =&\hat{J}_1(\bar{u}) + \langle\nabla\hat{J}_1(\bar{u}), \bar{u}^\ell - \bar{u}\rangle_U + \frac{1}{2}\langle\nabla^2\hat{J}_1(\bar{u})(\bar{u}^\ell - \bar{u}), \bar{u}^\ell - \bar{u}\rangle_U\\ &- \hat{J}_1^\ell(\bar{u}) - \langle\nabla\hat{J}_1^\ell(\bar{u}), \bar{u}^\ell - \bar{u}\rangle_U - \frac{1}{2}\langle\nabla^2\hat{J}_1^\ell(\bar{u})(\bar{u}^\ell - \bar{u}), \bar{u}^\ell - \bar{u}\rangle_U\\ =&\hat{J}_1(\bar{u}) - \hat{J}_1^\ell(\bar{u}) + \langle\nabla\hat{J}_1(\bar{u}) - \nabla\hat{J}_1^\ell(\bar{u}), \bar{u}^\ell - \bar{u}\rangle_U + \frac{1}{2}\left\|S(\bar{u}^\ell - \bar{u})\right\|^2_{L^2(0,T;H)}\\ &- \frac{1}{2}\left\|S^\ell(\bar{u}^\ell - \bar{u})\right\|^2_{L^2(0,T;H)}.\end{aligned} \tag{5.29}$$

By rearranging (5.27) and (5.28) as well as using (5.29), we end up with

$$\begin{aligned}
&\left\|\hat{J}^\ell(\bar{u}^\ell) - \hat{J}(\bar{u})\right\|_{\mathbb{R}^2}^2 + \left(\frac{\hat{J}_1^\ell(\bar{u}^\ell) + \hat{J}_1(\bar{u})}{2} - z_1\right) \left\|S^\ell(\bar{u}^\ell - \bar{u})\right\|_{L^2(0,T;H)}^2 \\
&\quad + \left(\frac{\hat{J}_2(\bar{u}) + \hat{J}_2(\bar{u}^\ell)}{2} - z_2\right) \left\|\bar{u}^\ell - \bar{u}\right\|_U^2 \\
&\leq \left(\hat{J}_1(\bar{u}) - z_1\right) \left(\hat{J}_1(\bar{u}) - \hat{J}_1^\ell(\bar{u}) + \langle \nabla \hat{J}_1(\bar{u}) - \nabla \hat{J}_1^\ell(\bar{u}), \bar{u}^\ell - \bar{u}\rangle_U\right) \\
&\quad + \left(\hat{J}_1^\ell(\bar{u}^\ell) - z_1\right) \left(\hat{J}_1^\ell(\bar{u}) - \hat{J}_1(\bar{u})\right) \\
&\leq \left(\hat{J}_1(\bar{u}) - z_1\right) \left(\left|\hat{J}_1(\bar{u}) - \hat{J}_1^\ell(\bar{u})\right| + \left\|\nabla \hat{J}_1(\bar{u}) - \nabla \hat{J}_1^\ell(\bar{u})\right\|_U \left\|\bar{u}^\ell - \bar{u}\right\|_U\right) \\
&\quad + \left(\hat{J}_1^\ell(\bar{u}^\ell) - z_1\right) \left|\hat{J}_1^\ell(\bar{u}) - \hat{J}_1(\bar{u})\right|. \qquad (5.30)
\end{aligned}$$

Now the claim follows, as $(\bar{u}^\ell)_{\ell\in\mathbb{N}}$ and $(\hat{J}_1^\ell(\bar{u}^\ell))_{\ell\in\mathbb{N}}$ are bounded by Lemma 5.40, from the Theorems 5.31 and 5.39. □

Corollary 5.42. *Let $z \in \bigcap_{\ell\in\mathbb{N}} \left(P_{Y^\ell} + \mathbb{R}^2_{\leq}\right)$ be a reference point and denote by $\bar{u}^\ell$ the minimizer of F_z^ℓ for every $\ell \in \mathbb{N}$ as well as by $\bar{u}$ the minimizer of F_z. If $S\bar{u}, \mathcal{A}\bar{u} \in H^1(0,T;V)$, it holds*

(i) $\hat{J}_1(\bar{u}^\ell) \to \hat{J}_1(\bar{u})$ *and* $\hat{J}_1^\ell(\bar{u}^\ell) \to \hat{J}_1(\bar{u})$ *as* $\ell \to \infty$.

(ii) $\nabla \hat{J}_1(\bar{u}^\ell) \to \nabla \hat{J}_1(\bar{u})$ *and* $\nabla \hat{J}_1^\ell(\bar{u}^\ell) \to \nabla \hat{J}_1(\bar{u})$ *as* $\ell \to \infty$.

(iii) $\nabla F_z(\bar{u}^\ell) \to \nabla F_z(\bar{u})$ *and* $\nabla F_z^\ell(\bar{u}^\ell) \to \nabla F_z(\bar{u})$ *as* $\ell \to \infty$.

Proof. The corollary is a consequence of the Theorems 5.41, 5.31 and 5.39. □

5.5.5 A-Posteriori Analysis

In the previous section we could show the convergence of the solutions $\bar{u}^\ell$ of $(\mathrm{ERPP})_z^\ell$ to the solution $\bar{u}$ of $(\mathrm{ERPP})_z$ for an increasing dimension of the POD space, but we did not get any a-priori estimates respective the convergence rate or a guaranteed quality of the approximation. Therefore, it is crucial to carry out the a-posteriori analysis. By applying the results from Chapter 3, we are able to show an a-posteriori error estimate for $\left\|\bar{u}^\ell - \bar{u}\right\|_U$ and using the convergence results of the previous section, it is possible to show that the a-posteriori error estimate converges to 0 if we increase the rank of our POD basis. Again, the outline of this section follows [15, 37], where the a-posteriori analysis was done for the weighted sum method.

In particular, we will work with the a-posteriori error estimate which was proved in Theorem 3.51. Note that in general we only know $\hat{J}_2(\bar{u}) - z_2 \geq 0$, where $\bar{u} \in U_{\mathrm{ad}}$ is the minimizer of F_z, if we want to use Theorem 3.51 as an a-posteriori error estimate. Thus, in the present case the estimate reads as follows.

Corollary 5.43. *Let $z \in P_Y + \mathbb{R}^2_{\leq}$ be an arbitrary reference point, $\bar{u} \in U_{ad}$ be the minimizer of F_z and $u_p \in U_{ad}$ be arbitrary. If $\hat{J}(u_p) \geq z$ and $\hat{J}_2(u_p) > z_2$, then it holds*

$$\|\bar{u} - u_p\|_U \leq \left(\frac{\hat{J}_2(u_p) - z_2}{2}\right)^{-1} \|\xi\|_U \quad \text{and}$$

$$\left\|\hat{J}(\bar{u}) - \hat{J}(u_p)\right\|_{\mathbb{R}^2} \leq \frac{1}{2}\left(\frac{\hat{J}_2(u_p) - z_2}{2}\right)^{-\frac{1}{2}} \|\xi\|_U ,$$

where $\xi \in U$ is given such that

$$\langle \nabla F_z(u_p) + \xi, u - u_p \rangle_U \geq 0 \text{ for all } u \in U_{ad} \tag{5.31}$$

is fulfilled.

To use this result as an a-posteriori error estimate, it is clear that we want to apply it to the solution $\bar{u}^\ell$ of $(\mathrm{ERPP})^\ell_z$, i.e. $u_p = \bar{u}^\ell$. The only remaining question is how we can construct $\xi \in U$ such that (5.31) is fulfilled. The canonical choice would be $\xi := -\nabla F_z(u_p)$, but then we cannot expect $\|\xi\|_U$ to be small even if $u_p = \bar{u}$ holds. In the next lemma we will introduce another possible choice of ξ, for which we can expect a smaller norm. The underlying idea of this definition of ξ is for instance explained in [15].

Lemma 5.44. *Let $u_p \in U_{ad}$ be arbitrary. If we define $\xi = \xi(u_p) \in U$ by*

$$\xi_i(t) := \begin{cases} -\min(0, \nabla F_z(u_p)_i(t)) & \text{a.e. in } \{t \in [0,T] \mid (u_p)_i(t) = (u_a)_i(t)\}, \\ \max(0, \nabla F_z(u_p)_i(t)) & \text{a.e. in } \{t \in [0,T] \mid (u_p)_i(t) = (u_b)_i(t)\}, \\ -\nabla F_z(u_p)_i(t) & \text{else}, \end{cases} \tag{5.32}$$

for all $i \in \{1, \dots, m\}$, then (5.31) holds.

Proof. This follows immediately by plugging (5.32) into (5.31). □

With this formula for ξ we can finally show the following result.

Theorem 5.45. *Let $z \in \bigcap_{\ell \in \mathbb{N}} \left(P_{Y^\ell} + \mathbb{R}^2_{\leq}\right)$ be a reference point and denote by $\bar{u}^\ell$ the minimizer of F^ℓ_z for every $\ell \in \mathbb{N}$. Suppose that it holds $S\bar{u}, \mathcal{A}\bar{u} \in H^1(0,T;V)$ for the unique minimizer $\bar{u}$ of F_z. Let $\xi^\ell := \xi(\bar{u}^\ell)$ be chosen as in (5.32). Then it holds $\left\|\xi^\ell\right\|_U \to 0$ as $\ell \to \infty$.*

Proof. By Theorem 5.41 and Corollary 5.42 we can conclude that there exists a subsequence $(\bar{u}^{\ell_k})_{k \in \mathbb{N}}$ such that

$$\lim_{k \to \infty} \bar{u}^{\ell_k}(s) = \bar{u}(s) \tag{5.33}$$

$$\lim_{k \to \infty} \nabla F_z(\bar{u}^{\ell_k})(s) = \nabla F_z(\bar{u})(s) \tag{5.34}$$

$$\lim_{k \to \infty} \nabla F^\ell_z(\bar{u}^{\ell_k})(s) = \nabla F_z(\bar{u})(s) \tag{5.35}$$

for almost all $s \in [0,T]$. For an arbitrary $i \in \{1, \dots, m\}$ we distinguish three cases.

Case 1: $s \in \{t \in [0,T] \mid u_{a_i}(t) < \bar{u}_i(t) < u_{b_i}(t)\}$
This immediately implies $\nabla(F_z(\bar{u}))_i(s) = 0$. Besides, (5.33) implies that there is a $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$ it holds $\bar{u}^{\ell_k}(s) \in (u_{a_i}(s), u_{b_i}(s))$. But then by definition $(\xi^{\ell_k})_i(s) = -(\nabla F_z(\bar{u}^{\ell_k}))_i(s)$ for all $k \geq k_0$. Hence, (5.34) yields

$$(\xi^{\ell_k})_i(s) = -(\nabla F_z(\bar{u}^{\ell_k}))_i(s) \to -(\nabla F_z(\bar{u}))_i(s) = 0.$$

Case 2: $s \in \{t \in [0,T] \mid u_{a_i}(t) = \bar{u}_i(t)\}$
Then it holds $(\nabla F_z(\bar{u}))_i(s) \geq 0$. Assume first that $(\nabla F_z(\bar{u}))_i(s) = 0$. Using again (5.34), we can conclude $(\xi^{\ell_k})_i(s) \to 0$.
So let us assume that $(\nabla F_z(\bar{u}))_i(s) > 0$. Applying (5.35) yields the existence of a $k_0 \in \mathbb{N}$ such that $(\nabla F_z^{\ell}(\bar{u}^{\ell_k}))_i(s) > 0$ for all $k \geq k_0$ holds. But this implies $(\bar{u}^{\ell_k})_i(s) = u_{a_i}(t)$ for all $k \geq k_0$, and hence $(\xi^{\ell_k})_i(s) = -\min(0, (\nabla F_z(\bar{u}^{\ell_k}))_i(s))$ for all $k \geq k_0$. From (5.34) we get that there is a $k_1 \in \mathbb{N}$ such that $(\nabla F_z(\bar{u}^{\ell_k}))_i(s) > 0$ holds for all $k \geq k_1$, and thus $(\xi^{\ell_k})_i(s) = 0$ for all $k \geq \max(k_0, k_1)$.

Case 3: $s \in \{t \in [0,T] \mid u_{b_i}(t) = \bar{u}_i(t)\}$
Analogous to Case 2.

As this holds for all $i \in \{1, \dots, m\}$, we get in total $(\xi^{\ell_k})(s) \to 0$ for almost all $s \in [0,T]$. Lebesgue's dominated convergence theorem now yields $\left\|\xi^{\ell_k}\right\|_U \to 0$. By a standard argument we finally get $\left\|\xi^{\ell}\right\|_U \to 0$. □

Using this result, we can show that the a-posteriori error estimate from Corollary 5.43 also converges to 0 under suitable assumptions. This verifies that it is reasonable to use this a-posteriori error estimate as a measure for the quality of the solutions of $(\mathrm{ERPP})_z^{\ell}$.

Theorem 5.46. *Let $z \in \bigcap_{\ell \in \mathbb{N}} \left(P_{Y^\ell} + \mathbb{R}^2_{\leq}\right)$ be a reference point and denote by $\bar{u}^{\ell}$ the minimizer of F_z^{ℓ} for every $\ell \in \mathbb{N}$. Moreover, let $\bar{u}$ be the minimizer of F_z and suppose that $S\bar{u}, \mathcal{A}\bar{u} \in H^1(0,T;V)$ holds. Let $\xi^{\ell} := \xi(\bar{u}^{\ell})$ be chosen as in (5.32). If $\hat{J}_2(\bar{u}) > z_2$, then it holds*

$$\left\|\bar{u} - \bar{u}^{\ell}\right\|_U \leq \left(\frac{\hat{J}_2(\bar{u}^{\ell}) - z_2}{2}\right)^{-1} \left\|\xi^{\ell}\right\|_U \to 0 \quad (\ell \to \infty),$$

$$\left\|\hat{J}(\bar{u}) - \hat{J}(\bar{u}^{\ell})\right\|_{\mathbb{R}^2} \leq \frac{1}{2}\left(\frac{\hat{J}_2(\bar{u}^{\ell}) - z_2}{2}\right)^{-\frac{1}{2}} \left\|\xi^{\ell}\right\|_U \to 0 \quad (\ell \to \infty).$$

Proof. We already know from Theorem 5.45 that $\left\|\xi^{\ell}\right\|_U \to 0$ as $\ell \to \infty$. By Theorem 5.41 it holds $\left\|\bar{u}^{\ell} - \bar{u}\right\|_U \to 0$ as $\ell \to \infty$ and thus $\hat{J}_2(\bar{u}^{\ell}) \to \hat{J}_2(\bar{u})$ as $\ell \to \infty$. Now the claim follows because $\hat{J}_2(\bar{u}) > z_2$ holds. □

Chapter 6

Numerical Results

In this chapter we investigate the bicriterial optimal control problem which was introduced in the previous chapter numerically by using Algorithm 1. To recall, the problem is given by

$$\min_{(u,y)} \quad J(u,y) = \begin{pmatrix} \frac{1}{2}\int_0^T \int_\Omega \left(y(t,x) - y_d(t,x)\right)^2 \, dx \, dt \\ \frac{1}{2}\int_0^T \|u(t)\|_{\mathbb{R}^m}^2 \, dt \end{pmatrix} \tag{BOCP}$$

subject to (s.t.)

$$\begin{aligned} y_t(t,x) - \kappa \Delta y(t,x) + c_b b(x) \cdot \nabla y(t,x) &= \textstyle\sum_{i=1}^m u_i(t)\chi_i(x) && \text{for } (t,x) \in Q := (0,T) \times \Omega \\ \tfrac{\partial y}{\partial \eta}(t,x) + \alpha_i y(t,x) &= \alpha_i y_a(t) && \text{for } (t,x) \in \Sigma_i := (0,T) \times \Gamma_i \\ y(0,x) &= y_0(x) && \text{for } x \in \Omega. \end{aligned} \tag{PPDE}$$

and

$$u_a(t) \leq u(t) \leq u_b(t) \text{ for almost all } t \in [0,T], \tag{BC}$$

where we use the following parameter and function values in this example:

- The domain Ω is given by the two-dimensional unit square, i.e. $\Omega := (0,1)^2$, and the end time is $T = 1$.
- The diffusion parameter is given by $\kappa = 0.5$.
- For the convection term b we use a stationary solution of an incompressible Navier-Stokes equation with a stream from the left top to the right bottom corner (see Figure 6.1). It holds $\|b\|_{L^\infty(\Omega;\mathbb{R}^2)} \approx 6.6$. The constant $c_b \geq 0$ is used to change the influence of the convection term.
- We impose a floor heating of the whole room with four uniformly distributed heaters in the domains $A_1 = (0, 0.5)^2$, $A_2 = (0, 0.5) \times (0.5, 1)$, $A_3 = (0.5, 1) \times (0, 0.5)$ and $A_4 = (0.5, 1)^2$, i.e. $m = 4$. We refer to the different regions of the heaters by calling the set A_i the region i.
- The box constraints on the control u are given by $u_a(t) = 0$ and $u_b(t) = 3$ for all $t \in [0,1]$. This yields $U_{\text{ad}} = \{u \in L^2(0,T;\mathbb{R}^4) \mid 0 \leq u(t) \leq 3 \text{ for almost all } t \in [0,1]\}$.
- The room is supposed to be perfectly isolated, i.e. it holds $\alpha_1 = 0$ on the whole boundary $\Gamma_1 = \partial\Omega$. This yields a homogeneous Neumann boundary condition.

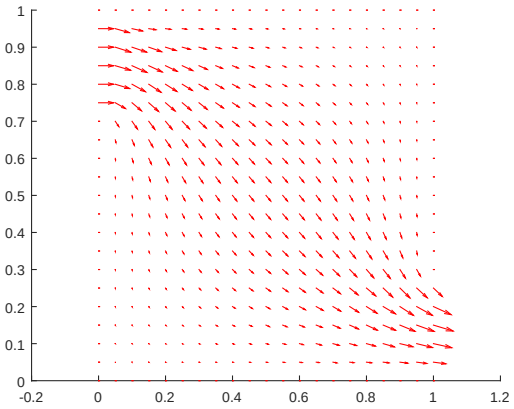


Figure 6.1: Convection in the room

- As an initial condition we suppose that there is a constant temperature of 16° in the whole room, i.e. $y_0(x) = 16$ for all $x \in \Omega$.
- For the desired temperature we want a uniform increase of the temperature from 16° at the starting time until 18° at the end time $T = 1$, i.e. $y_d(t, x) = 16 + 2t$ for all $(t, x) \in Q$.

In the following we will use all the notations which were introduced in the previous chapters.

6.1 Description of the Implementation

In this section we want to provide details about the implementation.

6.1.1 Discretization of the PDEs

The arising state and adjoint equations are discretized using linear finite elements with $N_x = 712$ nodes. The time interval is discretized into $N_t = 100$ equidistantly distributed time instances. For the time integration of the PDEs we use the Crank-Nicolson method.

6.1.2 Solving the Scalar Optimization Problems

The numerical solving of the arising scalar optimization problems $(\text{ERPP})_z$ for different reference points z is done by using a projected Newton-CG method, see for instance [19]. It uses evaluations of ∇F_z and $\nabla^2 F_z$. These can be computed by using the adjoint equation as demonstrated in Section 5.4. Note that the operator $\mathcal{B}^*$ is given by the representation in Lemma 5.20 and can thus also be evaluated by using numerical integration methods.
As described in Section 3.4.1, we can use the a-posteriori error result from Corollary 5.43 to formulate a stopping condition for the optimization routine. If $\bar{u}$ denotes the solution of $(\text{ERPP})_z$

and u_n the current iterative, we stop if

$$\|u_n - \bar{u}\|_U \leq \left(\frac{\hat{J}_2(u_n) - z_2}{2}\right)^{-1} \|\xi(u_n)\|_U \leq \varepsilon_U \text{ and}$$

$$\left\|\hat{J}(u_n) - \hat{J}(\bar{u})\right\|_{\mathbb{R}^2} \leq \frac{1}{2}\left(\frac{\hat{J}_2(u_n) - z_2}{2}\right)^{-\frac{1}{2}} \|\xi(u_n)\|_U \leq \varepsilon_Y$$

are fulfilled. Here, we use the thresholds $\varepsilon_U := 10^{-4}$ for the error bound in the control space and $\varepsilon_Y := 10^{-5}$ for the error bound in the objective space. Note that in the case of solving the full problem $(\text{ERPP})_z$, $\xi(u_n)$ needs to be computed such that

$$\langle \nabla F_z(u_n) + \xi(u_n), u - u_n \rangle_U \geq 0 \text{ for all } u \in U_{\text{ad}}$$

holds, but in the case of solving the POD approximated problem $(\text{ERPP})_z^\ell$, $\xi(u_n)$ needs to fulfil

$$\langle \nabla F_z^\ell(u_n) + \xi(u_n), u - u_n \rangle_U \geq 0 \text{ for all } u \in U_{\text{ad}}.$$

So in case of the full problem $(\text{ERPP})_z$ we can simply use (5.32) and for the POD approximated problem $(\text{ERPP})_z^\ell$ we have to replace ∇F_z by ∇F_z^ℓ in (5.32). In both cases we do not need to solve an additional state or adjoint equation to be able to compute $\xi(u_n)$ because the function values and the gradient at the current iterative are already available.

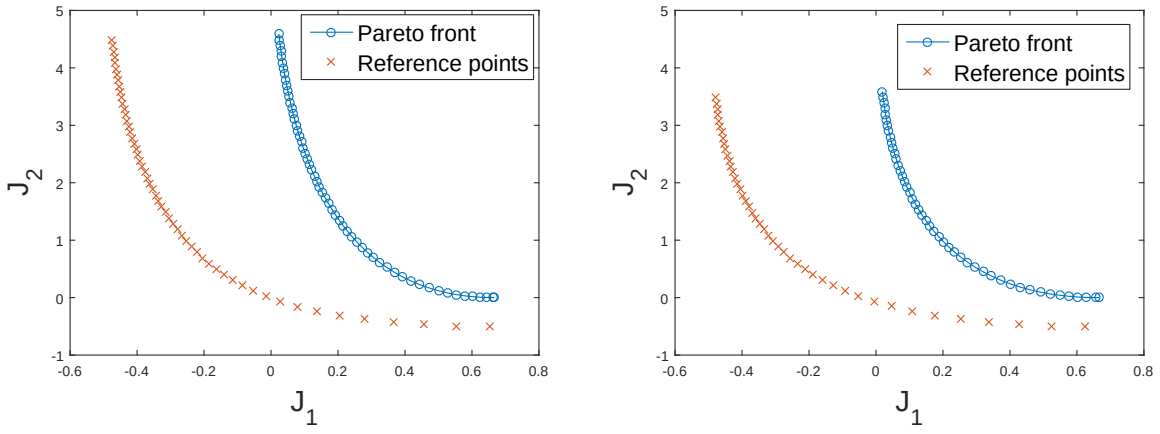
6.1.3 Details for Algorithm 1

We deviate from Algorithm 1 by not using the weight $\alpha = (1, 0)$ but the weight $\alpha = (1, 0.02)$ to compute the first Pareto optimal point. The reason for this is that $\nabla^2 \hat{J}_1$ is only positive definite and not strictly positive definite. This makes the numerical optimization of $\hat{J}_1$ difficult to handle. So instead we regularize the problem by putting a small weight on the cost function $\hat{J}_2$. Even for this optimization problem we observe that the optimization routine does not converge using the projected Newton-CG method, so that we have to use the gradient method instead. On the other hand it is easy to get the minimum of the function $\hat{J}_2$ because $u = 0$ fulfils the box constraints.
For generating the reference points, we use $h_p = 0.5$ and vary h_x in the course of the experiments to see the influence of this parameter on the quality of the approximation.

All computations were carried out on a standard PC, Intel(R) Core(TM)2 Duo CPU P8700 @ 2.53GHz, 4 GB RAM.

6.2 Results for the Full Problem

At first we analyse the solutions of the bicriterial optimal control problem, if we solve the appearing state and adjoint equations with the finite element method.
By running the algorithm, we get a family of solutions of the Euclidean reference point method $(\bar{u}^n, \hat{J}(\bar{u}^n), z^n)_{n=1,\ldots,N_{\text{P}}}$, where $(\bar{u}^n)_{n=1,\ldots,N_{\text{P}}}$ are the optimal controls, $(\hat{J}(\bar{u}^n))_{n=1,\ldots,N_{\text{P}}}$ the corresponding Pareto optimal points in the objective space and $(z^n)_{n=1,\ldots,N_{\text{P}}}$ the utilised reference



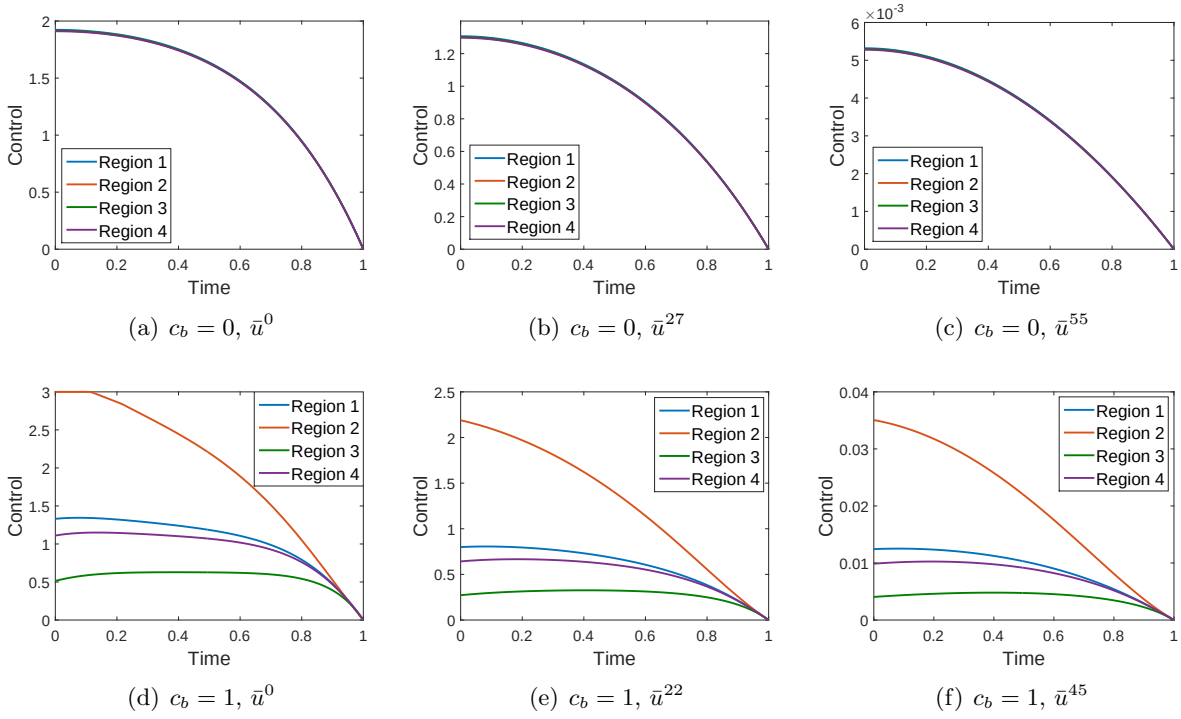
(a) $c_b = 0$ (b) $c_b = 1$

Figure 6.2: Pareto fronts for different values of c_b

points. Consequently, the number N_P denotes the number of Pareto optimal points computed by the Euclidean reference point method. The solutions of the weighted sum method are denoted by $(\bar{u}^0,\ \hat{J}(\bar{u}^0))$ for $\alpha = (1, 0.02)$ and by $(\bar{u}^{N_\mathrm{P}+1},\ \hat{J}(\bar{u}^{N_\mathrm{P}+1}))$ for $\alpha = (0, 1)$. Furthermore, we define $P^n := \hat{J}(\bar{u}^n)$ for all $n \in \{0, \dots, N_\mathrm{P} + 1\}$.

6.2.1 Influence of the Convection Term

In our first experiment we want to analyse the influence of the convection term on the solutions of the bicriterial optimal control problem. For this reason we first run the algorithm with the convection constant $c_b = 0$, i.e. without any convection, and compare the results with the ones we get for $c_b = 1$. For this experiment we use $h_x = 0.1$.
In Figure 6.2 (a) the Pareto front of the problem with $c_b = 0$ can be seen. First of all, we observe that the algorithm provides us with a smooth approximation of the Pareto front, for which 57 Pareto optimal points $P^0, \dots, P^{56}$ have been computed, i.e. $N_\mathrm{P} = 55$.
The range of Pareto optimal points reaches from $P^0 = (0.0231, 4.5894)$ to $P^{56} = (0.6667, 0)$, i.e. it holds $y^\mathrm{id} \approx (0.0231, 0)$ and $y^\mathrm{nad} \approx (4.5894, 0.6667)$. In particular, it holds $\min_{u \in U_\mathrm{ad}} \hat{J}_1(u) \approx 0.0231$, in which the heating costs $\hat{J}_2$ have to amount to 4.5894 to reach this value. Note that there is still a weight of 0.02 on the cost function $\hat{J}_2$ in the computation of the first Pareto optimal point P^0, so that these values are only approximations. However, the slope of the Pareto front at P^0 is -50, so that a small improvement in $\hat{J}_1$ already implies a huge increase of the heating costs $\hat{J}_2$. Therefore, we are satisfied with this approximation for our purposes. We continue the analysis by looking at some optimal controls. The Figures 6.3 (a),(b) and (c) show the optimal controls $\bar{u}^0$, $\bar{u}^{27}$ and $\bar{u}^{55}$. In all controls we can see that all four heaters have the same heating strategy, at least up to computational inaccuracies. This is because of the symmetry of the problem due to constant diffusion in the whole domain and homogeneous Neumann boundary conditions on the entire boundary.
Furthermore, the same qualitative behaviour can be noticed for all three optimal controls and they just differ in their scales. The controls have their maximum in the beginning of the time



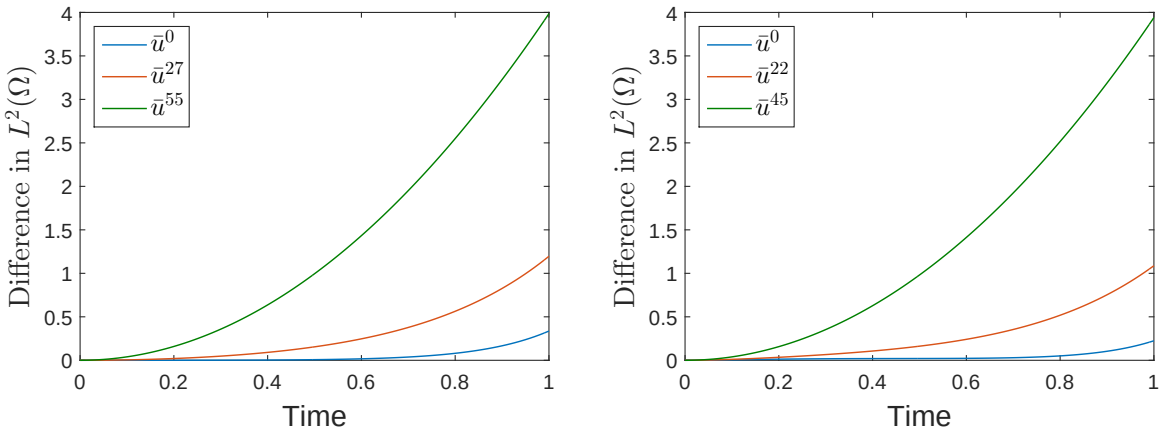
(a) $c_b = 0$, $\bar{u}^0$ (b) $c_b = 0$, $\bar{u}^{27}$ (c) $c_b = 0$, $\bar{u}^{55}$

(d) $c_b = 1$, $\bar{u}^0$ (e) $c_b = 1$, $\bar{u}^{22}$ (f) $c_b = 1$, $\bar{u}^{45}$

Figure 6.3: Optimal controls for different values of c_b

interval and decrease in a parabolic way until they reach zero at $T = 1$. This can be explained by the fact that we want the temperature in the room to follow a given temperature distribution during the whole time interval and to not only coincide with a given temperature distribution at the end time point $T = 1$. Therefore, not heating in the beginning would lead to a deviation, which could only be corrected by a disproportionately high heating input. The effect of the optimal heating strategies can be seen in Figure 6.4 (a), where the L^2-deviation between the temperature and the desired temperature in the whole time interval, i.e. the graph of the mapping $t \mapsto \|S\bar{u}(t,\cdot) + \hat{y}(t,\cdot) - y_d(t,\cdot)\|^2_{L^2(\Omega)}$, is shown. In the beginning the temperature distribution approximates the desired temperature distribution quite accurately for the two optimal controls $\bar{u}^0$ and $\bar{u}^{27}$. Only after some time we can actually see a deviation, which is of course bigger in the case of $\bar{u}^{27}$ than in $\bar{u}^0$. As the heating input decreases to 0 for all three optimal controls, the L^2-deviation is increasing in time.
By looking at Figure 6.5 (a), it can be seen that the achieved temperature is below the desired temperature at $T = 1$ for $\bar{u}^0$. In fact, we observe that at each node and each time instance, the achieved temperature is lower or equal to the desired temperature. As no control is active on the upper boundary in the whole time domain, we can conclude that the cost function $\hat{J}_1$ could be decreased by just heating more, and only the weight of $\alpha_2 = 0.02$ on the heating costs is preventing this from happening in the optimization routine.

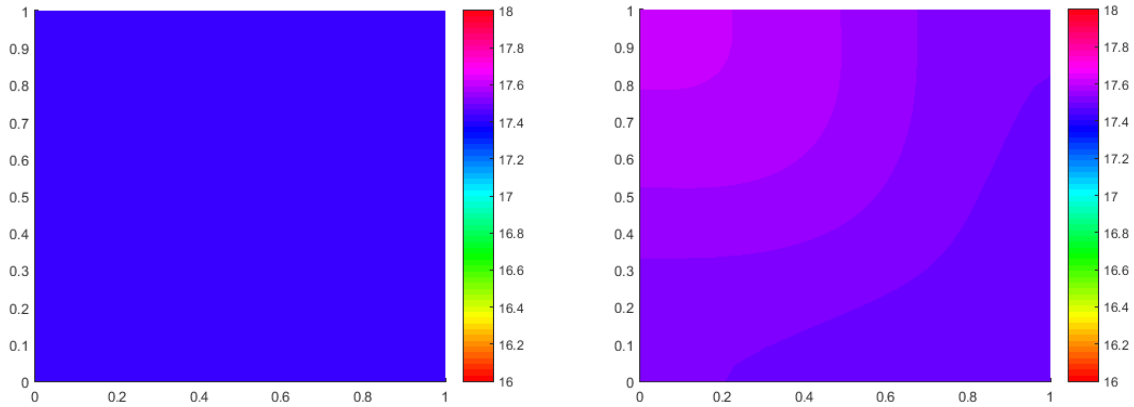
Now, we turn to analysing the results we get for $c_b = 1$ and comparing them with the previous results. Again, we first look at the Pareto front for this case in Figure 6.2 (b). We observe again a smooth approximation of the Pareto front, this time with 47 Pareto optimal points. By



(a) $c_b = 0$ (b) $c_b = 1$

Figure 6.4: Deviation from the desired temperature in $L^2(\Omega)$

looking at $P^0 = (0.0191, 3.5876)$, we find that the desired temperature distribution can actually be reached better with less heating costs than in the case of $c_b = 0$. We will explain this in a moment, when we look at the optimal controls. Of course, the end of the Pareto front is again at $P^{46} = (0.6667, 0)$. Since we start with a constant initial temperature, not controlling just leaves the temperature constant, even if there is an air flow in the room. This leads to the approximate values $\min_{u \in U_{\text{ad}}} \hat{J}_1(u) \approx 0.0191$, $y^{\text{id}} \approx (0.0191, 0)$ and $y^{\text{nad}} \approx (3.5876, 0.6667)$ (again with a weight of 0.02 on the heating costs for computing P^0).
By looking at the optimal controls $\bar{u}^0$, $\bar{u}^{22}$ and $\bar{u}^{45}$ in Figure 6.3 (d),(e) and (f), we can clearly see the influence of the convection term. Figure 6.1 illustrates that the air flow goes from the top left corner of the room to the right bottom corner. Consequently, heater two needs to heat the most in order to reach a uniquely distributed rise in temperature in the whole room because the warm air is transported from the second region into the other ones, mainly region three. This is also the reason why heater three has to heat the least, whereas the heaters one and four show similar heating strategies in between the other two heaters. This behaviour can be observed in all three optimal controls and they just differ from each other in the quantitative strength of heating, but not in the qualitative heating strategy. So this characteristic occurs for both cases $c_b = 0$ and $c_b = 1$.
In the optimal control $\bar{u}^0$ one can see that the control of the second heater is active on the upper bound of the constraints in the beginning of the heating process. The consequence is that the temperature in this region is actually overshooting the desired temperature distribution in the beginning. In the further progress the excessive heat of this region is transported into the other regions by the air flow, so that heaters one, three and four actually have to heat way less than in the case $c_b = 0$. In Figure 6.4 (b) one can see that this strategy leads to a slightly bigger deviation from the desired temperature distribution in the beginning compared to the case $c_b = 0$ because the temperature is too high in region two and too low in regions one, three and four. In the end, however, the strategy of 'overshooting' pays off and the deviation from the desired temperature is smaller than in the case $c_b = 0$, while using less heating input. This leads



(a) $c_b = 0$ (b) $c_b = 1$

Figure 6.5: Temperature at $T = 1$ for the optimal control $\bar{u}^0$ for different values of c_b

to the observed effect that in the problem with convection the desired temperature distribution can be reached slightly better than in the problem with only diffusion, but with way less heating costs.

Yet, we also observe in this case that the desired end temperature is not reached by using any of the optimal control inputs. Figure 6.5 (b) illustrates the temperature at $T = 1$ that is reached by using $\bar{u}^0$. In contrast to the temperature distribution at $T = 1$ in the case $c_b = 0$, the temperature is not homogeneous, but has its maximum in the top left corner and decreases towards the right left corner. Again, this can be explained by the direction of the air flow in the room.

In a last step we want to compare the computation times of both problems. In total, the computation time for solving all Euclidean reference point problems with $c_b = 0$ was 566 s and for the problem with $c_b = 1$ it was 722 s. Additionally, one has to take into consideration that in the case of $c_b = 0$, 55 Euclidean reference point problems have been solved in contrast to 45 Euclidean reference point problems in the case of $c_b = 1$. So on average the computation time for one Euclidean reference point problem is approximately 10 s for $c_b = 0$ and 16 s for $c_b = 1$. This is an expected result because on the one hand, including a convection term adds dynamics to the optimization problem which are more difficult to handle, i.e. more Newton-CG iterations are needed than without the convection term. On the other hand, solving the state and adjoint equation gets more costly due to the fact that the arising linear equation systems become non-symmetric.

Figure 6.6 shows the computational time for each Euclidean reference point problem in both cases. To be able to show both plots on the same scale, although 56 Pareto optimal points are computed in the case of $c_b = 0$ and only 46 Pareto optimal points for $c_b = 1$, we use the relative location on the Pareto front on the x-axis, i.e. $P^0 \widehat{=} 0$ and $P^{N_P+1} \widehat{=} 1$.

While for $c_b = 0$ the plot does not show a strong pattern but only a slight decrease of the computational time with increasing number of the Pareto point, one can clearly see a pattern in the plot for $c_b = 1$. With the exception of two Pareto optimal points, the computation

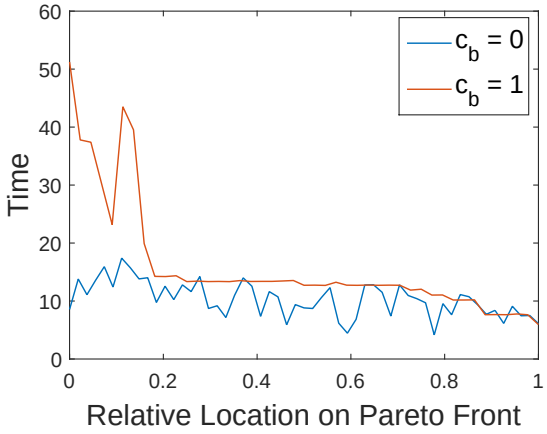


Figure 6.6: Computation times for $c_b = 0$ and $c_b = 1$

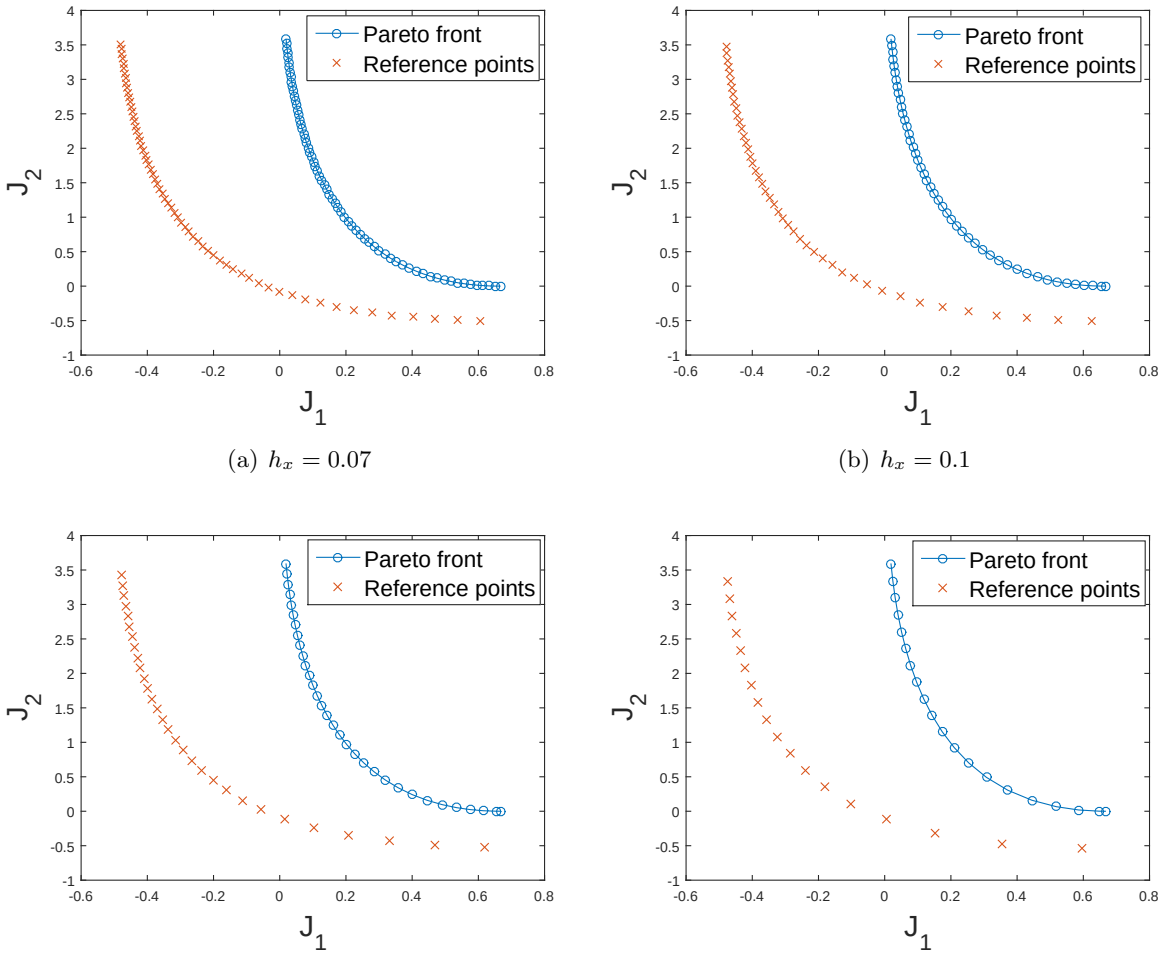
time is monotonically decreasing and there are several steps which correspond to the decrease of needed Newton-CG iterations in the optimization routine. The reason why the number of needed iterations is decreasing while traversing the Pareto front is that the optimization problem gets smoother if the factor $\hat{J}_2(u) - z_2$ increases in comparison to $\hat{J}_1(u) - z_1$ because $\nabla^2 F_z(u)$ is strictly positive definite with coercivity constant $\hat{J}_2(u) - z_2$, if $\hat{J}(u) - z > 0$ holds (see Lemma 3.38).

6.2.2 Controlling the Fineness of the Approximation

In a next experiment we want to investigate whether and how precisely the fineness of the approximation of the Pareto front can be controlled by the parameter h_x which appears in the computation of the reference points (see (3.26)).
In estimate (3.27) we stated that

$$\left\| \hat{J}(\bar{u}^{n+1}) - \hat{J}(\bar{u}^n) \right\|_{\mathbb{R}^2} \leq |h_x|$$

holds for our choice of reference points. However, we could not show a lower bound on the distance $\left\| \hat{J}(\bar{u}^{n+1}) - \hat{J}(\bar{u}^n) \right\|_{\mathbb{R}^2}$. So from a theoretical point of view it is not clear if or to which degree increasing the value of h_x leads to an increasing distance between consecutive Pareto optimal points. To investigate this aspect, we take the previous result for $c_b = 1$ and $h_x = 0.1$ and compare it to the results we get by setting $h_x = 0.07$, $h_x = 0.15$ and $h_x = 0.25$, respectively. Figure 6.7 shows the Pareto fronts for these different values. It can be seen that the approximation gets indeed coarser with increasing h_x. However, no irregular jumps in the approximation of the Pareto front can be seen. This is confirmed by Figure 6.8 (a), where the distances between consecutive Pareto optimal points in the objective space are shown for all the values of h_x. We observe on the one hand that the continuity result (3.27) is verified, i.e. the distances are always smaller than the parameter h_x, and on the other hand that h_x is a close upper bound on the distance between consecutive Pareto optimal points in the parts where the Pareto front is nearly a straight line. In regions in which the Pareto front has a high curvature, the distance of consecutive Pareto optimal points gets automatically smaller without having to

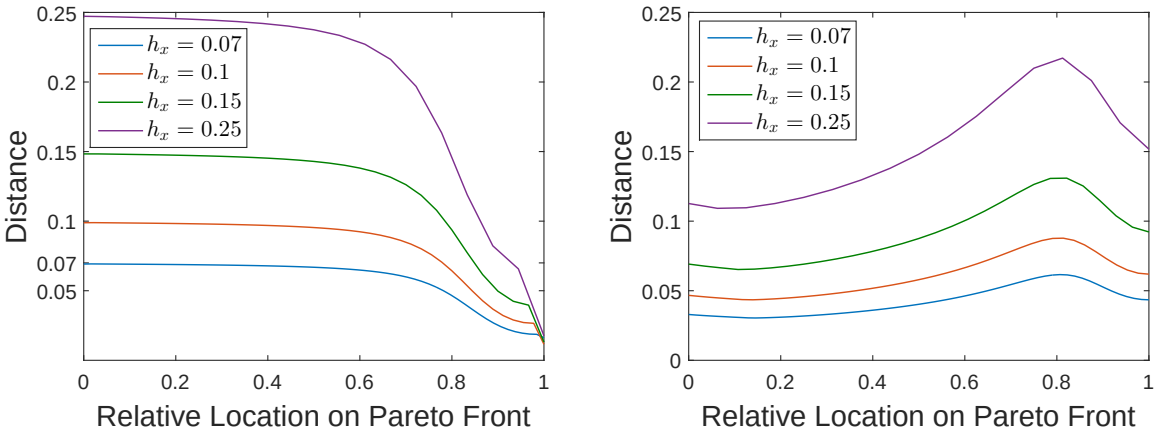


(a) $h_x = 0.07$ (b) $h_x = 0.1$

(c) $h_x = 0.15$ (d) $h_x = 0.25$

Figure 6.7: Pareto fronts for different values of h_x

change the parameter h_x. This is important because these regions need a finer approximation to capture the behaviour of the Pareto front than regions with low curvature.
As a direct consequence the number of computed Pareto optimal points decreases with increasing h_x, which leads to a smaller total computation time (see the second and third column of Table (6.1)). However, the fourth column of Table 6.1 shows that the average computation time per Pareto optimal point increases with increasing h_x. This can be explained by the fact that a bigger distance between Pareto optimal points in the objective space obviously also leads to a bigger distance between the Pareto optimal controls (see Figure 6.3 (b)). As we take the previous optimal control as an initial guess for the computation of the new Pareto optimal control, the algorithm needs more steps to converge in the case of an increased h_x.



(a) Distance $\left\|\hat{J}(\bar{u}^{n+1}) - \hat{J}(\bar{u}^n)\right\|_{\mathbb{R}^2}$ $(n = 1, \ldots, N_{\mathrm{P}})$ (b) Distance $\left\|\bar{u}^{n+1} - \bar{u}^n\right\|_U$ $(n = 1, \ldots, N_{\mathrm{P}})$

Figure 6.8: Distances between consecutive Pareto optimal points for different values of h_x

h_x	Number of Pareto Points	Comp. Time	Avg. Comp. Time per Pareto Point
0.07	66	969,7	15.2
0.1	47	721.8	16.0
0.15	32	678.6	22.6
0.25	20	361.9	20.1
0.5	11	205.0	22.8

Table 6.1: Data for different values of h_x

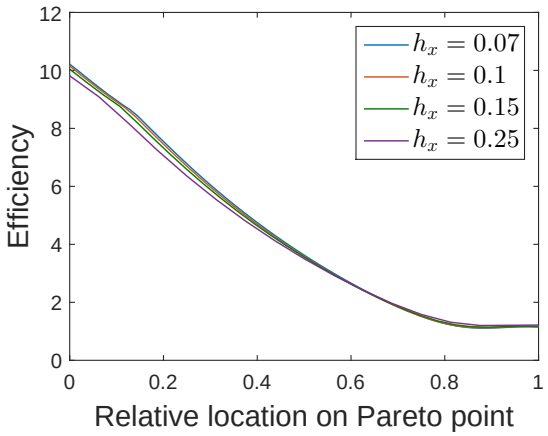
6.2.3 Quality of the Continuity Estimate on the Controls

In this section we want to investigate the quality of the continuity estimate on optimal controls, which was shown in Theorem 3.49. Therefore, we use the data obtained in the previous section for different values of h_x.
To begin with, we recall the result of Theorem 3.49. We showed that

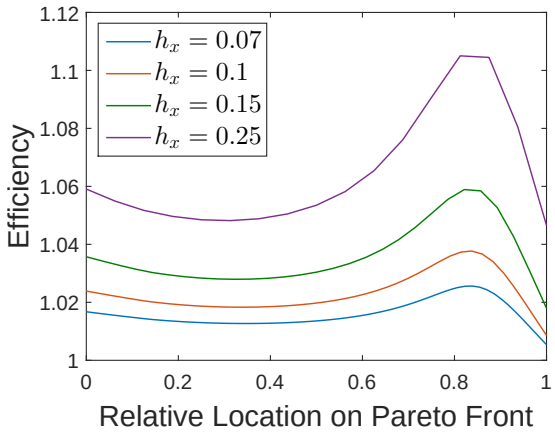
$$\left\|\bar{u}^1 - \bar{u}^2\right\|_U \leq \left[2\left(\hat{J}_2(\bar{u}^1) - z_2^1 + \hat{J}_2(\bar{u}^2) - z_2^2\right)\right]^{-\frac{1}{2}} \left\|z^1 - z^2\right\|_{\mathbb{R}^2}$$

holds in our framework for two arbitrary reference points z^1 and z^2 and the corresponding optimal controls $\bar{u}^1$ and $\bar{u}^2$. In the situation of the algorithm we have the reference points z^n and z^{n+1} as well as the optimal control $\bar{u}^n$ corresponding to z^n, and want to estimate the distance $\left\|\bar{u}^{n+1} - \bar{u}^n\right\|_U$. A priori we thus only know the function value of $\hat{J}_2(\bar{u}^n)$ and that

$$\hat{J}_2(\bar{u}^{n+1}) - z_2^{n+1} \geq \frac{h_p}{\left\|\hat{J}(\bar{u}^n) - z^n\right\|_{\mathbb{R}^2}} (\hat{J}_2(\bar{u}^n) - z_2^n) \tag{6.1}$$



(a) $\eta_{\text{apr}}(h_x, \cdot)$ for different values of h_x



(b) Quotient $\eta_{\text{apo}}(h_x, \cdot)/\eta_{\text{apr}}(h_x, \cdot)$ for different values of h_x

Figure 6.9: Efficiencies for different values of h_x

holds (see Remark 3.54), so that the a-priori estimate obtained from Theorem 3.49 reads

$$\left\|\bar{u}^n - \bar{u}^{n+1}\right\|_U \leq \left[2\left(1 + \frac{h_p}{\left\|\hat{J}(\bar{u}^n) - z^n\right\|_{\mathbb{R}^2}}\right)\left(\hat{J}_2(\bar{u}^n) - z_2^n\right)\right]^{-\frac{1}{2}} \left\|z^{n+1} - z^n\right\|_{\mathbb{R}^2}. \tag{6.2}$$

A posteriori we know of course the difference $\hat{J}_2(\bar{u}^{n+1}) - z_2^{n+1}$, so that

$$\left\|\bar{u}^n - \bar{u}^{n+1}\right\|_U \leq \left[2\left(\hat{J}_2(\bar{u}^n) - z_2^n + \hat{J}_2(\bar{u}^{n+1}) - z_2^{n+1}\right)\right]^{-\frac{1}{2}} \left\|z^{n+1} - z^n\right\|_{\mathbb{R}^2}. \tag{6.3}$$

holds.
In the following we want to investigate and compare the quality of both the a-priori and the a-posteriori continuity estimates (6.2) and (6.3).
Therefore, we introduce a measure for the quality of an estimate, which will be used again later in the context of the POD a-posteriori error estimate.

Definition 6.1. Let $a, b \in \mathbb{R}_{\geq}$ and b be an overestimate for a, i.e. it holds $a \leq b$. Then we define the so-called estimate efficiency η by the formula

$$\eta := \frac{b}{a}. \tag{6.4}$$

It always holds $\eta \in [1, \infty)$, where $\eta = 1$ would be the ideal efficiency.

In our situation we define the two estimate efficiencies $\eta_{\text{apo}}(n)$ and $\eta_{\text{apr}}(n)$. For $\eta_{\text{apo}}(n)$ we set $a := \left\|\bar{u}^n - \bar{u}^{n+1}\right\|_U$ and b as the right-hand side of (6.3), and for $\eta_{\text{apr}}(n)$ we set $a := \left\|\bar{u}^n - \bar{u}^{n+1}\right\|_U$ and b as the right-hand side of (6.2).
Figure 6.9 (a) shows the a-priori estimate efficiencies for different values of h_x. To indicate the dependence on h_x, we write $\eta_{\text{apr}}(h_x, n)$ and $\eta_{\text{apo}}(h_x, n)$ in the following. First of all, we observe

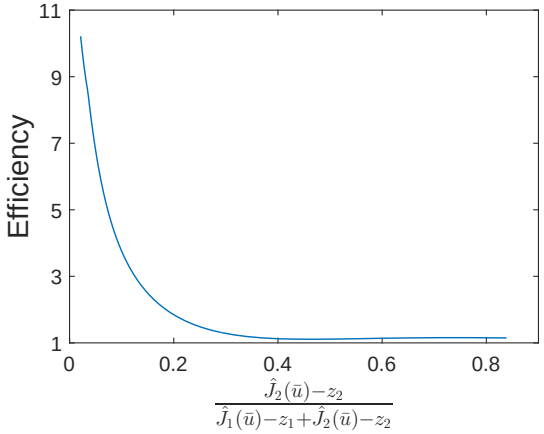


Figure 6.10: Correlation between the share of $\hat{J}_2(\bar{u}) - z_2$ and the efficiency

that $\eta_{\text{apr}}(h_x, n) \geq 1$ holds for all $h_x \in \{0.07, 0.1, 0.15, 0.25\}$ and $n \in \{1, \dots, N_{\text{P}(h_x)}\}$. This verifies the theoretical result of Theorem 3.49. Furthermore, it also shows that the efficiency of the a-priori estimate (6.2) behaves identically for different values of h_x. One can see that the efficiency of the estimate gets better while traversing the Pareto front. While starting with a value of $\eta_{\text{apr}}(h_x, 1) \approx 10$, it decreases to $\eta_{\text{apr}}(h_x, N_{\text{P}(h_x)}) \approx 1$ for all $h_x \in \{0.07, 0.1, 0.15, 0.25\}$. This can be explained by the fact that for an optimal control $\bar{u}$ in the top left region of the Pareto front, it holds $\hat{J}_1(\bar{u}) - z_1 \gg \hat{J}_2(\bar{u}) - z_2$ and in the bottom right region $\hat{J}_1(\bar{u}) - z_1 \ll \hat{J}_2(\bar{u}) - z_2$. When looking at the proof of Theorem 3.49, we could only estimate

$$-\frac{\hat{J}_1(\bar{u}^1) - z_1^1 + \hat{J}_1(\bar{u}^2) - z_1^2}{2} \langle \nabla^2 \hat{J}_1(\bar{u}^1)(\bar{u}^2 - \bar{u}^1), \bar{u}^2 - \bar{u}^1 \rangle_U \leq 0$$

because $\nabla^2 \hat{J}_1$ is not strictly positive definite. Consequently, we lose in some sense the factor $\frac{\hat{J}_1(\bar{u}^1)-z_1^1+\hat{J}_1(\bar{u}^2)-z_1^2}{2} \left\|\bar{u}^1 - \bar{u}^2\right\|_U^2$ in the estimate of Theorem 3.49. Hence, it is reasonable to suspect a connection between the efficiency of the a-priori estimate and the ratio $\frac{\hat{J}_2(\bar{u}^1)-z_2^1}{\hat{J}_1(\bar{u}^1)-z_1^1+\hat{J}_2(\bar{u}^1)-z_2^1}$. Figure 6.10 shows this correlation for the data we get for $h_x = 0.07$. We can clearly see that the efficiency gets rapidly better if the ratio increases from nearly 0 to 0.4. After 0.4, the efficiency does not increase anymore, as it has already reached a value of ≈ 1. This correlation supports the considerations from above.

In Figure 6.9 (b) the quotient between the a-priori and a-posteriori estimate efficiencies for different values of h_x can be seen. It shows that the efficiency of the a-priori estimate is almost as good as the one of the a-posteriori estimate for all values of h_x, i.e. (6.1) is an efficient lower bound. However, the quotient of the efficiency of both estimates gets bigger for increasing h_x, implying that the efficiency of the estimate (6.1) gets slightly worse for increasing h_x.

Additionally, it can be observed that the maximum of the ratio of the two efficiencies is at the same location of the Pareto front as the maximum of the distances of consecutive Pareto optimal controls (see Figure 6.8). This suggests a connection between the efficiency of the estimate (6.1) and the distance of consecutive optimal controls.

6.3 Results for the POD Approximated Problem

In this section we apply the POD method to the bicriterial optimal control problem as shown in Section 5.5 to reduce the computational effort. In the subsequent experiments the quality of the solutions of the POD approximated problem is investigated.

To generate the POD basis we use the snapshots $y^1 := S\bar{u}^0$, $y^2 := \mathcal{A}\bar{u}^0$, where $\bar{u}^0$ is the optimal control corresponding to the first Pareto optimal point P^0 obtained by solving the weighted sum problem $\min_{u\in U_{\text{ad}}} \hat{J}_1(u) + 0.02\hat{J}_2(u)$ with the full-order equations. The computation of the POD basis is done with the method described in Section 4.2.2 using trapezoidal weights for $\alpha_1, \ldots, \alpha_n$ (see (4.11)) and the singular value decomposition to solve the occurring eigenvalue problem (4.9).

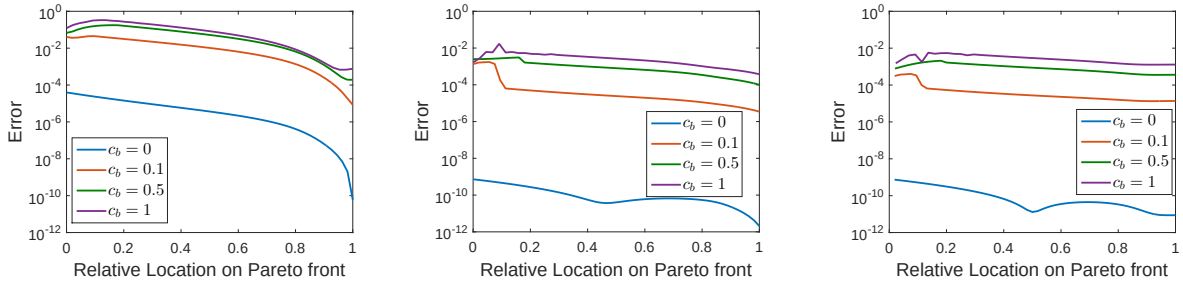
We run the algorithm for the POD approximated problem independently of the computations for the full problem, i.e. the new reference points are generated using the Pareto optimal points obtained from solving $(\text{ERPP})_z^\ell$ and not $(\text{ERPP})_z$. In this way we get two families of solutions $(\bar{u}^n_{\text{POD}}, P^n_{\text{POD}}, z^n_{\text{POD}})_{n=1,\ldots,N_{\text{POD}}}$ and $(\bar{u}^n_{\text{Full}}, P^n_{\text{Full}}, z^n_{\text{Full}})_{n=1,\ldots,N_{\text{Full}}}$. The tuples $(\bar{u}^0, \hat{J}(\bar{u}^0))$ and $(\bar{u}^{N_{\text{P}}+1}, \hat{J}(\bar{u}^{N_{\text{P}}+1}))$ are set as before.
First of all we observe in all our runs that $N_{\text{POD}} = N_{\text{Full}} =: N_{\text{P}}$ holds, i.e. the number of computed Pareto optimal points is always the same. Therefore, we can consider the three errors

$$\|\bar{u}^n_{\text{POD}} - \bar{u}^n_{\text{Full}}\|_U\,,\ \|P^n_{\text{POD}} - P^n_{\text{Full}}\|_{\mathbb{R}^2}\,,\ \|z^n_{\text{POD}} - z^n_{\text{Full}}\|_{\mathbb{R}^2}\,,\ n = 1, \ldots, N_{\text{P}}. \tag{6.5}$$

The focus of our experiments here is to investigate how these errors depend on the strength of the convection term, i.e. in our case on the constant c_b, and on the number of POD basis functions. Additionally, we want to test the quality of the a-posteriori estimate of Theorem 3.51. These results will be used to propose an adaptive POD basis extension strategy, which will be tested and compared to solving the full problem and the POD approximated problem with a fixed number of POD basis functions.

6.3.1 Influence of the Convection Term

To start with, we choose $\ell = 2$ basis functions for the POD basis and run the algorithm for $c_b \in \{0, 0.1, 0.5, 1\}$. In [2] experiments for $c_b = 0$ have been conducted, which show that two POD basis functions already yield good results in this case. Here, we want to see how the errors (6.5) behave if the convection term is added to the problem.
Figure 6.11 shows that there is a big jump in all three errors when we add convection to the problem. Comparing the errors for $c_b = 0$ and $c_b = 0.1$, we observe that the error of the optimal controls increases by a factor of about 10^3 and the errors in the objective space and the reference points by a factor of about 10^6. Understandably, increasing the strength of the convection term also increases the errors (approximately by a factor of 10^2 in all three errors if we compare $c_b = 0.1$ with $c_b = 1$), but the effect of switching on and off the convection has a much stronger influence than increasing its strength. This can be explained by the fact that adding convection to the problem completely changes the dynamics of the state and the adjoint equation, whereas increasing the strength of the convection only intensifies these dynamics. Mathematically, this can be seen by looking at the eigenvalues $\{\bar{\lambda}^n_i\}_{i=1}^m$ of the operator $\mathcal{R}^n$ (see Definition and Theorem 4.15 and Theorem 4.16) for the different values of c_b. By Theorem 4.16 we know that these



(a) Error in control space (b) Error in objective space (c) Error of reference points

Figure 6.11: Errors between the solutions of the POD approximated problem and the full problem for different values of c_b

eigenvalues are a measure for the energy of the snapshot space and that in particular the sum $\sum_{i=\ell+1}^{d} \bar{\lambda}_i^n$ is an inverse measure for the quality of the POD approximation by equation (4.10). Figure 6.12 (a) shows the first 20 eigenvalues of $\mathcal{R}^n$ for the different values of c_b. It can be seen that the eigenvalues for $c_b = 0$ decrease much faster than the ones for $c_b \in \{0.1, 0.5, 1\}$. This confirms our argumentation from above. By looking at Figure 6.12 (b), we can see that value of $\sum_{i=\ell+1}^{d} \bar{\lambda}_i^n$ for $c_b = 0$ is already at 10^{-6} for $\ell = 2$, whereas we need approximately $\ell = 10$ basis functions to reach this value for $c_b = 1$.

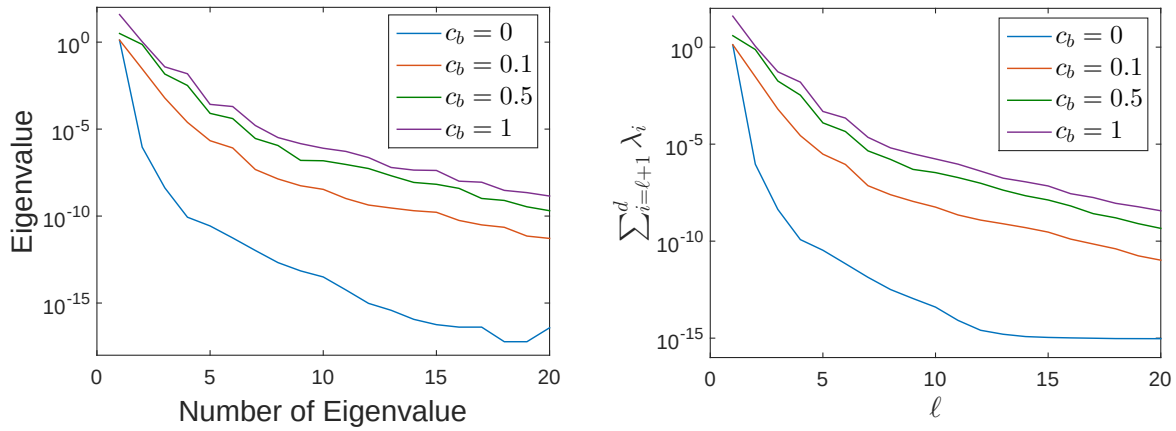
So in a next step we want to test if our theoretical considerations are confirmed and using more basis functions for the POD approximation really reduces the errors (6.5). For this reason we take $c_b = 1$ fixed and run our algorithm with $\ell \in \{2, 5, 10\}$.
By looking at Figure 6.12 (b) again, we see that $\sum_{i=\ell+1}^{d} \bar{\lambda}_i^n \approx 1$ for $\ell = 2$, $\sum_{i=\ell+1}^{d} \bar{\lambda}_i^n \approx 10^{-3}$ for $\ell = 5$ and $\sum_{i=\ell+1}^{d} \bar{\lambda}_i^n \approx 10^{-6}$ for $\ell = 10$.
The resulting errors can be seen in Figure 6.13. As predicted by the theory, all three errors decrease with an increasing number of POD basis functions. Note that there are some Pareto optimal points for which the error in the reference points is actually smaller for $\ell = 5$ than for $\ell = 10$. But as the error in the objective space is bigger for $\ell = 5$ than for $\ell = 10$ on the whole Pareto front, we can conclude that this is only due to a coincidence in the computation of the reference points and not due to a better approximation.

6.3.2 Quality of the POD A-Posteriori Error Estimate

In Theorem 3.51 we showed a result which was used to construct a POD a-posteriori error estimate in Corollary 5.43 for the current bicriterial optimal control problem. It reads in its original form

$$\|\bar{u} - u_p\|_U \leq \left(\frac{\hat{J}_2(\bar{u}) + \hat{J}_2(u_p)}{2} - z_2 \right)^{-1} \|\xi(u_p)\|_U =: \mu_U(\bar{u}, u_p, z), \tag{6.6}$$

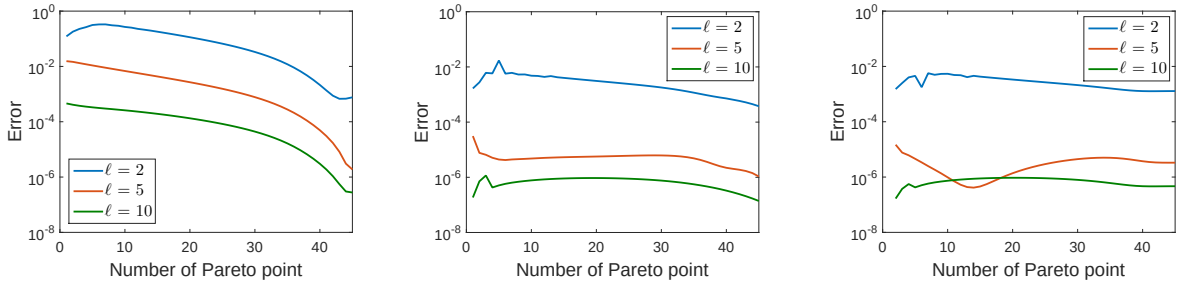
$$\left\| \hat{J}(\bar{u}) - \hat{J}(u_p) \right\|_{\mathbb{R}^2} \leq \frac{1}{2} \left(\frac{\hat{J}_2(\bar{u}) + \hat{J}_2(u_p)}{2} - z_2 \right)^{-\frac{1}{2}} \|\xi(u_p)\|_U =: \mu_Y(\bar{u}, u_p, z), \tag{6.7}$$



(a) Eigenvalues for different values of c_b

(b) Sum of remaining eigenvalues for different values of c_b

Figure 6.12: Eigenvalues values for different values of c_b



(a) Error in control space

(b) Error in objective space

(c) Error of reference points

Figure 6.13: Errors between the solutions of the POD approximated problem and the full problem for $c_b = 1$ and different values of ℓ

where $\xi(u_p)$ is given by (5.32). If we want to use this estimate as an a-posteriori error estimate for the error between the solution $\bar{u}$ of $(\mathrm{ERPP})_z$ and the solution $\bar{u}^\ell$ of $(\mathrm{ERPP})_z^\ell$, it is clear that we have to set $u_p = \bar{u}^\ell$. But then we notice that the estimates μ_U and μ_Y still depend on the function value $\hat{J}_2(\bar{u})$, which is not known if we only solve the POD approximated problem. We will deal with this issue in a moment.

In the following we want to analyse the quality of this a-posteriori error estimate by using the notion of efficiency from Definition 6.1 again. In order to do so, we have to change the structure of our algorithm. So far the POD approximated problem and the full problem are solved independently of each other, which results in two families of solutions $(\bar{u}^n_{\mathrm{POD}}, P^n_{\mathrm{POD}}, z^n_{\mathrm{POD}})_{n=1,\dots,N_{\mathrm{P}}}$ and $(\bar{u}^n_{\mathrm{Full}}, P^n_{\mathrm{Full}}, z^n_{\mathrm{Full}})_{n=1,\dots,N_{\mathrm{P}}}$. In particular, the reference points of both solutions do not coincide, so that the differences $\|\bar{u}^n_{\mathrm{Full}} - \bar{u}^n_{\mathrm{POD}}\|_U$ and $\left\|\hat{J}(\bar{u}^n_{\mathrm{Full}}) - \hat{J}(\bar{u}^n_{\mathrm{POD}})\right\|_{\mathbb{R}^2}$ cannot be estimated by $\mu_U(\bar{u}^n_{\mathrm{Full}}, \bar{u}^n_{\mathrm{POD}}, z^n_{\mathrm{POD}})$ and $\mu_Y(\bar{u}^n_{\mathrm{Full}}, \bar{u}^n_{\mathrm{POD}}, z^n_{\mathrm{POD}})$, but there is an additional error term depend-

ing on $\|z^n_{\text{Full}} - z^n_{\text{POD}}\|_{\mathbb{R}^2}$. Of course, this falsifies the quality of the a-posteriori error estimates. Instead, we first solve the full problem and take the reference points $(z^n_{\text{Full}})_{n=1,\dots,N_{\text{P}}}$ as reference points for the POD approximated problem. This guarantees that we can measure the quality of the a-posteriori error estimates. So in the following we denote by $(z^n)_{n=1,\dots,N_{\text{P}}}$ the reference points used for both problems.

To get rid of the terms depending on the solution of the full problem in (6.6) and (6.7), we can use the estimate $\hat{J}_2(\bar{u}) - z_2 \geq 0$ in the case of $n = 1$ and the estimate (6.1) for $n = 2, \dots, N_{\text{P}}$. This was already used in a similar way to improve the continuity estimate on the controls in Section 6.2.3. Thus, we define new estimates

$$\mu_U(\bar{u}^1_{\text{POD}}, z^1) := \left(\frac{\hat{J}_2(\bar{u}^1_{\text{POD}}) - z^1_2}{2} \right)^{-1} \left\| \xi(\bar{u}^1_{\text{POD}}) \right\|_U \tag{6.8}$$

$$\mu_Y(\bar{u}^1_{\text{POD}}, z^1) := \frac{1}{2} \left(\frac{\hat{J}_2(\bar{u}^1_{\text{POD}}) - z^1_2}{2} \right)^{-\frac{1}{2}} \left\| \xi(\bar{u}^1_{\text{POD}}) \right\|_U \tag{6.9}$$

for $n = 1$ and by using (6.1) and (6.7) we can conclude

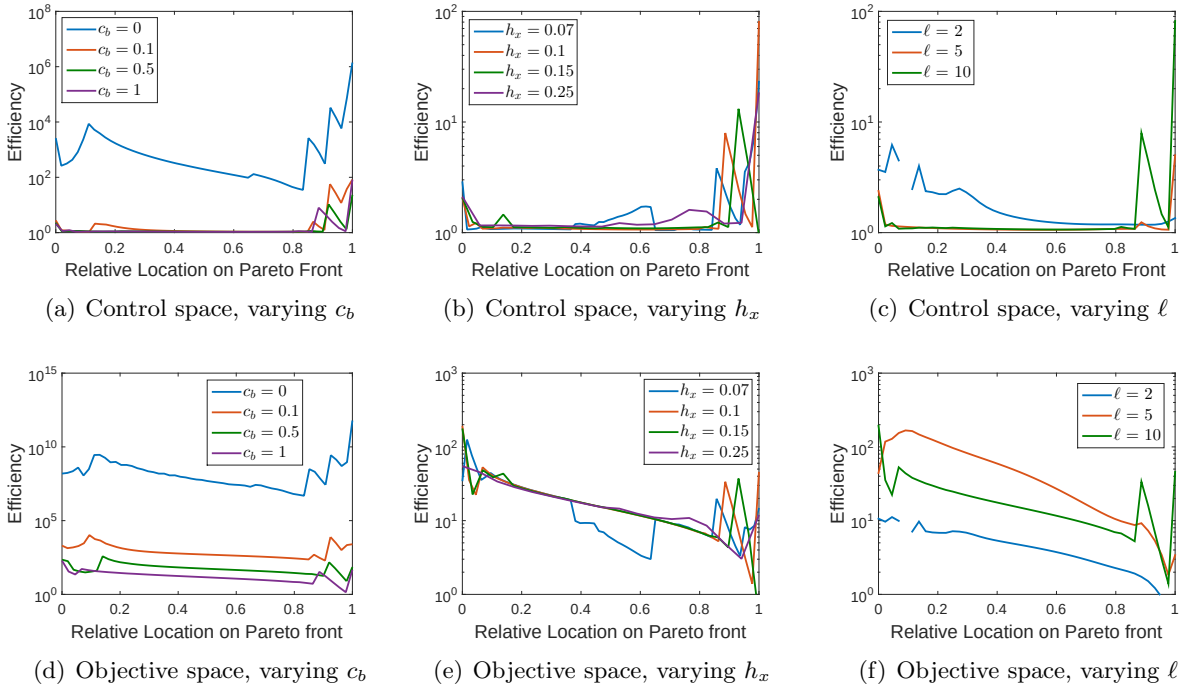
$$\begin{aligned}
\hat{J}_2(\bar{u}^n_{\text{Full}}) - z^n_2 &\geq \frac{h_p}{\left\| \hat{J}(\bar{u}^{n-1}_{\text{Full}}) - z^{n-1} \right\|_{\mathbb{R}^2}} (\hat{J}_2(\bar{u}^{n-1}_{\text{Full}}) - z^{n-1}_2) \\
&\geq \frac{h_p}{\left\| \hat{J}(\bar{u}^{n-1}_{\text{POD}}) - z^{n-1} \right\|_{\mathbb{R}^2} + \mu_Y(\bar{u}^{n-1}_{\text{POD}}, z^{n-1})} (\hat{J}_2(\bar{u}^{n-1}_{\text{POD}}) - \mu_Y(\bar{u}^{n-1}_{\text{POD}}, z^{n-1}) - z^{n-1}_2) \\
&=: \mu_J(\bar{u}^{n-1}_{\text{POD}}, z^{n-1})
\end{aligned}$$

for $i = 2, \dots, N_{\text{P}}$, which leads to the estimates

$$\mu_U(\bar{u}^n_{\text{POD}}, z^n) := \left(\frac{\mu_J(\bar{u}^{n-1}_{\text{POD}}, z^{n-1}) + \hat{J}_2(\bar{u}^n_{\text{POD}})}{2} - z^n_2 \right)^{-1} \|\xi(\bar{u}^n_{\text{POD}})\|_U \tag{6.10}$$

$$\mu_Y(\bar{u}^n_{\text{POD}}, z^n) := \frac{1}{2} \left(\frac{\mu_J(\bar{u}^{n-1}_{\text{POD}}, z^{n-1}) + \hat{J}_2(\bar{u}^n_{\text{POD}})}{2} - z^n_2 \right)^{-\frac{1}{2}} \|\xi(\bar{u}^n_{\text{POD}})\|_U \tag{6.11}$$

for $i = 2, \dots, N_{\text{P}}$.
For these estimates we measure the efficiency for varying parameters in the sense of Definition 6.1. To measure the efficiency of the a-posteriori error estimate in the control space, we set $a := \|\bar{u}^n_{\text{Full}} - \bar{u}^n_{\text{POD}}\|_U$ and $b := \mu_U(\bar{u}^n_{\text{POD}}, z^n)$. For the efficiency of the a-posteriori error estimate in the objective space we set analogously $a := \left\| \hat{J}(\bar{u}^n_{\text{Full}}) - \hat{J}(\bar{u}^n_{\text{POD}}) \right\|_{\mathbb{R}^2}$ and $b := \mu_Y(\bar{u}^n_{\text{POD}}, z^n)$. In a first test we set $h_x = 0.1$, use $\ell = 10$ basis functions for the POD approximation and compare the efficiency for varying convection constants $c_b \in \{0, 0.1, 0.5, 1\}$. The results for the efficiency of (6.10) can be seen in Figure 6.14 (a) and for the efficiency of (6.11) in Figure 6.14 (d). Comparing both figures, it can be seen that the efficiency of the a-posteriori error estimate in the control space is better than the one in the objective space by a factor of $10^2 - 10^4$. Both figures have in common that there is a big jump in the efficiency between the problem with



(a) Control space, varying c_b (b) Control space, varying h_x (c) Control space, varying ℓ

(d) Objective space, varying c_b (e) Objective space, varying h_x (f) Objective space, varying ℓ

Figure 6.14: Efficiencies of the POD a-posteriori error estimates

and without convection - namely the efficiency of both a-posteriori error estimates is worse by a factor of $10^3 - 10^4$ for the problem without convection than for the problem with $c_b = 0.1$. In the plot of the efficiencies of the estimates in the objective space in Figure 6.14 (d) one can actually see that the efficiency gets even better with increasing c_b, whereas the efficiency of the a-posteriori error estimates in the control space are close to 1 for all $c_b \in \{0.1, 0.5, 1\}$ in large parts of the Pareto front.

In a second test we fix $c_b = 1$, $\ell = 10$ and look at the results for varying $h_x \in \{0.07, 0.1, 0.15, 0.25\}$. The Figures 6.14 (b) and (e) show the results in this case. In the control space the efficiency is close to 1 for all values of h_x in the main part of the Pareto front. Again, it can be observed that the efficiency of the estimates in the objective space is worse in comparison. Moreover, the efficiency of the estimate for $h_x = 0.07$ is much better than the estimates for $h_x \in \{0.1, 0.15, 0.25\}$ in the middle of the Pareto front. So far, we were not able to find an explanation for this.

Lastly, we set $c_b = 1$, $h_x = 0.1$ and compare the results for varying $\ell \in \{2, 5, 10\}$ in Figures 6.14 (c) and (f). First of all, there is a gap in the efficiency of the a-posteriori error estimates for $\ell = 2$. This is due to the fact that in this case, the algorithm did not converge properly in the beginning of the Pareto front. As a consequence, it holds $\hat{J}_2(\bar{u}^5_{\text{POD}}) < z^5_2$, so that the a-posteriori error estimates could not be computed in this case. Apart from that we see in the control space that the efficiency for $\ell = 2$ is worse than for $\ell = 5$ and $\ell = 10$, which are again close to 1. In the objective space, however, we can observe a reversed behaviour, namely that the estimate for $\ell = 2$ is more efficient than the ones for $\ell = 5$ and $\ell = 10$. Nonetheless, the efficiencies of the estimates in the objective space are again worse than of the estimates in the control space, exactly as in the two other parameter settings.

Note that there are Pareto optimal points in the end of the Pareto front for $h_x = 0.15$ (Figure 6.14 (e)) and for $\ell = 2$ (Figure 6.14 (f)), for which the efficiency is actually below 1. This is due to the inaccuracy in the computation of the optimal controls of the full problem, for which the stopping threshold $\varepsilon_U = 10^{-4}$ was used. So by not reaching the optimal control of the full problem exactly, we get an additional error between the solution of the full and the POD approximated problem, which cannot be captured by the a-posteriori error estimate. This can lead to a situation in which the real error is actually bigger than the a-posteriori error estimate. All three parameter settings have in common that a zigzag behaviour of the efficiencies can be observed at the end of the Pareto front. This can be explained by the fact that the real errors $\|\bar{u}^n_{\text{POD}} - \bar{u}^n_{\text{Full}}\|_U$ and $\left\|\hat{J}(\bar{u}^n_{\text{POD}}) - \hat{J}(\bar{u}^n_{\text{POD}})\right\|_{\mathbb{R}^2}$ decrease drastically in the end of the Pareto front. This decrease cannot be totally captured by the a-posteriori error estimates (6.10) and (6.11).
As a conclusion we can say that the efficiency of the a-posteriori error estimate in the control space is remarkably good for all parameters except $c_b = 0$ and $\ell = 2$. This observation is used in the next section to propose a basis extension algorithm using the a-posteriori error estimate.

6.3.3 Adaptive POD Basis Extension

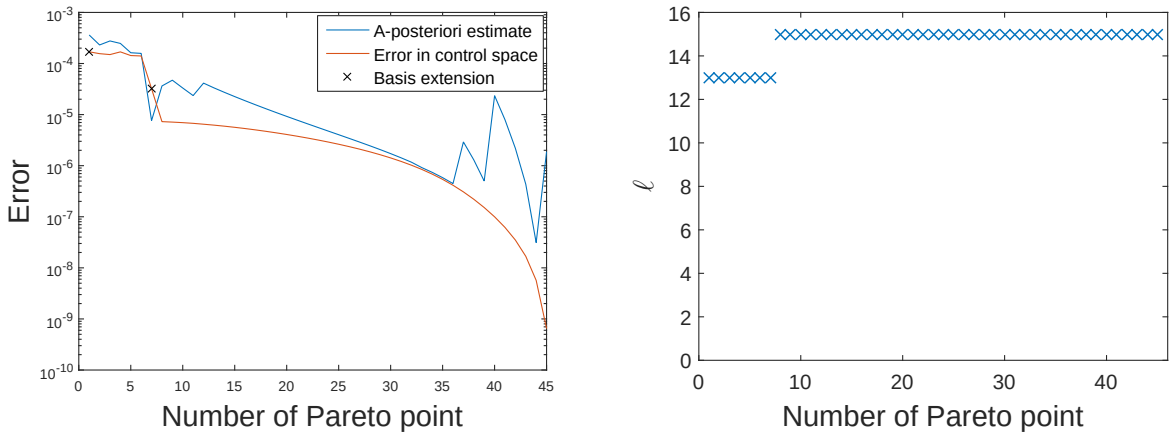
The insights from the previous sections motivate an adaptive strategy to choose the number of POD basis functions for a problem with convection. On the one hand, we saw in Section 6.3.1 that increasing the number of POD basis functions decreases the error between the solutions of the POD approximated problem and of the full problem. On the other hand, the a-posteriori error estimate for the error in the control space has a good efficiency, if $c_b > 0$. Therefore, we propose the following basis extension algorithm. It shows the optimization routine for the n-th Pareto optimal point.

Algorithm 2: POD Basis Extension

Data: μ: Threshold for a-posteriori error estimate

begin
- Set check = 0;
- **while** *check = 0* **do**
  - Solve $(\text{ERPP})^{\ell}_{z^n}$;
  - Compute $\mu_U(\bar{u}^n_{\text{POD}}, z^n)$;
  - **if** $\mu_U(\bar{u}^n_{POD}, z^n) > \mu$ **then**
    - Set $\ell = \ell + 1$;
  - **else**
    - Set check = 1;

To put the algorithm into words, we choose an initial number of POD basis functions ℓ_0 in the beginning of the algorithm and increase this number every time the a-posteriori error estimate is larger than a certain threshold μ. In this way we ensure that the error in the controls never rises above this threshold.
To test this adaptive strategy, we set $c_b = 1$, $h_x = 0.1$, $\ell_0 = 2$ and as a threshold $\mu = 4 \cdot 10^{-4}$. Note that it does not make sense to set $\mu < \varepsilon_U$. Recall that ε_U was the stopping condition for the distance of the current suboptimal control to the optimal control of the optimization



(a) Error in the control space

(b) Number of used POD basis functions

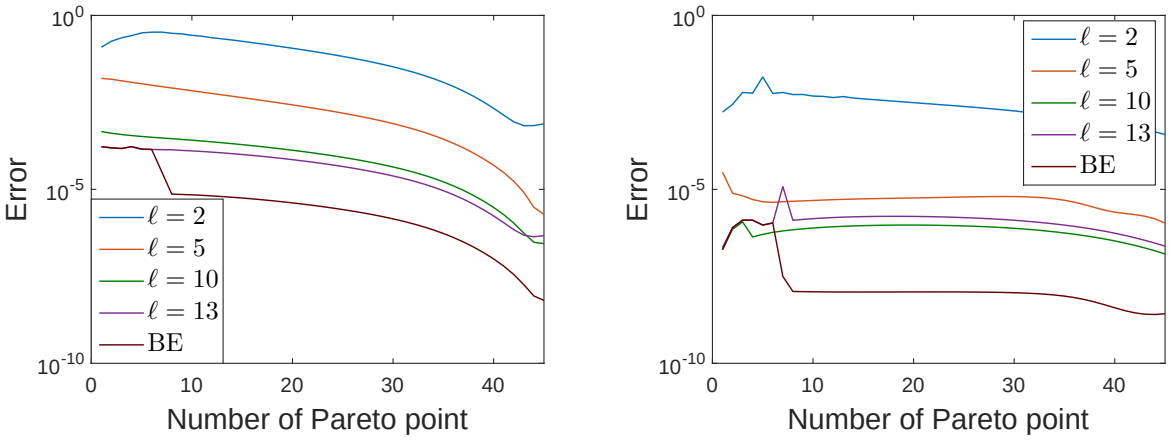
Figure 6.15: Results for the POD basis extension algorithm

problem $(\text{ERPP})_z^\ell$. So if we set $\mu < \varepsilon_U$, we would only ensure that the computed control is closer than ε_U to the optimal control of $(\text{ERPP})_z^\ell$, while expecting that the suboptimal control is closer than μ to the optimal control of $(\text{ERPP})_z$. This would not be reasonable, if $\mu < \varepsilon_U$, independent of the quality of the POD approximation.

Figure 6.15 (a) shows that the a-posteriori error estimate and consequently the real error in the control space adapt nicely to the set threshold μ. Note that there is a point for which the a-posteriori error estimate is smaller than the actual error in the controls. This can be explained by the same argument as in the end of Section 6.3.2, where we argued why the efficiency of the a-posteriori error estimate can be below 1. Note that the real error can only overshoot the a-posteriori error estimate by the value of ε_U.

The instances in which the POD basis has been extended are marked in Figure 6.15 (a) as well. Here, the basis had to be extended from two to 13 basis functions at the first Pareto optimal point and by additional two basis functions at the seventh Pareto optimal point to reach the desired threshold (see Figure 6.15 (b)). So in the end a total of 15 basis functions are used.

Naturally, we want to compare these results with the ones that we get by setting a fixed number of POD basis functions. For this experiment we use again the setting in which the reference points of the POD approximated problem and the full problem are computed separately and thus differ from each other. As the number of POD basis functions is extended to 13 in the computation of the first Pareto optimal point, the results for the POD basis extension algorithm should be better than the ones for $\ell = 2$, $\ell = 5$ and $\ell = 10$ from Section 6.3.1. Indeed, Figure 6.16 shows that both the error in the control space and the error in the objective space are the lowest for the basis extension algorithm. As an additional comparison Figure 6.16 also shows the errors for a fixed number of $\ell = 13$ POD basis functions to show the influence of the basis extension to 15 POD basis functions at the seventh Pareto optimal point. We see that the basis extension was not necessary to stay under the threshold μ, however, it yields much better results in both the control and the objective space.



(a) Error in the control space

(b) Error in the objective space

Figure 6.16: Comparison of the results for the POD basis extension algorithm and for a fixed number of POD basis functions

	$c_b = 0$	$c_b = 0.1$	$c_b = 0.5$	$c_b = 1$
Full Problem	566.3 s	678,8 s	729.2 s	701.5 s
$\ell = 2$	23.6 s	69.6 s	103.8 s	164.2 s
$\ell = 5$	27.3 s	20.5 s	27.9 s	39.4 s
$\ell = 10$	29.2 s	22.6 s	56.6 s	49.3 s
Adaptive Basis Extension	49.0 s	101.4 s	150.8 s	184.1 s

Table 6.2: Computation times for the different methods and different values of c_b.

6.3.4 Comparison of the Computational Time

In a last step we want to compare the computation times of the different algorithms we presented in the last sections to see how much time we can save by using the POD approximation. We can see in Table 6.2 that the solving of all POD approximated problems with a fixed number of basis functions are faster than solving the full problem by a factor of 10 - 20. However, using too few basis functions leads to an increased computation time, especially if the strength of the convection term is increased because the optimization algorithm does not converge properly in this case. Therefore, it might be beneficial to choose the number of POD basis functions according to the energy of the snapshot space (see Section 6.3.1) to save time on the one hand and to get better results on the other hand.
Compared to the POD approximated problems with a fixed number of basis functions, the basis extension algorithm needs more computation time. There are two obvious reasons for this: On the one hand the a-posteriori error estimate has to be computed in each time step to be able to decide whether more basis functions are needed. To compute the estimate, both the state and the adjoint equation have to be solved once using the finite element method, which is costly. On the other hand, extending the basis needs time, as it requires the repeated computation of the same Pareto optimal point. Hence, the more extensions are needed, the higher is the

computational time. An idea to save some time would be again to look at the energy of the snapshot space to get an idea how many basis functions might be needed. By setting the initial number of basis functions ℓ_0 to this number, we might avoid some extensions. Another approach are greedy or optimal sampling algorithms for the generation of the POD basis, see e.g. [18]. Nevertheless, the POD basis extension algorithm is still much faster than solving the full problem by a factor of 4-10, depending on the number of needed basis extensions and hence on c_b in this case.

Chapter 7

Conclusion and Outlook

In this thesis we have investigated the application of the Euclidean reference point method combined with the POD method for model order reduction to the bicriterial optimal control problem (BOCP) governed by the linear heat equation with convection term (PPDE) and the bilateral box constraints (BC).

First of all we have dealt with general convex multiobjective optimization problems and one possible notion of optimality in this framework, namely the Pareto optimality. It has been shown that scalarizing the multiobjective function by using suitable functions is an appropriate strategy to compute Pareto optimal points. Sufficient conditions for the Pareto optimality of the solution of the scalarized optimization problem have been proved. Moreover, it has been illustrated how solving a scalarized optimization problem can provide us with geometrical insight on the shape of the Pareto front (see Remark 3.9 and the Lemmata 3.23 and 3.25).
In the further course two parameter-dependent strategies to approximate the whole Pareto front have been investigated closer, the weighted sum method on the one hand and reference point methods, in particular the Euclidean reference point method, on the other hand. Both approaches have been studied before and it has been observed that it is difficult to choose appropriate weights for the weighted sum method to get a uniform approximation of the Pareto front (see e.g. [6]), whereas this can be achieved for the Euclidean reference point method (see for instance [28, 2]). This has been confirmed by our theoretical investigations. In those we have been able to show a continuous dependence of the Pareto optimal points on the reference points for the Euclidean reference point method (see Theorem 3.47), but to the best of our knowledge it is not possible to prove a similar result for the weighted sum method. This has allowed us to derive a strategy for generating reference points for the Euclidean reference point method, such that it is possible to predict a-priori where the new Pareto optimal point will be situated in a very precise way, only based on the previous reference point, the previous Pareto optimal point and the two parameters h_p and h_x in (3.26).
Additionally, we have justified the use of the Euclidean reference point method in situations, where the reference point does not fulfil $z \leq y^{\mathrm{id}}$ by viewing the Euclidean reference point method as a projection of z onto the set $Y + \mathbb{R}^k_{\geq}$ (see Theorem 3.35). This new point of view has furthermore enabled us to show that all Pareto optimal points can be computed by using the Euclidean reference point method (see Theorem 3.43).
Lastly, we have improved the continuity result of the Pareto optimal controls in dependence on the reference points (see Theorem 3.49) and the a-posteriori error estimate (see Theorem 3.51) in comparison to the corresponding results in [2, Theorems 7 and 8].

In Chapter 5 we have shown that all results that we have proven in Chapter 3 for the Euclidean reference point method can be applied to the reduced bicriterial optimal control problem (RBOCP). Furthermore, it has been demonstrated that the POD method, which has been

introduced in Chapter 4, can be successfully applied to the problem at hand and that under certain assumptions the convergence of the solutions of the POD approximated problem to the solution of the full problem can be expected. Here, we have strongly followed the procedure in [15, 37] by adapting the results which were shown for the weighted sum method to the Euclidean reference point method. The same has been done for the a-posteriori analysis, in which we could show that the a-posteriori error estimate converges to zero under certain assumptions.

In our numerical experiments we have focused on tree aspects: Testing Algorithm 1 for approximating the Pareto front, verifying the theoretical results and estimates that we have shown and testing the quality of the solutions obtained by the POD approximation.
As an extension to the results in [2] we have demonstrated that the reference point method can be successfully applied to a bicriterial optimal control problem, where the linear heat equation in the side condition contains an additional convection term. Moreover, it has been shown that the fineness of the algorithm can be controlled in a very precise way by adapting the parameter h_x in the formula for generating the reference points (3.26).
Concerning the POD approximation, our experiments have shown that the approximation error is quite small while the computing time is much smaller than for the full problem. However, the approximation error is highly dependent on the strength of the convection term and the number of utilised POD basis functions. The stronger the convection, the bigger the approximation error. We have presented two different strategies to choose an appropriate number of basis functions:
The first approach measures the energy in the snapshot space by looking at the eigenvalues of the corresponding operator $\mathcal{R}^n$. Now, the number of POD basis functions can be chosen according to the desired 'ratio of energy', which shall be contained in the POD basis. The drawback of this approach is that it cannot be ensured that the quality of the POD approximation remains on a good level during the process of approximating the whole Pareto front. Therefore, we have proposed a second approach, which utilises the a-posteriori error estimate from Theorem 3.51 to adaptively increase the number of POD basis functions if necessary. For this approach it has been essential that the a-posteriori error estimate from Theorem 3.51 has been shown to be very efficient. In our experiments we have achieved good approximation errors for this approach making up for the higher computation time in this case in comparison with the methods using a fixed number of basis functions.

Looking at both the theoretical and numerical investigations, there are several points of contact for further investigation.
From a theoretical point of view it is also possible to consider a time-dependent convection term $b : (0,T) \times \Omega \to \mathbb{R}^2$, if b additionally fulfils certain regularity properties in the time space in such a way that the bilinear form a still fulfils Lemma 5.1. This is for example satisfied, if $b \in C([0,T], L^\infty(\Omega))$ holds. From the numerical viewpoint we expect that both the state and adjoint equation show more complicated dynamics, which might not be captured that easily by a single POD basis for optimal controls all over the Pareto front. Already for a time-independent convection term we observe that increasing the constant c_b (i.e. the strength of the convection) leads to a requirement of more POD basis functions. Thus, if we allow the convection to be time-dependent, it will probably not be enough to increase the number of used POD basis functions during the computation of the Pareto front, but to make basis updates. Work in this

direction has already been done for instance in [30] for a semilinear heat equation, where a trust region POD algorithm was presented.
Besides, an extension to a domain $\Omega \subset \mathbb{R}^3$ would allow us to model a three-dimensional room. Again, the theoretical investigation can easily be extended to this case because all the computations that we have made also hold for a three-dimensional domain.
Further investigations considering Algorithm 1 might also be worthwhile.
In this thesis we only considered the bicriterial case, in which the choice of reference points can be derived from some geometrical considerations. The extension of the generation of reference points to a higher dimensional case is not straightforward.
Furthermore, a parallelisation of the algorithm might be of interest to save computing time. To be able to implement this, an a-priori choice of all reference points is needed. In the bicriterial case one idea could be to choose uniformly distributed reference points on the line segments between $(y_1^{\mathrm{id}}, y_2^{\mathrm{nad}})^T$ and y^{id} as well as on the line segment between y^{id} and $(y_1^{\mathrm{nad}}, y_2^{\mathrm{id}})^T$, where y^{id} and y^{nad} can be computed by solving the weighted sum problems for the weights $\alpha_1 = (1, 0)$ (or $\alpha_1 = (1, \varepsilon)$ for a small $\varepsilon > 0$) and $\alpha_2 = (0, 1)$.
For our problem we observed that the qualitative control strategies of the heaters were the same for optimal controls all over the Pareto front and they just differed in their scales (see Figure 6.3). An idea to make use of this could be to reduce the dimension of the feasible space by choosing basis functions for the optimal controls which are supposed to model the set of all optimal controls well. This would lead to a similar approach as the POD method for model order reduction of PDEs.

Bibliography

[1] J. Atwell, J. Borggaard, and B. King. Reduced order controllers for Burgers' equation with a nonlinear observer. *Int. J. Appl. Math. Comput. Sci.*, 11:1311–1330, 2001.

[2] S. Banholzer, D. Beermann, and S. Volkwein. POD-based bicriterial optimal control by the reference point method. *IFAC-PapersOnLine*, 49(8):210–215, 2016.

[3] R. Bellman. *Dynamic Programming.* Princeton University Press, Princeton, NJ, USA, 1st edition, 1957.

[4] H. Bock and K. Plitt. A multiple shooting algorithm for direct solution of optimal control problems. *Proceedings of the 9th IFAC world congress*, pages 242–247, 1984.

[5] A. J. Bryson. Optimal Control - 1950 to 1985. *IEEE Control Systems*, 16:26–33, 1996.

[6] I. Das and J. Dennis. A closer look at drawbacks of minimizing weighted sums of objectives for Pareto set generation in multicriteria optimization problems. *Structural Optimization*, 14:63–69, 1997.

[7] I. Das and J. Dennis. Normal-boundary intersection: A new method for generating the Pareto surface in nonlinear multicriteria optimization problems. *SIAM J. Optim.*, 8:631–657, 1998.

[8] K. Deb. *Multi-Objective Optimization Using Evolutionary Algorithms.* John Wiley & Sons, 1st edition, 2001.

[9] R. Denk. Skript zur Vorlesung Theorie partieller Differentialgleichungen II. `https://cms.uni-konstanz.de/math/denk/home/publikationen-und-skripten/skripten/`, 2013. Accessed: 7 February 2017.

[10] R. Denk and R. Racke. *Kompendium der Analysis Band 2.* Springer Spektrum, Wiesbaden, 1st edition, 2012.

[11] M. Ehrgott. *Multicriteria Optimization.* Springer, Berlin, Heidelberg, 2nd edition, 2005.

[12] L. Evans. *Partial Differential Equations.* American Mathematical Society, Providence, Rhode Island, 2nd edition, 2010.

[13] K. Fong, V. Hanby, and T.-T. Chow. HVAC system optimization for energy management by evolutionary programming. *Energy and Buildings*, (38):220–231, 2006.

[14] M. Grepl, Y. Maday, N. Nguyen, and A. Patera. Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations. *ESAIM: M2AN*, 41(3):575–605, 2007.

[15] M. Gubisch and S. Volkwein. Proper orthogonal decomposition for linear-quadratic optimal control. *Model Reduction and Approximation: Theory and Algorithms. SIAM*, 2017. to appear. Benner, P., Cohen, A., Ohlberger, M. and Willcox, K. (eds.).

[16] L. Iapichino, S. Trenz, and S. Volkwein. Reduced-order multiobjective optimal control of semilinear parabolic problems. *Numerical Mathematics and Advanced Applications (ENUMATH 2015), Lecture Notes in Computational Science and Engineering*, 112:389–397, 2016.

[17] L. Iapichino, S. Ulbrich, and S. Volkwein. Multiobjective PDE-constrained optimization using the reduced-basis method. *Advances in Computational Mathematics*, 2017. to appear.

[18] M. Kärcher and M. Grepl. A posteriori error estimation for reduced order solutions of parametrized parabolic optimal control problems. *ESAIM*, 48:1615–1638, 2014.

[19] C. Kelley. *Iterative Methods for Optimization.* Society for Industrial and Applied Mathematics, Philadelphia, 1999.

[20] K. Kunisch and S. Volkwein. Proper orthogonal decomposition for optimality systems. *ESAIM: Math. Model. Numer. Anal.*, 42:1–23, 2008.

[21] J. Lions. *Optimal Control of Systems Governed by Partial Differential Equations.* Springer, Berlin, 1971.

[22] F. Logist, B. Houska, M. Diehl, and J. van Impe. Fast Pareto set generation for nonlinear optimal control problems with multiple objectives. *Struct Multidisc Optim*, 42:591–603, 2010.

[23] K. Malanowski, C. Büskens, and H. Maurer. Convergence of approximations to nonlinear control problems. *CRC Press*, 195:253–284, 1997.

[24] K. Miettinen. *Nonlinear Multiobjective Optimization.* Kluwer Academic Publishers, AH Dordrecht, 1st edition, 1999.

[25] J. Nocedal and S. Wright. *Numerical Optimization.* Springer, 2nd edition, 2006.

[26] S. Ober-Blöbaum, M. Ringkamp, and G. zum Felde. Solving multiobjective optimal control problems in space mission design using discrete mechanics and reference point techniques. *IEEE 51st Annual Conference on Decision and Control (CDC)*, pages 5711–5716, 2012.

[27] V. Pareto. *Manual of Political Economy.* Augustus M. Kelly Publishers, New York, 1971.

[28] S. Peitz, S. Ober-Blöbaum, and M. Dellnitz. Multiobjective optimal control methods for fluid flow using reduced order modeling. `https://arxiv.org/pdf/1510.05819v2.pdf`, 2015. Accessed: 7 February 2017.

[29] L. Pontryagin, V. Boltyanskii, R. Gamkrelidze, and E. Mishchenko. *The Mathematical Theory of Optimal Processes.* Pergamon Press Ltd., 1964.

[30] S. Rogg. Trust Region POD for Optimal Boundary Control of a Semilinear Heat Equation. `http://nbn-resolving.de/urn:nbn:de:bsz:352-0-255879`, 2014. Diploma Thesis – Accessed: 7 February 2017.

[31] C. Romaus, J. Böcker, K. Witting, A. Seifried, and O. Znamenshchykov. Optimal energy management for a hybrid energy storage system combining batteries and double layer capacitors. *Proc. of IEEE ECCE*, pages 1640–16471, 2009.

[32] R. Sargent and G. Sullivan. The development of an efficient optimal control package. *Proceedings of the 8th IFIP conference on optimization techniques*, pages 158–168, 1978.

[33] W. Schilders, H. van der Vorst, and J. Rommes. *Model Order Reduction*. Springer, Berlin Heidelberg, 1st edition, 2008.

[34] W. Stadler. *Multicriteria Optimization in Engineering and in the Sciences*. Plenum Press, New York, 1988.

[35] D. Struik. *A Source Book in Mathematics, 1200-1800*. Harvard University Press, Cambridge, Massachusetts, 1969.

[36] H. Sussmann and J. Willems. 300 years of optimal control: From the brachystochrone to the maximum principle. *IEEE Control Systems*, 17(3):32–44, 1997.

[37] F. Tröltzsch and S. Volkwein. POD a-posteriori error estimates for linear-quadratic optimal control problems. *Comput. Optim. Appl.*, 44:83–115, 2009.

[38] F. Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen. Theorie, Verfahren und Anwendungen*. Vieweg+Teubner, Wiesbaden, Berlin, 2nd edition, 2009.

[39] F. Tröltzsch. *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, Rhode Island, 2010.

[40] D. Werner. *Funktionalanalysis*. Springer, Berlin Heidelberg, 7th edition, 2011.