A Stochastic Primal-Dual Proximal Splitting Method for Risk-Averse Optimal Control of PDEs

---

Ein Stochastisches Primal-Duales Proximal-Splittingverfahren für Risiko-Averse Optimalsteuerung von PDGln

Dissertation

zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften (Dr. rer. nat.)

der Mathematischen Fakultät

der Universität Duisburg-Essen

vorgelegt von

Sebastian Angerhausen, M. Sc.

im Oktober 2022

DOI: 10.17185/duepublico/78165
URN: urn:nbn:de:hbz:465-20230412-081808-2

1. Gutachter: Prof. Dr. Christian Clason

2. Gutachter: Prof. Tuomo Valkonen, PhD

Tag der mündlichen Prüfung: 20.12.2022

Abstract

In this thesis we consider a non-convex optimization problem that is constrained by a partial differential equation (PDE) with uncertain coefficients. The random field PDE solution is taken into account in the objective function by means of the *Conditional Value-at-Risk* (CVaR), which is a well-known risk measure. A particularly useful feature of CVaR comes to light when it is used in the context of a proximal point method, since the proximal operator of its Fenchel conjugate is just the metric projection onto the so-called *bounded probability simplex*. Consequently, we propose a stochastic primal-dual proximal splitting method which is adapted from the well-known *Chambolle-Pock method* and solves the aforementioned problem. The stochasticity or randomness of the algorithm arises from what we call *component-wise gradient freezing* or *CGF*. It is motivated by randomized coordinate descent methods and requires that only a subset of the coordinates of an occurring gradient is recalculated in each iteration. We provide an abstract proof of almost sure weak convergence of the algorithm and specify the results for the case of scalar and deterministic step sizes. Furthermore, we present an algorithm for computing the aforementioned simplex projection and prove its convergence. The reduction of iteration costs due to CGF in terms of saved PDE solutions is presented by means of two numerical examples which are implemented in the Julia programming language.

Zusammenfassung

In dieser Arbeit betrachten wir ein nicht-konvexes Optimierungsproblem, das durch eine partielle Differentialgleichung (PDGl) mit zufälligen Koeffizienten beschränkt wird. Die Lösung der PDGl in Form eines Zufallfeldes wird in der Zielfunktion mit Hilfe des Conditional Value-at-Risk (CVaR) berücksichtigt, welcher ein bekanntes Risikomaß ist. Eine besonders nützliche Eigenschaft des CVaR tritt zutage, wenn er im Zusammenhang mit einem Proximalpunktverfahren verwendet wird, da die Proximalpunktabbildung seiner Fenchel-Konjugierten lediglich die metrische Projektion auf den so genannten *beschränkten Wahrscheinlichkeitssimplex* ist. Folglich präsentieren wir ein stochastisches primal-duales Proximal-Splittingverfahren, das von dem bekannten Chambolle-Pock-Verfahren abgeleitet ist und das oben genannte Problem löst. Die Stochastizität oder Zufälligkeit des Algorithmus ergibt sich aus was wir *komponentenweises Gradienteneinfrieren* oder CGF (vom englischen *component-wise gradient freezing*) nennen. Der Algorithmus ist durch randomisierte Koordinatenabstiegsverfahren motiviert und erfordert, dass nur eine Teilmenge der Koordinaten eines auftrenden Gradienten in jeder Iteration neu berechnet wird. Wir liefern einen abstrakten Beweis für die fast sichere schwache Konvergenz des Algorithmus und spezifizieren die Ergebnisse für den Fall skalarer und deterministischer Schrittweiten. Darüber hinaus stellen wir einen Algorithmus zur Berechnung der oben erwähnten Simplex-Projektion vor und beweisen dessen Konvergenz. Die Reduktion der Iterationskosten durch CGF in Form von eingesparten PDGl-Lösungen wird anhand von zwei numerischen Beispielen dargestellt, die in der Programmiersprache Julia implementiert sind.

Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Dr. Christian Clason. He gave me the freedom I needed to pursue my own ideas and yet helped me not to lose sight of the final goal. He had, at any time of the day, an open ear for my questions and problems, and was always committed to helping me in the best possible way (even if I asked the same stupid question twice). Furthermore, he encouraged me to attend two conferences, which helped me think outside the box and learn about similar work.

I am also thankful to the Faculty of Mathematics for funding my PhD scholarship, which allowed me to fully focus on my research.

Since working on a PhD thesis for more than three years naturally has its impact on the private life as well, I would also like to thank my family and friends for their continuous support. I thank my parents for always believing in me and encouraging me to follow my dreams. I thank my mother in law for her unconditional help with the kids. I thank my friend Timo for the many relaxed and fun evenings that gave my mind a bit of a break. And last but not least I would like to thank my wonderful wife for always having my back and taking good care of our children, for always listening to me when I was frustrated, for supporting my work, no matter what, and, of course, for always cooking excellent food.

Contents

Chapter 1

Introduction

The mathematical field of convex analysis emerged in the 1960s mainly due to the work of R. T. Rockafellar [Roc70] and provides the right tools to deal with the optimization of non-smooth functions. For this reason, convex optimization has become one of the main application areas in the following years. A typical problem for convex optimization is

$$\min_{u \in \mathcal{U}} \quad F(K(u)) + G(u), \tag{1.0.1}$$

where $\mathcal{U}$ and $\mathcal{V}$ are Hilbert spaces. The functions $F\colon \mathcal{V} \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ and $G\colon \mathcal{U} \to \overline{\mathbb{R}}$ are assumed to be proper, convex and lower semi-continuous[1], and $K\colon \mathcal{U} \to \mathcal{V}$ is (at first) required to be a continuous linear operator, which makes the resulting objective function convex. Furthermore, F is assumed to be non-smooth. Problems like this appear in many application fields like mathematical imaging, signal processing, machine learning, and optimal control, to name just a few. If $F \circ K + G$ is proper, convex and lower semi-continuous as well (i.e. if K is a continuous linear operator), then the optimal solution $\bar{u} \in \mathcal{U}$ to problem (1.0.1) satisfies

$$0 \in T(\bar{u}), \tag{1.0.2}$$

for some maximally monotone operator $T\colon \mathcal{U} \rightrightarrows \mathcal{U}^*$ (see Definition 2.2.10). Rockafellar proposed the *proximal point algorithm* [Roc76] to solve this optimality condition based on the property that the operator $(\mathrm{Id} + \gamma T)^{-1}$ with $\gamma > 0$ is single-valued, which follows from [Min62]. The inclusion in (1.0.2) can then be expressed as

$$\bar{u} = (\mathrm{Id} + \gamma T)^{-1}(\bar{u})\,. \tag{1.0.3}$$

The term *proximal point algorithm* stems from the representation of $(\mathrm{Id} + \gamma T)^{-1}$ as a *proximal operator* (as we will see later in Section 2.2.4), which can be seen as the "natural extension of the notion of a projection operator onto a convex set" [CP11b]. Equation (1.0.3) enabled the development of many different fixed-point iteration algorithms for a wide variety of problems, depending on the properties of the involved functions. For example, if K is just the identity, one could use *explicit* or *forward-backward splitting* [CW05; CR97; DS09], which is especially useful if F is differentiable. Another method is called *implicit* or *Douglas–Rachford splitting* [Lio71; GB17; BCH15; CP07; EB92], which is also a special case of the proximal point algorithm [Eck89]. The alternating direction method of multipliers (ADMM) [Gab83; Boy+11] is in turn a special case of the Douglas–Rachford splitting [Eck89] and incorporates the augmented Lagrangian of (1.0.1), as described in [CV20, Section 8.6]. This method requires the inversion of relatively complicated set-valued mappings, which is why additional terms are often added in practice. The resulting method is called *preconditioned ADMM* [ZBO11]. All these *splitting methods* have in common

[1] See Section 2.2 for a definition of the terms *proper*, *convex*, and *lower semi-continuous*.

that they use the particular sum structure of (1.0.1) to split the whole objective function into two parts that can be computed separately.

If K is a continuous linear operator, then it can be shown under mild assumptions [CV20, Theorem 5.10] that the problem (1.0.1) is equal to its dual formulation

$$\min_{v \in \mathcal{V}} \quad -F^*(v) - G^*\left(-K^* v\right), \tag{1.0.4}$$

where $F^*\colon \mathcal{V} \to \overline{\mathbb{R}}$ and $G^*\colon \mathcal{U} \to \overline{\mathbb{R}}$ denote the Fenchel conjugates (see Definition 2.2.6) of F and G, respectively, and K^* is the adjoint operator of K. Both the primal formulation in (1.0.1) and the dual formulation in (1.0.4) can additionally be written as the *saddle-point problem*

$$\min_{u \in \mathcal{U}} \quad \max_{v \in \mathcal{V}} \quad G(u) + \langle v, K(u) \rangle_{\mathcal{V}} - F^*(v), \tag{1.0.5}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{V}}$ denotes the inner product on $\mathcal{V}$[2]. A derivation of the saddle-point formulation can be found in Section 3.1. It explicitly involves the dual variable $v \in \mathcal{V}$ and is the starting point for the analysis of many solution algorithms [CP11a; Cha+18; DL14; HY10; MJV19; Val14; CMV19]. These algorithms are often called *primal-dual methods*, since they update the primal and the dual iterates alternately (we refer to [CP16, Chapter 5] or [Val21] for a comprehensive comparison). One particularly useful algorithm was proposed in 2009 by Pock et al. [Poc+09] for minimizing the so-called *Mumford-Shah functional*, which is "one of the most studied variational approaches to image segmentation", according to the authors. The more general formulation by Chambolle and Pock in 2011 [CP11a] gained massive popularity, which is why this algorithm is also referred to as the *Chambolle-Pock method*. The authors required K to be a continuous linear operator so that the resulting problem is convex. It was shown in [HY12] that this method is also a special case of the proximal point algorithm. However, the requirement of K being a continuous linear operator is quite strong, and prevents the algorithm from being used for many interesting problems. Therefore, some research has recently focused on primal-dual methods for non-convex optimization problems [Val14; CV17; CMV19; MJV19], where K is allowed to be non-linear. These papers, especially [CMV19], also form the basis of the convergence analysis in this thesis. Furthermore, the original Chambolle-Pock method was developed for finite-dimensional real vector spaces $\mathcal{U}$ and $\mathcal{V}$ with an inner product and a norm. Hence, the results in [CP11a] do not cover the case where $\mathcal{U}$ and $\mathcal{V}$ are of infinite dimension. However, this requirement is needed for the optimal control of partial differential equations (PDEs) we consider in this thesis, which is why it has recently been addressed by some papers as well [CV17; VP17; CMV19; MJV19; CMV21].

To be more specific, we consider a constraint in form of a PDE with uncertain coefficients, hence the mathematical area of stochastic programming [BL11; SDR14] is also relevant here. PDEs themselves are of interest for a wide range of applications e.g. in science and engineering, and the modeling of random coefficients makes sense whenever possible uncertainties are to be included into the analysis. In this case, the PDE solution (i.e. state) becomes a random field making the objective function itself random. In general, there are different ways how to substitute this random objective by a scalar value allowing us to apply one of the primal-dual proximal splitting methods mentioned above. For example, one could use the expectation, probabilistic functions, distributionally robust optimization, or risk measures, as discussed in [KS18b]. In this thesis, we focus on the latter and consider a particular risk measure called the *Conditional Value-at-Risk* (CVaR), which is also known as *Average Value-at-Risk*, *Mean Excess Loss*, *Expected Shortfall*, or *Expected Tail Loss*. Its definition involves the Value-at-Risk, which is, for a given probability level β,

[2] Note that we have used the Riesz-Fréchet representation theorem [Bre11, Theorem 5.5] to identify $\mathcal{V}$ and $\mathcal{V}^*$.

the lowest amount α such that, with probability β, the loss will not exceed α. CVaR[3] is the conditional expectation of the losses greater than α. The term *loss* stems from the fact that the random function to be measured often represents a financial loss. Areas of application include portfolio management [RU00; RUZ02], credit risk optimization [And+01], supply chain management [WB09], medical treatment planning [CMP14], and unit commitment of power plants [BPP16]. Rockafellar and Uryasev [RU00] showed that, if CVaR is applied to the objective function of an otherwise linear problem, the approximation by Monte Carlo sampling has an equivalent linear programming formulation. While this approach also works if we only consider discrete random variables with finitely many outcomes [ST06], it can no longer be used when a problem is not linear. Although CVaR is a coherent risk measure (see Proposition 2.5.10), it is not differentiable, which is why a smoothing approach is often used in practice [KS16; ACL06]. However, since the proximal point algorithms presented above do not require the objective function to be differentiable, they are well suited to be applied to such problems without the need of smoothing [MB21; BH15; dOli21]. Furthermore, it turns out that the Fenchel conjugate of CVaR has a particularly useful representation [SDR14; RS06; LD05; KS16], which results in a proximal operator that is simply the metric projection onto the so-called *bounded probability simplex* (see Chapter 5). This projection is easy to compute and hence makes the proximal splitting approach especially useful in this case.

Another component of this work is motivated by coordinate descent methods [Wri15], which in their original meaning are characterized by the fact that "only a subset of the coordinates of the primal and dual iterates is updated at each iteration" [FB19]. Coordinate descent methods are especially suited for parallel computation [BT97a] and have gained massive importance in recent years as they have been successfully applied in various fields, most notably probably machine learning [CHL08; Hsi+08]. If the coordinates to be updated are chosen randomly as proposed in [Nes12], the resulting algorithms are referred to as *randomized* or *stochastic coordinate descent methods* [RT14; Cha+18]. An assumption that is often made in this context is that parts of the objective function are separable or at least block-separable, in the sense that they can be expressed as the sum of functions, each of which is applied to only one coordinate [Nes12] or the coordinates of one block [RT14; Val19; Cha+18; DL14]. Chambolle et al. considered saddle point problems that are separable in the dual variable [Cha+18]. In this case, the proximal operators that are applied to the dual variables can also be split up and computed separately. Therefore, the authors proposed a stochastic extension of the Chambolle-Pock method [CP11a] where only an arbitrarily sampled subset of these variables is updated. The proof of almost sure convergence can be found in [AFC19; GDE20]. Similar works are for example [ZL15; ZS15; DL14; GS21]. In this thesis we choose a slightly different approach and introduce a framework where only a subset of the coordinates of the gradient of K is recalculated in each iteration. Because we still update all the coordinates of the primal and dual variables, we call this procedure *component-wise gradient freezing* or *CGF*. As a result, it is possible to drastically reduce iteration costs by eliminating the time-consuming process of solving PDEs for many samples, as we will see in Chapter 6.

Outline

This thesis is structured as follows:

- In Chapter 2, we summarize some essential results from the mathematical fields that are involved in this thesis, starting with differential calculus in Banach spaces. As mentioned

[3] See Section 2.5.4 for a formal definition of CVaR.

before, we are also dealing with non-smooth functions, which is why some basics of convex analysis are needed here. Furthermore, parts of non-convex analysis are relevant as well due to the non-linearity of K. Building on measure theory, the necessary results from stochastic programming are presented together with some basics of uncertainty, sampling, risk-aversion, and the Conditional Value-at-Risk. We conclude this chapter with elementary results about PDEs.

- The concrete definition of problem (1.0.1) which we consider in this thesis is presented in Chapter 3. We introduce some fundamental assumptions on the functions involved and focus specifically on how the PDE constraint is included. Finally, we prove the existence of a solution and derive the necessary optimality condition.
- In Chapter 4, we present the stochastic primal-dual proximal splitting method to solve the problem described in the previous chapter. We formalize the randomization in order to allow for an abstract proof of almost sure weak convergence in the subsequent sections, for which the so-called *stochastic quasi-Fejér monotonicity* is crucial. In the last section of this chapter, we examine how the previously stated conditions can be satisfied in the case of scalar and deterministic step sizes in order to develop assumptions that are easier to verify in practice.
- As mentioned before, the proximal operator of the Fenchel conjugate of CVaR is the metric projection onto the so-called *bounded probability simplex*. This assertion is proven in Chapter 5, where we also explain the discretization of the probability space. After that, we derive the optimality condition of the problem that has to be solved in order to compute this projection. We present a solution algorithm that was originally developed in the Master's thesis [Ang18] and prove its convergence in the last section.
- In Chapter 6, we first present some examples of how the step sizes and the indices for CGF can be chosen in practice to satisfy the theoretical assumptions of the convergence analysis. After that, we consider two exemplary problems to show how our algorithm performs and how CGF can reduce the iteration costs. For the first problem, which is constrained by an elliptic PDE with a discontinuous and uncertain coefficient, we show in detail that the necessary assumptions are indeed satisfied. We explain the discretization of the PDE and show some numerical results. The constraint of the second problem is the steady Burgers' equation. For this example, we also describe the discretization and present numerical results. All algorithms are implemented in Julia[4], and the corresponding code can be found in [Ang22].
- Chapter 7 contains some concluding remarks and highlights how future research could be based on this work.

[4] For information about the Julia programming language, see https://julialang.org/.

Chapter 2

Background

In this chapter, we summarize some essential results from the mathematical fields that are involved in this thesis. After introducing the main concepts of differential calculus in Banach spaces in the first section, we cover convex analysis, which emerged in the 1960s mainly due to the work of R. T. Rockafellar [Roc70]. It uses convexity instead of differentiability assumptions and is therefore the right toolbox we need to handle non-smooth functionals. However, since only parts of the objective function we will consider are convex, we also use concepts from non-convex analysis. Instead of differentiability and convexity, it requires a continuity property, which is explained in the third section. In the fourth section, we summarize the most important results from measure theory, followed by some basics regarding stochastic programming in section five. In this context, we also address the topic of risk awareness and focus especially on the Conditional Value-at-Risk as a risk measure. Finally, the last section covers the theory of partial differential equations, as they form the constraints of our optimization problem later in this thesis.

Before starting with the first subsection, we introduce some notations that are used throughout this thesis. We will denote the extended real numbers by $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$. If $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ are normed vector spaces, then $\mathbb{L}(X, Y)$ denotes the space of all linear and continuous operators from X to Y, endowed with the operator norm $\|\cdot\|_{\mathbb{L}(X,Y)}$. Furthermore, we denote by $X^* := \mathbb{L}(X, \mathbb{R})$ the dual space of X. The mapping $X^* \times X \ni (x^*, x) \mapsto \langle x^*, x\rangle_{X^*,X} := x^*(x) \in \mathbb{R}$ is called *duality pairing*. The closed ball with radius $r \geq 0$ around an element $x \in X$ is denoted by $\mathbb{B}_r(x) := \{z \in X \mid \|z - x\|_X \leq r\}$. If X is not only a normed vector space but also a Hilbert space (i.e. a Banach space together with an inner product), we denote its inner product by $\langle\cdot,\cdot\rangle_X : X \times X \to \mathbb{R}$. If $A \subset X$ is some set, then its interior is denoted by $\operatorname{int}(A)$, its closure by $\bar{A}$ or $\operatorname{cl}(A)$, and its complement by $A^c := X \setminus A$. The characteristic function χ_A of A is defined as

$$\chi_A(x) := \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{if } x \notin A. \end{cases} \tag{2.0.1}$$

The indicator function δ_A of A is defined as

$$\delta_A(x) := \begin{cases} 0, & \text{if } x \in A, \\ \infty, & \text{if } x \notin A. \end{cases} \tag{2.0.2}$$

The identity function is denoted by Id, and sometimes, if it is helpful for understanding, we use a subscript to make clear in which space the function is defined, e.g. $\mathrm{Id}_X \colon X \to X$. Vectors in finite dimensional spaces will be denoted by boldface letters, e.g. $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ ($n \in \mathbb{N}$).

If $\mathbf{x} \in \mathbb{R}^n$ is such a vector, we denote by $\|\cdot\|_2$ the 2-norm on $\mathbb{R}^n$ and by

$$\operatorname{diag}(\mathbf{x}) := \begin{pmatrix} x_1 & & \\ & \ddots & \\ & & x_n \end{pmatrix} \in \mathbb{R}^{n\times n} \tag{2.0.3}$$

the diagonal matrix of $\mathbf{x}$.

2.1 Differential Calculus in Banach Spaces

Since not all functions we encounter in this thesis are non-smooth, we use this section to briefly introduce two concepts of differentiation in Banach spaces. It is based on [CV20, Section 2.2.].

We consider a mapping $F\colon X \to Y$ between two Banach spaces X and Y. If the one-sided limit

$$F'(x;h) := \lim_{t \searrow 0} \frac{F(x+th) - F(x)}{t} \in Y \tag{2.1.1}$$

exists, it is called the *directional derivative* of F at $x \in X$ in direction $h \in X$. With this preparation, we can now define the two main concepts of differentiation in Banach spaces.

Definition 2.1.1 (Derivatives in Banach Spaces)

Let X and Y be Banach spaces, $F\colon X \to Y$ a mapping, and $x \in X$ be given.

(i) If the directional derivative $F'(x;h)$ exists for all $h \in X$ and

$$\mathrm{D}F(x)\colon X \to Y, \quad h \mapsto \mathrm{D}F(x)h := F'(x;h)$$

is a bounded linear operator, then F is called *Gâteaux differentiable* at x with Gâteaux derivative $\mathrm{D}F(x) \in \mathbb{L}(X,Y)$. If F is Gâteaux differentiable at every $x \in \operatorname{dom}(F)$, it is just called *Gâteaux differentiable*.

(i) If additionally

$$\lim_{\|h\|_X \to 0} \frac{\|F(x+h) - F(x) - \mathrm{D}F(x)h\|_Y}{\|h\|_X} = 0,$$

then F is called *Fréchet differentiable* at x with Fréchet derivative $F'(x) := \mathrm{D}F(x) \in \mathbb{L}(X,Y)$. If F is Fréchet differentiable at every $x \in \operatorname{dom}(F)$, it is just called *Fréchet differentiable*.

(i) If F is Fréchet differentiable and the mapping $F'\colon X \to \mathbb{L}(X,Y)$ is continuous, then F is called *continuously differentiable*.

This definition immediately implies that every Fréchet differentiable mapping is also Gâteaux differentiable. However, the approximation error of F near x by $F(x)+\mathrm{D}F(x)h$ has to be superlinear in $\|h\|_X$ if F is Fréchet differentiable.

2.2 Convex Analysis

Unless stated otherwise, this section is based on [CV20, Part II]. Although most of the results are also valid for a normed vector space X, we require throughout this section that X is a real Hilbert space with inner product $\langle \cdot, \cdot \rangle_X$, so that we can use the Riesz-Fréchet representation theorem [Bre11, Theorem 5.5] to identify X and its dual space X^*.

2.2.1 Convex Functions and Subdifferentials

First, recall that a proper function $F\colon X \to \overline{\mathbb{R}}$, where *proper* means

$$\operatorname{dom}(F) := \{x \in X \mid F(x) < \infty\} \neq \emptyset, \tag{2.2.1}$$

is called *convex* if for all $x, y \in X$ and $\lambda \in [0,1]$ the inequality

$$F\big(\lambda x + (1-\lambda)y\big) \leq \lambda F(x) + (1-\lambda)F(y) \tag{2.2.2}$$

holds. The following lemma summarizes some operations which preserve the convexity of a given functional. We refer to [BC17, Section 8.2] or [CV20, Section 3] for a proof.

Lemma 2.2.1 (Construction of Convex Functions)

Let $F\colon X \to \overline{\mathbb{R}}$ be convex. Then the following functions are convex as well:

(i) αF for all $\alpha \geq 0$;

(ii) $F + G$ for a convex $G\colon X \to \overline{\mathbb{R}}$;

(iii) $F \circ A$ for a linear $A\colon Y \to X$ and a normed vector space Y;

(iv) $x \mapsto \sup_{i \in I} F_i(x)$ with $F_i\colon X \to \overline{\mathbb{R}}$ convex for all $i \in I$ and an arbitrary set I.

Since we also consider non-smooth functions in this work, the classical derivative concepts for Banach spaces introduced in Section 2.1 are not sufficient here. Therefore, we need to define some sort of *generalized derivative*.

Definition 2.2.2 (Convex Subdifferential)

We define the (convex) subdifferential of $F\colon X \to \overline{\mathbb{R}}$ at $x \in \operatorname{dom}(F)$ as

$$\partial F(x) := \left\{ x^* \in X^* \;\middle|\; \langle x^*, \tilde{x} - x \rangle_{X^*,X} \leq F(\tilde{x}) - F(x) \quad \text{for all } \tilde{x} \in X \right\}.$$

Note that $\partial F(x)$ is always a convex set, hence the name *convex* subdifferential. It can be shown that this definition is indeed a generalization in the sense that, if $F\colon X \to \overline{\mathbb{R}}$ is convex and Gâteaux differentiable at $x \in X$ with derivative $DF(x)$, then $\partial F(x) = \{DF(x)\}$ [CV20, Theorem 4.5]. Furthermore, the definition of the subdifferential gives us the following Fermat principle [CV20, Theorem 4.2].

Theorem 2.2.3 (Fermat Principle)

Let $F\colon X \to \overline{\mathbb{R}}$ and $\bar{x} \in \operatorname{dom}(F)$. Then

$$0 \in \partial F(\bar{x}) \quad \Leftrightarrow \quad F(\bar{x}) = \min_{x \in X} F(x).$$

2.2.2 Lower Semi-Continuity

In order to recall the definition of lower semi-continuity, we first introduce the concept of weak convergence: A sequence $(x_n)_{n\in\mathbb{N}} \subset X$ converges weakly to $x \in X$, denoted by $x_n \overset{n\to\infty}{\rightharpoonup} x$, if

$$\langle x^*, x_n\rangle_{X^*,X} \overset{n\to\infty}{\to} \langle x^*, x\rangle_{X^*,X} \qquad \text{for all } x^* \in X^*. \tag{2.2.3}$$

If X is finite-dimensional, the notions of strong and weak convergence coincide [Kes09, Proposition 5.1.3].

Definition 2.2.4 (Lower Semi-Continuity)

Let $F\colon X \to \overline{\mathbb{R}}$ and $x \in X$. If F satisfies

$$F(x) \le \liminf_{n\to\infty} F(x_n)$$

(i) for every sequence $(x_n)_{n\in\mathbb{N}} \subset X$ with $x_n \overset{n\to\infty}{\to} x$, it is called *lower semi-continuous in* $x \in X$;

(ii) for every sequence $(x_n)_{n\in\mathbb{N}} \subset X$ with $x_n \overset{n\to\infty}{\rightharpoonup} x$, it is called *weakly lower semi-continuous in* $x \in X$.

If a function is lower semi-continuous in all $x \in X$, we will just say that it is lower semi-continuous. Since the property of lower semi-continuity will be used frequently in this work, it is useful to have a result which shows how other lower semi-continuous functions can be constructed from a given one. The proof can be found in [CV20, Lemma 2.3].

Lemma 2.2.5 (Construction of Lower Semi-Continuous Functions)

Let $F\colon X \to \overline{\mathbb{R}}$ be lower semi-continuous. Then the following functions are lower semi-continuous as well:

(i) αF for all $\alpha \ge 0$;

(ii) $F + G$ for a lower semi-continuous $G\colon X \to \overline{\mathbb{R}}$;

(iii) $F \circ A$ for a continuous $A\colon Y \to X$ and a Banach space Y;

(iv) $x \mapsto \sup_{i\in I} F_i(x)$ with lower semi-continuous $F_i\colon X \to \overline{\mathbb{R}}$ for all $i \in I$ and an arbitrary set I.

2.2.3 Fenchel Duality

The so-called *Fenchel conjugate*, which is named after the German mathematician Werner Fenchel, allows us to use an equivalent expression of the optimality condition in Theorem 2.2.3, as we will see in Lemma 2.2.8. This is useful if the subdifferential of this transformation is easier to calculate than the subdifferential of the original function.

Definition 2.2.6 (Fenchel Conjugate)

(i) Let $F\colon X \to \overline{\mathbb{R}}$ be proper. We define the Fenchel conjugate F^* of F as

$$F^*\colon X^* \to \overline{\mathbb{R}}, \quad x^* \mapsto F^*(x^*) := \sup_{x \in X} \left(\langle x^*, x \rangle_{X^*,X} - F(x) \right).$$

(ii) Let both $F\colon X \to \overline{\mathbb{R}}$ and its Fenchel conjugate $F^*\colon X^* \to \overline{\mathbb{R}}$ be proper. We define the Fenchel biconjugate F^{**} of F as

$$F^{**}\colon X \to \overline{\mathbb{R}}, \quad x \mapsto F^{**}(x) := \sup_{x^* \in X^*} \left(\langle x^*, x \rangle_{X^*,X} - F^*(x^*) \right).$$

The required properness in the above definitions ensures that the defined functionals are indeed greater than $-\infty$ for all elements of their respective domains. The properness of F^* in (ii) can be guaranteed by assuming that F is bounded from below by an affine functional. Furthermore, F^* is always convex and lower semi-continuous, according to Lemma 2.2.1 (iv) and Lemma 2.2.5 (iv).

The next theorem shows that the biconjugate is, under certain assumptions, equal to the original function [CV20, Theorem 5.1].

Theorem 2.2.7 (Fenchel-Moreau-Rockafellar)

Let $F\colon X \to \overline{\mathbb{R}}$ be proper. Then,

$$F^{**}(x) \leq F(x) \qquad \text{for all } x \in X$$

with equality if and only if F is convex and lower semi-continuous.

The following lemma stems from [CV20, Lemma 5.8] and shows the connection between a function, its Fenchel conjugate, and their respective subdifferentials.

Lemma 2.2.8 (Fenchel-Young)

Let $F\colon X \to \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous. Then the following statements are equivalent for any $x \in X$ and $x^* \in X^*$:

(i) $\langle x^*, x \rangle_{X^*,X} = F(x) + F^*(x^*)$;

(ii) $x^* \in \partial F(x)$;

(iii) $x \in \partial F^*(x^*)$.

For the next result, we need to recall that a function $F\colon X \to \overline{\mathbb{R}}$ is called *coercive*, if for every sequence $(x_n)_{n\in\mathbb{N}} \subset X$ with $\|x_n\|_X \to \infty$ as $n \to \infty$, we also have $F(x_n) \to \infty$. With this property, we can show under which assumptions a solution of an optimization problem exists [CV20, Theorem 2.1].

Theorem 2.2.9 (Existence of Minimizer)

Let $F\colon X \to \overline{\mathbb{R}}$ be proper, coercive and weakly lower semi-continuous. Then the problem

$$\min_{x\in X} F(x)$$

has a solution $\bar{x} \in \operatorname{dom}(F)$.

2.2.4 Monotone and Proximal Operators

We denote by $A\colon X \rightrightarrows Y$ a set-valued mapping $A\colon X \to 2^Y$, where Y is a normed vector space and 2^Y its power set. Its graph is defined as

$$\operatorname{graph}(A) := \left\{ (x, y) \in X \times Y \;\middle|\; y \in A(x) \right\}. \tag{2.2.4}$$

In this section we focus on the case where $Y = X^*$, which includes for example the convex subdifferential mapping $x \mapsto \partial F(x)$ as in Definition 2.2.2. One important property of set-valued mappings is monotonicity, which is defined as follows.

Definition 2.2.10 (Monotonicity of Set-Valued Mappings)

Let $A\colon X \rightrightarrows X^*$ be a set-valued mapping with $\operatorname{graph}(A) \neq \emptyset$. We call A

(i) *monotone*, if

$$\left\langle x_1^* - x_2^*, x_1 - x_2 \right\rangle_{X^*,X} \geq 0$$

for all $(x_1, x_1^*), (x_2, x_2^*) \in \operatorname{graph}(A)$;

(ii) *maximally monotone*, if it is monotone and there does not exist another monotone operator $B\colon X \rightrightarrows X^*$ such that

$$\operatorname{graph}(A) \subsetneq \operatorname{graph}(B);$$

(iii) *strongly monotone* with factor $\gamma > 0$, if

$$\left\langle x_1^* - x_2^*, x_1 - x_2 \right\rangle_{X^*,X} \geq \gamma \|x_1 - x_2\|_X^2$$

for all $(x_1, x_1^*), (x_2, x_2^*) \in \operatorname{graph}(A)$.

Note that the definition immediately implies that $\partial F\colon X \rightrightarrows X$, $x \mapsto \partial F(x)$ is monotone if $F\colon X \to \overline{\mathbb{R}}$ is convex. Moreover, if F is proper, convex, and lower semi-continuous, then ∂F is even maximally monotone [CV20, Theorem 6.11], since X is reflexive. The following lemma from [CV20, Lemma 7.4] shows that there is a similar connection between strong monotonicity of ∂F and strong convexity of F.

Lemma 2.2.11 (Strong Convexity/Monotonicity)

Let $F\colon X \to \overline{\mathbb{R}}$. We consider the following properties:

(i) F is strongly convex with factor $\gamma > 0$, i.e.

$$F\big(\lambda x + (1-\lambda)y\big) + \lambda(1-\lambda)\frac{\gamma}{2}\,\|x-y\|_X^2 \leq \lambda F(x) + (1-\lambda)F(y)$$

for all $x, y \in X$ and $\lambda \in [0,1]$;

(ii) ∂F is strongly monotone with factor $\gamma > 0$.

Then (i) $\Rightarrow$ (ii). If F is proper, convex and lower semi-continuous, then also (ii) $\Rightarrow$ (i).

Since X is a Hilbert space with inner product $\langle\cdot,\cdot\rangle_X$, we can also define the so-called *weighted inner product*, which will be useful later when we consider step size operators. For every linear operator $R \in \mathbb{L}(X,X)$, let

$$\langle x, y\rangle_R := \langle Rx, y\rangle_X \quad \text{for all } x, y \in X, \tag{2.2.5}$$

which is, due to the linearity of R, also a bilinear form. If, in addition to that, R is self-adjoint and positive definite, then $\langle\cdot,\cdot\rangle_R$ is an inner product and we call it *weighted inner product*. Like any inner product it induces a norm.

Definition 2.2.12 (Weighted Norm)

Let $R \in \mathbb{L}(X,X)$ be self-adjoint and positive definite. The weighted norm of $x \in X$ is given by

$$\|x\|_R := \sqrt{\langle x, x\rangle_R}.$$

If R is self-adjoint but only positive semi-definite, which we write as $R \geq 0$, then $\langle\cdot,\cdot\rangle_R$ is a positive semi-definite Hermitian form, which induces the semi-norm $\|\cdot\|_R$. Furthermore, if $S \in \mathbb{L}(X,X)$ is another linear operator, we use the notation $S \geq R$ to express that $S - R \geq 0$, i.e. $S - R$ is positive semi-definite. Note that, in the following, we may use one of the introduced notations without having proven or assumed the desired properties of the linear operator so far. However, we will make up for this afterwards.

The definition of strongly monotone set-valued mappings given in Definition 2.2.10 can analogously be used for continuous linear operators; see [BC17, Definition 2.23].

Definition 2.2.13 (Strongly Monotone Operator)

Let $R \in \mathbb{L}(X,X)$ be a linear operator. We call R *strongly monotone* with factor $\gamma > 0$, if

$$\langle Rx, x\rangle_X \geq \gamma\,\|x\|_X^2$$

for all $x \in X$.

If, at any point, the factor γ is not relevant for the argumentation, then we also use the term *strongly monotone* without explicitly mentioning γ. Note that the definition of the norm directly implies that $R \in \mathbb{L}(X,X)$ is strongly monotone with factor $\gamma > 0$ if and only if $R - \gamma\mathrm{Id}$ is positive semi-definite, which is a slightly stronger assumption than R being positive definite. However,

the benefit of requiring strong monotonicity of $R \in \mathbb{L}(X, X)$ is the resulting invertibility of R, i.e. R^{-1} exists and is an element of $\mathbb{L}(X, X)$ [Cla20, Theorem 15.11]. It is moreover positive definite, since for every $x \in X$ we can define $x' := R^{-1}x$ such that

$$\left\langle R^{-1}x, x \right\rangle_X = \langle Rx', x' \rangle_X > 0 \tag{2.2.6}$$

due to the positive definiteness of R. If R is additionally self-adjoint, then the same is true for R^{-1} [Cla20, Theorem 9.4]. Hence, we can use R^{-1} in Definition 2.2.12 to define a weighted norm if $R \in \mathbb{L}(X, X)$ is strongly monotone and self-adjoint. We will exploit this in the next definition.

Since the set-valued subdifferential mapping is often not very useful for algorithms, we can use the so-called *proximal operator* to find an equation that implicitly defines a subgradient in Lemma 2.2.15. Based on the weighted norm, we present a definition of the weighted proximal operator in the following.

Definition 2.2.14 (Weighted Proximal Operator)

Let $F\colon X \to \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous, and $\Sigma \in \mathbb{L}(X, X)$ a self-adjoint and strongly monotone operator. The weighted proximal operator of F is defined as

$$\operatorname{prox}_F^{\Sigma}\colon X \to X, \quad x \mapsto \operatorname{prox}_F^{\Sigma}(x) := \operatorname*{arg\,min}_{z \in X} \left(\frac{1}{2} \| z - x \|_{\Sigma^{-1}}^2 + F(z) \right).$$

The operator Σ will later play the role of a step size operator. This definition is slightly different from the one in [CV20, (6.12)], where

$$\operatorname{prox}_F\colon X \to X, \quad x \mapsto \operatorname{prox}_F(x) := \operatorname*{arg\,min}_{z \in X} \left(\frac{1}{2} \| z - x \|_X^2 + F(z) \right). \tag{2.2.7}$$

does not involve the weighted norm. However, the two definitions coincide for a scalar step size $\sigma > 0$, since $\operatorname{prox}_{\sigma F} = \operatorname{prox}_F^{\sigma \mathrm{Id}}$. This extension allows us to express the weighted proximal operator of a proper, convex, and lower semi-continuous function $F\colon X \to \overline{\mathbb{R}}$ in terms of the so-called *resolvent*, which is for a set-valued operator $A\colon X \rightrightarrows X$ defined as

$$\mathcal{R}_A\colon X \rightrightarrows X, \quad x \mapsto \mathcal{R}_A(x) := (\mathrm{Id} + A)^{-1}(x). \tag{2.2.8}$$

The unique minimizer $\bar{z}$ in the definition of the weighted proximal operator satisfies the Fermat principle in Theorem 2.2.3, i.e.

$$0 \in \partial \left(\frac{1}{2} \| \cdot - x \|_{\Sigma^{-1}}^2 + F(\cdot) \right)(\bar{z}) = \partial \left(\frac{1}{2} \| \cdot \|_{\Sigma^{-1}}^2 \right)(\bar{z} - x) + \partial F(\bar{z}), \tag{2.2.9}$$

where we have used the sum rule with equality [CV20, Theorem 4.14] and [CV20, Lemma 4.13 (ii)] for the last equality. Since the function within the second subdifferential is differentiable, we obtain

$$0 \in \Sigma^{-1}\{\bar{z} - x\} + \partial F(\bar{z}), \tag{2.2.10}$$

and therefore

$$\mathcal{R}_{\Sigma \partial F}(x) = (\mathrm{Id} + \Sigma \partial F)^{-1}(x) = \bar{z} = \operatorname{prox}_F^{\Sigma}(x). \tag{2.2.11}$$

Since ∂F is maximally monotone and $\Sigma \in \mathbb{L}(X, X)$ a self-adjoint and strongly monotone operator, it follows from [BC17, Proposition 20.24] that $\Sigma \partial F$ is maximally monotone as well. Hence, $\mathcal{R}_{\Sigma \partial F}$ is

indeed single-valued [CV20, Corollary 6.14]. With the following lemma, we have an equivalent expression of the relation in Lemma 2.2.8 (ii). As mentioned before, this will help us find the solution to an optimization problem if we do not explicitly know the subdifferential but are able to calculate the proximal operator instead. The lemma is based on [CV20, Lemma 6.18] but reformulated to allow for linear step size operators.

Lemma 2.2.15

Let $F\colon X \to \overline{\mathbb{R}}$ be proper, convex, and lower semi-continuous, $x, x^* \in X$, and $\Sigma \in \mathbb{L}(X, X)$ a self-adjoint and strongly monotone operator. Then

$$x^* \in \partial F(x) \quad \Leftrightarrow \quad x = \operatorname{prox}_F^{\Sigma}(x + \Sigma x^*).$$

Proof: If we apply Σ to both sides of the subdifferential inclusion and add x, we get

$$\begin{aligned} x^* \in \partial F(x) &\Leftrightarrow x + \Sigma x^* \in (\mathrm{Id} + \Sigma \partial F)(x) \\ &\Leftrightarrow x \in (\mathrm{Id} + \Sigma \partial F)^{-1}(x + \Sigma x^*) \\ &\Leftrightarrow x = \operatorname{prox}_F^{\Sigma}(x + \Sigma x^*), \end{aligned} \tag{2.2.12}$$

where we have used (2.2.11) for the last equivalence. □

2.3 Non-Convex Analysis

If we are dealing with a non-convex function, most results from the previous section can not be applied any more. Therefore, we take a look at some tools from the field of non-convex analysis in this section, which is based on [CV20, Part III]. In the following, we consider a Banach space X. Since we can not rely on differentiability or convexity any more, we need to introduce a form of continuity.

Definition 2.3.1 (Local Lipschitz Continuity)

The function $F\colon X \to \mathbb{R}$ is locally Lipschitz continuous near $x \in X$ if there exist $\delta > 0$ and a Lipschitz constant $L > 0$ such that

$$|F(x_1) - F(x_2)| \leq L \|x_1 - x_2\|_X$$

for all $x_1, x_2 \in \mathbb{O}_\delta(x) := \{x' \in X \mid \|x' - x\|_X < \delta\}$. If F is locally Lipschitz continuous near every $x \in U \subset X$, we call F locally Lipschitz continuous on U.

An important connection between local Lipschitz continuity and the properties introduced in the previous section (i.e. convexity and lower semi-continuity) is given by the following proposition [CV20, Theorem 3.13].

Proposition 2.3.2

If $F\colon X \to \mathbb{R}$ is convex and lower semi-continuous, then it is locally Lipschitz continuous on $\operatorname{int}(\operatorname{dom}(F))$.

Based on this continuity property, we can define the so-called *Clarke subdifferential*, which was first introduced by Frank H. Clarke [Cla90] and is a generalization of the convex subdifferential from Definition 2.2.2.

Definition 2.3.3 (Clarke Subdifferential)

Let $F\colon X \to \mathbb{R}$ be locally Lipschitz continuous near $x \in X$. The Clarke subdifferential of F in x is given by

$$\partial_C F(x) := \left\{ x^* \in X^* \;\middle|\; \langle x^*, h\rangle_{X^*,X} \leq F^\circ(x;h) \text{ for all } h \in X \right\},$$

where $F^\circ(x;h)$ is the generalized directional derivative given by

$$F^\circ(x;h) := \limsup_{\substack{y \to x \\ t \searrow 0}} \frac{F(y+th) - F(y)}{t}.$$

It can be shown that the Clarke subdifferential is indeed a generalization of the convex subdifferential, since the equality $\partial_C F(x) = \partial F(x)$ holds for all $x \in \operatorname{int}(\operatorname{dom}(F))$ if $F\colon X \to \mathbb{R}$ is convex and lower semi-continuous [CV20, Theorem 13.8]. The proof of this statement also shows that every convex and lower semi-continuous function $F\colon X \to \mathbb{R}$ is *regular* in the sense that its directional derivative and generalized directional derivative coincide for all directions $h \in X$, i.e.

$$F^\circ(x;h) = F'(x;h) = \lim_{t \searrow 0} \frac{F(x+th) - F(x)}{t} \tag{2.3.1}$$

for all $x \in \operatorname{int}(\operatorname{dom}(F))$. Now, Definition 2.3.3 immediately implies the following optimality condition, which stems from [CV20, Theorem 13.4].

Theorem 2.3.4 (Fermat Principle)

If $F\colon X \to \mathbb{R}$ has a local minimum in $\bar{x} \in \operatorname{dom}(F)$, then $0 \in \partial_C F(\bar{x})$.

The Fermat principle is the starting point of many non-smooth optimization algorithms. Therefore, it is reasonable to develop some rules that will make it easier to calculate the Clarke subdifferential of the composition of two functions. For example, we will use the following sum rule from [CV20, Theorem 13.20] in Section 3.3, where the objective function is indeed a sum of two functions.

Theorem 2.3.5 (Sum Rule)

Let $F, G\colon X \to \mathbb{R}$ be locally Lipschitz continuous near $x \in X$. Then

$$\partial_C (F+G)(x) \subset \partial_C F(x) + \partial_C G(x).$$

If F and G are regular at x, then $F+G$ is regular at x, and equality holds.

Before we present the next calculus rule, we recall the definition of the adjoint operator. Let therefore Y be another normed vector space. For any continuous linear operator $A \in \mathbb{L}(X,Y)$, the *adjoint operator* $A^* \in \mathbb{L}(Y^*, X^*)$ is defined as

$$\langle A^* y^*, x\rangle_{X^*,X} = \langle y^*, Ax\rangle_{Y^*,Y} \qquad \text{for all } x \in X,\ y^* \in Y^*. \tag{2.3.2}$$

The following chain rule handles the composition of a locally Lipschitz continuous and a continuously Fréchet differentiable function [CV20, Theorem 13.23].

Theorem 2.3.6 (Chain Rule)

Let Y be a separable Banach space, $K\colon X \to Y$ continuously Fréchet differentiable at $x \in X$, and $F\colon Y \to \mathbb{R}$ locally Lipschitz continuous near $K(x)$. Then,

$$\partial_C (F \circ K)(x) \subset K'(x)^* \partial_C F(K(x)) := \left\{ K'(x)^* y^* \;\middle|\; y^* \in \partial_C F(K(x)) \right\}.$$

If F is regular at $K(x)$, then $F \circ K$ is regular at x, and equality holds.

2.4 Measure Theory

This section is mainly based on both [Kle13, Chapters 1 and 7] and [Hyt+17, Chapter 1]. Let $(\Omega, \mathcal{A}, \mu)$ be a σ-finite measure space, i.e. Ω is a non-empty set, $\mathcal{A} \subset 2^\Omega$ a σ-algebra such that there exists a sequence $(\Omega_k)_{k\in\mathbb{N}} \subset \mathcal{A}$ with $\Omega = \cup_{k\in\mathbb{N}} \Omega_k$ and $\mu(\Omega_k) < \infty$ for all $k \in \mathbb{N}$, and $\mu\colon \mathcal{A} \to [0, \infty]$ is a measure.

Definition 2.4.1 (Measurable Function)

If $(\Omega', \mathcal{A}', \mu')$ is another σ-finite measure space, then we call a function $f\colon \Omega \to \Omega'$ ($\mathcal{A}$-)*measurable*, if

$$f^{-1}(A') \subset \mathcal{A} \qquad \text{for all } A' \in \mathcal{A}'.$$

In the following, if $\Omega' = \mathbb{R}$, we will always assume that $\mathcal{A}' = \mathcal{B}(\mathbb{R})$ is the Borel σ-algebra and μ' the Lebesgue measure, unless stated otherwise. Now, we consider for every measurable function $f\colon \Omega \to \mathbb{R}$ and $p \in [1, \infty)$ the mappings

$$\| f \|_p := \left(\int_\Omega | f(x) |^p \, \mathrm{d}\mu(x) \right)^{1/p} \tag{2.4.1}$$

as well as

$$\| f \|_\infty := \inf \{ K \geq 0 \mid \mu(|f| > K) = 0 \}. \tag{2.4.2}$$

Following [Hin+09, Section 1.2.2.3], we can then define the Lebesgue or L^p-space.

Definition 2.4.2 (L^p-Space)

Let $p \in [1, \infty]$. We define the *L^p-space* by the quotient space

$$L^p(\Omega) := L^p(\Omega, \mathcal{A}, \mu) := \mathcal{L}^p(\Omega, \mathcal{A}, \mu) / \sim$$

where

$$\mathcal{L}^p(\Omega, \mathcal{A}, \mu) := \mathcal{L}^p(\Omega) := \left\{ f\colon \Omega \to \mathbb{R} \;\middle|\; f \text{ is } \mu\text{-measurable and } \| f \|_p < \infty \right\}$$

and

$$f \sim g :\Leftrightarrow \| f - g \|_p = 0 \quad \text{for all } f, g \in \mathcal{L}^p(\Omega).$$

Therefore, $L^p(\Omega)$ is the space of all equivalence classes of μ-a.e. identical functions in $\mathcal{L}^p(\Omega)$. The term μ-a.e. is short for *μ-almost everywhere* and means that there exists a null set $N \in \mathcal{A}$ (i.e. $\mu(N) = 0$) such that the respective property (e.g. $f(x) = g(x)$) holds for all $x \in \Omega \setminus N$. If μ is a probability measure as in Section 2.5.1, we will use the term μ-a.s. instead, which stands for *μ-almost surely*.

In the following, we will also use the notation $\|\cdot\|_{L^p(\Omega)}$ instead of $\|\cdot\|_p$. It can be shown that $\|\cdot\|_p : L^p(\Omega, \mathcal{A}, \mu) \to [0, \infty)$ is indeed a norm for every $p \in [1, \infty]$. Furthermore, if we have $1 \leq p \leq q$, then the inclusion

$$L^q(\Omega, \mathcal{A}, \mu) \subset L^p(\Omega, \mathcal{A}, \mu) \tag{2.4.3}$$

holds. The case $p = 2$ will be of special interest in this thesis, because $Y := L^2(\Omega, \mathcal{A}, \mu)$ is a Hilbert space with the inner product $\langle \cdot, \cdot \rangle_Y : Y \times Y \to \mathbb{R}$ given by

$$\langle f, g \rangle_Y := \int_\Omega f(x) g(x) \, \mathrm{d}\mu(x) \tag{2.4.4}$$

for all $f, g \in Y$.

Since we are concerned with PDEs in this thesis, we will also use a more general concept where the image space of the function f is not $\mathbb{R}$. Let therefore X be a separable Banach space. Before stating the following definition adapted from [Hyt+17, Definition 1.1.14], we note that a function $f : \Omega \to X$ is called *μ-simple*, if there exist a number $N \in \mathbb{N}$, elements $z_k \in X$, and sets $A_k \in \mathcal{A}$ satisfying $\mu(A_k) < \infty$ for all $k \in \{1, \ldots, N\}$ such that

$$f(x) = \sum_{k=1}^{N} \chi_{A_k}(x) z_k \tag{2.4.5}$$

for all $x \in \Omega$ [Hyt+17, Definition 1.1.13].

Definition 2.4.3 (Strongly Measurable Function)

A function $f : \Omega \to X$ is called *strongly μ-measurable*, if there exists a sequence $(f_k)_{k \in \mathbb{N}}$ of μ-simple functions $f_k : \Omega \to X$ converging to f μ-a.e.

Note that, since X is separable, any strongly μ-measurable function is μ-almost everywhere equal to a measurable function in the sense of Definition 2.4.1; see [Hyt+17, Proposition 1.1.16] and [Hyt+17, Corollary 1.1.10]. Analogously to Definition 2.4.2, we can now define the so-called *Bochner space* [Hyt+17, Definition 1.2.15].

Definition 2.4.4 (Bochner Space)

Let $p \in [1, \infty]$. We define the *Bochner space* $L^p(\Omega; X)$ by the quotient space

$$L^p(\Omega; X) := L^p(\Omega, \mathcal{A}, \mu; X) := \mathcal{L}^p(\Omega, \mathcal{A}, \mu; X) / \sim$$

where

$$\mathcal{L}^p(\Omega, \mathcal{A}, \mu; X) := \left\{ f : \Omega \to X \;\middle|\; f \text{ is strongly } \mu\text{-measurable and } \|f\|_p < \infty \right\}$$

and

$$f \sim g :\Leftrightarrow \|f - g\|_p = 0 \quad \text{for all } f, g \in \mathcal{L}^p(\Omega, \mathcal{A}, \mu; X).$$

The so-defined space is a Banach space for every $p \in [1, \infty]$. Furthermore, the Bochner space $L^p(\Omega; X)$ and the Lebesgue space $L^p(\Omega)$ coincide if $X = \mathbb{R}$, i.e. Definition 2.4.4 is a generalization of Definition 2.4.2. Consequently, the norm $\|\cdot\|_p$ we used in Definition 2.4.4 is a generalization of (2.4.1) and (2.4.2), respectively, where the absolute values have been replaced by the norm on X, i.e.

$$\| f \|_p := \left(\int_\Omega \| f(x) \|_X^p \, \mathrm{d}\mu(x) \right)^{1/p} \tag{2.4.6}$$

for $p \in [1, \infty)$, and

$$\| f \|_\infty := \inf \left\{ K \geq 0 \,\middle|\, \mu\left(\| f \|_X > K \right) = 0 \right\} \tag{2.4.7}$$

for every strongly μ-measurable function $f \colon \Omega \to X$.

2.5 Stochastic Programming

In this section, we give a short introduction to the general framework of stochastic programming [BL11; SDR14]. First, we explain the concept of uncertainty and how it is incorporated into optimization models. Then, we briefly introduce a sample-based discretization approach which will be used for the numerical experiments in Chapter 6. After that, we show how risk can be accounted for in these models and focus especially on the Conditional Value-at-Risk as a risk measure. Some expressions from the fields of probability and measure theory are used without comprehensive explanation because this would go beyond the scope of this thesis. However, the interested reader is referred to the respective references.

2.5.1 Uncertainty

The difference between deterministic and stochastic programming problems is the uncertainty in the parameters of a stochastic program. In an application, this can be for example the return of a financial instrument or the amount of wind energy produced during the day ahead. This uncertainty is modeled as random variables on a probability space. The following summary is based on [Kle13, Chapters 1 and 5].

Definition 2.5.1 (Probability Space)

A triple $(\Xi, \mathcal{A}, \mathbb{P})$ is called a *probability space* if

(i) Ξ is a non-empty set,

(i) $\mathcal{A} \subset 2^\Xi$ is a σ-algebra, and

(i) $\mathbb{P} \colon \mathcal{A} \to [0, \infty]$ is a measure with $\mathbb{P}(\Xi) = 1$.

In this case, the measure $\mathbb{P}$ is called a *probability measure*.

If $(\Xi, \mathcal{A}, \mathbb{P})$ is a probability space, we use $\mathcal{A}^1$ to denote the set of all elements of $\mathcal{A}$ with probability 1, i.e.

$$\mathcal{A}^1 := \{ A \in \mathcal{A} \mid \mathbb{P}(A) = 1 \}. \tag{2.5.1}$$

For a proper definition of the used terms *σ-algebra* and *measure*, we refer to [Kle13, Chapter 1]. With this preparation we can now define a random variable.

Definition 2.5.2 (Random Variable)

Let $(\Xi, \mathcal{A}, \mathbb{P})$ be a probability space, i.e. Ξ is a non-empty set, $\mathcal{A} \subset 2^{\Xi}$ a σ-algebra (i.e. a *set of events*) and $\mathbb{P}\colon \mathcal{A} \to [0,1]$ a probability measure.

(i) A measurable function $\hat{z}\colon \Xi \to \Xi'$, where $(\Xi', \mathcal{A}')$ is a measurable space, is called *random variable*.

(ii) If $\mathcal{B} \subset \mathcal{A}$ is a sub-σ-algebra, we define the space of $\mathcal{B}$-measurable random variables $\hat{z}\colon \Xi \to \Xi'$ as

$$\mathcal{R}(\mathcal{B}, \Xi') := \{\hat{z}\colon \Xi \to \Xi' \mid \hat{z} \text{ is } \mathcal{B}\text{-measurable}\}.$$

(iii) If $\hat{z}\colon \Xi \to \Xi'$ is a random variable, and $(\Xi', \mathcal{A}') = (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, where $\mathcal{B}_{\mathbb{R}}$ is the Borel σ-algebra, then $\hat{z}$ is called a *real random variable*.

(iv) A vector of real random variables is called *random vector*.

Every random variable $\hat{z}$ induces an *image measure* $\mathbb{P}_{\hat{z}} := \mathbb{P} \circ \hat{z}^{-1}$ in the sense of [Kle13, Definition 1.98], which is called the *distribution* of $\hat{z}$ [Kle13, Definition 1.103]. If $\hat{z}$ is a random vector of dimension $d \in \mathbb{N}$, the mapping

$$F_{\hat{z}}\colon \mathbb{R}^d \to [0,1], \quad x \mapsto F_{\hat{z}}(x) := \mathbb{P}\big(\hat{z}_i \leq x_i \text{ for all } i \in \{1, \dots, d\}\big) \tag{2.5.2}$$

is called the (joint) *distribution function* of $\hat{z}$. Moreover, if this distribution function takes the form

$$F_{\hat{z}}(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f_{\hat{z}}(t_1, \dots, t_d)\, \mathrm{d}t_d \dots \mathrm{d}t_1 \tag{2.5.3}$$

for all $x \in \mathbb{R}^d$ and some integrable function $f_{\hat{z}}\colon \mathbb{R}^d \to [0, \infty)$, then $f_{\hat{z}}$ is called the *density* of $F_{\hat{z}}$ or the density of $\hat{z}$.

Since a probability space is a measure space, the concepts introduced in Section 2.4 are also applicable here. However, in Chapter 6 we will use a special notation for a Bochner space that is defined on a probability space with the image measure of a random vector. Therefore, we briefly introduce this notation in the following. Let

$$\hat{z}\colon \Xi \to S \subset \mathbb{R}^d \tag{2.5.4}$$

be a random vector with density $\rho\colon S \to [0, \infty)$. Following Definition 2.4.4, we define for every $q \in [1, \infty]$ the Bochner space

$$L^q_\rho(S; X) := L^q(S, \mathcal{B}(S), \mathbb{P}_{\hat{z}}; X), \tag{2.5.5}$$

where $\mathcal{B}(S)$ is the Borel σ-algebra of S and $\mathbb{P}_{\hat{z}} := \mathbb{P} \circ \hat{z}^{-1}$ the image measure of $\hat{z}$. Due to [Kle13, Theorem 4.10], this especially implies that

$$y \in L^q_\rho(S; X) \quad \Leftrightarrow \quad y \circ \hat{z} \in L^q(\Xi; X). \tag{2.5.6}$$

In the following definition, we present a way to *evaluate* (i.e. assign real numbers to) random variables, which is necessary since we want to incorporate random variables into the objective function of an optimization problem.

Definition 2.5.3 (Expected Value)

The *expected value* (also referred to as *expectation* or *mean*) of a real random variable $\hat{z} \in L^1(\Xi, \mathcal{A}, \mathbb{P})$ is given by

$$\mathbb{E}(\hat{z}) := \int_{\Xi} \hat{z} \, \mathrm{d}\mathbb{P}.$$

When we explain the randomization of the algorithm in Section 4.2, we will also use the *conditional expectation* [Kle13, Definition 8.11], which is defined as follows.

Definition 2.5.4 (Conditional Expectation)

Let $\mathcal{F} \subset \mathcal{A}$ be a sub-σ-algebra and $\hat{z} \in L^1(\Xi, \mathcal{A}, \mathbb{P})$. A random variable $\hat{y} \colon \Xi \to \mathbb{R}$ is called the *conditional expectation* of $\hat{z}$ given $\mathcal{F}$, if

(i) $\hat{y}$ is $\mathcal{F}$-measurable,

(ii) for any $A \in \mathcal{F}$, we have $\mathbb{E}(\hat{z}\mathbf{1}_A) = \mathbb{E}(\hat{y}\mathbf{1}_A)$.

In this case, we also write $\mathbb{E}(\hat{z} \mid \mathcal{F}) := \hat{y}$.

The so-called *filtration* will help us modeling the information available in each iteration of the algorithm. It stems from the theory of stochastic processes [Kle13, Chapter 9.1] and is defined as follows.

Definition 2.5.5 (Filtration)

A filtration $\mathfrak{A} := (\mathcal{A}_k)_{k \in \mathbb{N}_0}$ is a family of sub-σ-algebras $\mathcal{A}_k \subset \mathcal{A}$, $k \in \mathbb{N}_0$, such that $\mathcal{A}_k \subset \mathcal{A}_{k+1}$ for all $k \in \mathbb{N}_0$.

Note that we could use a more general index set here, but $\mathbb{N}_0$ is sufficient in our case.

With these preparations, we could define a simple stochastic optimization model. For example, if $\mathcal{V} := L^2(\Xi, \mathcal{A}, \mathbb{P})$ is the space of random variables, $\mathcal{U}$ is another Hilbert space, and $f \colon \mathcal{U} \to \mathcal{V}$ a mapping, then we could consider

$$\min_{u \in \mathcal{U}} \quad \mathbb{E}(f(u)). \tag{2.5.7}$$

In practice, this could correspond to the following scenario: $\mathcal{U}$ is a space whose elements represent an investment portfolio composition and $f \colon \mathcal{U} \to \mathcal{V}$ calculates the random financial loss of this portfolio. Then, the optimization problem would seek for a portfolio composition which results in a minimal expected loss. If we stick with this example, it quickly becomes clear that only minimizing the *average* loss is often not good enough (e.g. since large price fluctuations can cause a total loss, no matter what the average price of the respective instrument is). Therefore, concepts of *risk averse optimization* as introduced in Section 2.5.3 can help mitigating these kind of problems.

2.5.2 Sampling

As soon as we leave the theoretical analysis of stochastic programming algorithms, we need to think about a way how to discretize the random variables so as to solve the resulting problems by

deterministic optimization techniques. One way to do this is called *sample average approximation* or *SSA*, which we explain in the following, based on [SDR14, Chapter 5.1]. For that purpose, we modify the problem in (2.5.7) such that the function $f\colon \mathcal{U} \times \mathbb{R}^d \to \mathbb{R}$ explicitly takes the image of a d-dimensional ($d \in \mathbb{N}$) random vector $\hat{\xi} := \left(\hat{\xi}_1, \ldots, \hat{\xi}_d\right)$ with $\hat{\xi}_i \in \mathcal{V}$ for all $i \in \{1, \ldots, d\}$ as the second argument. The problem can then be written as

$$\min_{u \in \mathcal{U}} \quad \mathbb{E}\left(f\left(u, \hat{\xi}(\cdot)\right)\right). \tag{2.5.8}$$

Now, we assume that we have a sample $\left\{\xi^j \in \mathbb{R}^d \;\middle|\; j \in \{1, \ldots, S\}\right\}$ of $S \in \mathbb{N}$ realizations of the random vector $\hat{\xi}$. This can be achieved by an appropriate transformation of independently generated and uniformly distributed random numbers, which is called *Monte Carlo sampling* [SDR14, Section 5.3]. While this approach only works if one knows the distribution of the random vector $\hat{\xi}$, one could also view the sample as historical data of S observations. The sample average approximation of (2.5.8) is then given by computing the average of the values $f\left(u, \xi^j\right)$ for $j \in \{1, \ldots, S\}$, i.e.

$$\min_{u \in \mathcal{U}} \quad \frac{1}{S} \sum_{j=1}^{S} f\left(u, \xi^j\right). \tag{2.5.9}$$

If we recall the definition of the expected value in Definition 2.5.3, we see that the sum stems from the discretization of the integral. As a consequence of the *Uniform Law of Large Numbers* [SDR14, Section 7.2.5] it was shown in [SDR14, Section 5.1.1] that, under mild regularity conditions, the optimal solutions and the optimal value of the sample average approximation problem (2.5.9) converge $\mathbb{P}$-a.s. to the optimal solutions and the optimal value of the original problem (2.5.8) as the sample size increases. Hence, it is a reasonable discretization approach.

2.5.3 Risk Aversion

In Section 2.5.1, we have seen an example of an optimization problem in which the objective function is defined as the expected value of a random variable. This kind of structure is reasonable if the *Law of Large Numbers* [Kle13, Chapter 5] can be invoked and our major concern of the real-world problem is the long-term performance. However, this approach does not consider fluctuations of specific outcome realizations. This could cause, for example, the loss of all invested money in a portfolio optimization problem. Therefore, we need to extend the model from Section 2.5.1 by the capability of measuring and minimizing risk.

Let $\hat{z} \in L^1(\Xi, \mathcal{A}, \mathbb{P})$ be a random variable. According to [SDR14, Chapter 6.2], the main idea of a so-called *mean-risk model* is to characterize the uncertain outcome of $\hat{z}$ by two scalars: the expected value $\mathbb{E}(\hat{z})$ describing the expected outcome, and the uncertainty of that outcome in form of a risk or dispersion measure $\mathbb{D}(\hat{z})$. These two objectives are combined to one objective function of the form

$$\mathbb{E}(\hat{z}) + \gamma \mathbb{D}(\hat{z}), \tag{2.5.10}$$

where the coefficient $\gamma \geq 0$ can be varied in order to change the weight of the risk measure, i.e. the price of risk. By minimizing this objective function for a variety of values of γ, we generate a set of so-called *efficient* solutions. These solutions have the property that they, on the one hand, minimize the value of the risk measure for a given expected value, and, on the other hand, minimize the expected value for a given value of the risk measure. If we use this approach to

incorporate a risk measure $\mathbb{D}\colon L^1(\Xi, \mathcal{A}, \mathbb{P}) \to \overline{\mathbb{R}}$ into the problem (2.5.7), it becomes

$$\min_{u \in \mathcal{U}} \quad \mathbb{E}(f(u)) + \gamma \mathbb{D}(f(u)), \tag{2.5.11}$$

for $\gamma \geq 0$.

Another possible approach is to drop the expected value and only consider a risk measure within the objective function, i.e.

$$\min_{u \in \mathcal{U}} \quad \mathbb{D}(f(u)). \tag{2.5.12}$$

This is basically a part of the structure of our problem defined in Chapter 3. Although there is a large variety of risk measures [SDR14, Chapter 6.2], we will focus on one particular measure called *Conditional Value-at-Risk*, which is introduced in the following section.

2.5.4 Conditional Value-at-Risk

In this section, we introduce the Conditional Value-at-Risk (CVaR), which is the particular risk measure we focus on in this thesis. As mentioned in Chapter 1, CVaR is also referred to as *Average Value-at-Risk*, *Mean Excess Loss*, *Expected Shortfall*, or *Expected Tail Loss*. It has become an important tool for risk management in insurance and finance [RU00, Chapter 3] (e.g. for hedging a portfolio of investment instruments) and is a potential choice whenever it comes to measuring the risk of some monetary loss. Possible fields of application are supply chain management [WB09, Section 8.2] and unit commitment of power plants [BPP16], to name just a few. However, since the definition of CVaR does not restrict its use to financial losses, many other applications are conceivable. These include, for example, reinforcement learning [San+17], statistical learning [TK09], wireless data broadcast systems [Kam+14], and medical treatment planning [An+17]. While, in these examples, the CVaR is used either within a constraint or as part of the objective function, we will focus only on the latter case.

In the following, we will first give a formal definition of the CVaR based on both [RU00, Chapter 2] and [SDR14, Section 6.2.4], and then summarize some of its properties that are needed in subsequent chapters. Let

$$\hat{z} \in L^1(\Xi, \mathcal{A}, \mathbb{P}) \tag{2.5.13}$$

be a real-valued random variable. We assume that the distribution of $\hat{z}$ has a density $p_{\hat{z}}\colon \mathbb{R} \to [0, \infty)$ [Kle13, Definition 1.106], which does not necessarily need to take an analytical form (we have seen in Section 2.5.2 that, for applications, it is enough to have an algorithm which generates samples according to $p_{\hat{z}}$). The value $\hat{z}(\xi)$ for an event $\xi \in \Xi$ is referred to as *loss*; $\hat{z}$ itself is called the *loss-function*. The probability of $\hat{z}$ not exceeding a threshold $\alpha \in \mathbb{R}$ is given by its distribution function $F_{\hat{z}}\colon \mathbb{R} \to [0, 1]$, which is defined as

$$\alpha \mapsto F_{\hat{z}}(\alpha) := \mathbb{P}\big(\{\xi \in \Xi \mid \hat{z}(\xi) \leq \alpha\}\big) = \int_{-\infty}^{\alpha} p_{\hat{z}}(z)\, \mathrm{d}z. \tag{2.5.14}$$

$F_{\hat{z}}$ is non-decreasing and continuous from the right, and we assume for simplicity that it is also continuous from the left. With this, we can define the Value-at-Risk as described informally in Chapter 1. The usage of the minimum instead of the infimum is thereby justified by the continuity of $F_{\hat{z}}$ from the right.

Definition 2.5.6 (Value-at-Risk)

For a given probability level $\beta \in (0,1)$, the Value-at-Risk of a loss-function $\hat{z} \in L^1(\Xi, \mathcal{A}, \mathbb{P})$ is defined as

$$\mathrm{VaR}_\beta(\hat{z}) := \inf\{\alpha \in \mathbb{R} \mid F_{\hat{z}}(\alpha) \geq \beta\},$$

where $F_{\hat{z}}$ is the distribution function of $\hat{z}$.

Although β can take any value between 0 and 1, the three values commonly considered are 0.90, 0.95 and 0.99. Note that the VaR could also serve us as a risk measure, but Artzner et al. showed in [Art+99] that it has some undesirable mathematical properties such as the lack of subadditivity[1] (see (2.5.15)), which directly implies non-convexity. Therefore, we focus on the CVaR, which has much nicer properties. Apart from that, the definition of the CVaR (see Definition 2.5.7) ensures that it is always greater than or equal to the VaR, meaning that a low CVaR leads to a low VaR as well.

Since $F_{\hat{z}}$ is continuous and non-decreasing, the set $\{\alpha \in \mathbb{R} \mid F_{\hat{z}}(\alpha) = \beta\}$ is either a singleton or a non-empty interval. In the latter case, the value of $\mathrm{VaR}_\beta(\hat{z})$ is obviously the left endpoint of this interval [RU00, Page 5]. Consequently, the probability of the event $\left\{\xi \in \Xi \,\middle|\, \hat{z}(\xi) \geq \mathrm{VaR}_\beta(\hat{z})\right\}$ is equal to $1-\beta$, which explains the factor $(1-\beta)^{-1}$ in the following definition. Note that, in some publications, the concept of Conditional Value-at-Risk is called *Average Value-at-Risk*, for reasons that are well explained in [SDR14, Theorem 6.2]. One can also find notations where the subscript β is the complementary probability of our probability level.

Definition 2.5.7 (Conditional Value-at-Risk)

For a given probability level $\beta \in (0,1)$, the Conditional Value-at-Risk of a loss-function $\hat{z} \in L^1(\Xi, \mathcal{A}, \mathbb{P})$ is defined as its conditional expectation relative to the loss being greater than or equal to its VaR_β, i.e.

$$\begin{aligned}\mathrm{CVaR}_\beta(\hat{z}) &:= \mathbb{E}\left(\hat{z} \,\middle|\, \hat{z} \geq \mathrm{VaR}_\beta(\hat{z})\right) \\ &= (1-\beta)^{-1} \int_{z \geq \mathrm{VaR}_\beta(\hat{z})} z p_{\hat{z}}(z)\, \mathrm{d}z,\end{aligned}$$

where $p_{\hat{z}}$ is the density of $\hat{z}$.

Note that the reason why we do not consider the case $\beta = 0$ is that Definition 2.5.6 implies $\mathrm{VaR}_0(\hat{z}) = -\infty$ for all $\hat{z} \in L^1(\Xi, \mathcal{A}, \mathbb{P})$, and hence $\mathrm{CVaR}_0(\hat{z}) = \mathbb{E}(\hat{z})$, which is not a risk measure any more. For background information regarding conditional expectations beyond Definition 2.5.4 and how the second equality in Definition 2.5.7 is obtained, we refer to [Kle13, Chapter 8]. The condition $\hat{z} \in L^1(\Xi, \mathcal{A}, \mathbb{P})$ implies that $\mathrm{CVaR}_\beta(\hat{z})$ is finite-valued and thus well defined. If the distribution function $F_{\hat{z}}$ is not continuous as we have assumed (e.g. if the random variable $\hat{z}$ has a discrete probability distribution), then the definition of CVaR has to be altered [RU02].

As mentioned before, the CVaR has some nice properties which we will utilize later. Before presenting them, we follow [SDR14, Section 6.3] and briefly discuss the case of a general risk measure $\mathcal{R}\colon L^p(\Xi, \mathcal{A}, \mathbb{P}) \to \mathbb{R} \cup \{-\infty\} \cup \{\infty\}$ with $p \in [1, \infty)$ and properness as the only restriction. First, we introduce a pointwise partial order $\hat{z} \succcurlyeq \hat{z}'$ for $\hat{z}, \hat{z}' \in L^p(\Xi, \mathcal{A}, \mathbb{P})$, meaning $\hat{z}(\xi) \geq \hat{z}'(\xi)$ for $\mathbb{P}$-almost all $\xi \in \Xi$. This is equivalent to requiring $F_{\hat{z}}(x) \leq F_{\hat{z}'}(x)$ for all

[1] In fact, VaR is only subadditive if it is based on the standard deviation of normal distributions [Art+99, Remark 3.6].

$x \in \mathbb{R}$, where $F_{\hat{z}}$ and $F_{\hat{z}'}$ are the distribution functions of $\hat{z}$ and $\hat{z}'$, respectively. This ordering is also referred to as *first-order stochastic dominance*, which is explained in more detail in [MOA11, Chapter 17.A]. Since the term *risk* requires one to know what is good and what is bad, we assume throughout this section that the smaller the realizations of a random variable, the better. This is consistent with the previously mentioned convention of referring to real random variables as *loss-functions*.

The following definition summarizes some important features that can be associated with a risk measure.

Definition 2.5.8 (Properties of Risk Measures)

Let $p \in [1, \infty)$ and $\mathcal{R} \colon L^p(\Xi, \mathcal{A}, \mathbb{P}) \to \mathbb{R} \cup \{-\infty\} \cup \{\infty\}$ be a proper risk measure. Then, $\mathcal{R}$ can have the following properties:

(i) Convexity: For all $\hat{z}, \hat{z}' \in L^p(\Xi, \mathcal{A}, \mathbb{P})$ and $t \in [0, 1]$ it holds that

$$\mathcal{R}\big(t\hat{z} + (1-t)\hat{z}'\big) \leq t\mathcal{R}(\hat{z}) + (1-t)\mathcal{R}(\hat{z}').$$

(ii) Monotonicity: If $\hat{z}, \hat{z}' \in L^p(\Xi, \mathcal{A}, \mathbb{P})$ and $\hat{z} \succcurlyeq \hat{z}'$, then

$$\mathcal{R}(\hat{z}) \geq \mathcal{R}(\hat{z}').$$

(iii) Translation equivariance: If $c \in \mathbb{R}$ and $\hat{z} \in L^p(\Xi, \mathcal{A}, \mathbb{P})$, then

$$\mathcal{R}(\hat{z} + c) = \mathcal{R}(\hat{z}) + c.$$

(iv) Positive homogeneity: If $t > 0$ and $\hat{z} \in L^p(\Xi, \mathcal{A}, \mathbb{P})$, then

$$\mathcal{R}(t\hat{z}) = t\mathcal{R}(\hat{z}).$$

Convexity implies that $\mathcal{R}\left(\frac{1}{2}\hat{z} + \frac{1}{2}\hat{z}'\right) \leq \frac{1}{2}\mathcal{R}(\hat{z}) + \frac{1}{2}\mathcal{R}(\hat{z}')$ for all $\hat{z}, \hat{z}' \in L^p(\Xi, \mathcal{A}, \mathbb{P})$, and together with positive homogeneity we have

$$\mathcal{R}(\hat{z} + \hat{z}') \leq \mathcal{R}(\hat{z}) + \mathcal{R}(\hat{z}') \tag{2.5.15}$$

for all $\hat{z}, \hat{z}' \in L^p(\Xi, \mathcal{A}, \mathbb{P})$. A risk measure satisfying this inequality is called *subadditive*. Since, conversely, subadditivity together with positive homogeneity imply convexity, the following definition is still consistent with the original one in [Art+99, Definition 2.4], where property (i) of Definition 2.5.8 is replaced by the subadditivity condition (2.5.15).

Definition 2.5.9 (Coherent Risk Measure)

Let $p \in [1, \infty)$ and $\mathcal{R} \colon L^p(\Xi, \mathcal{A}, \mathbb{P}) \to \mathbb{R} \cup \{-\infty\} \cup \{\infty\}$ be a proper risk measure. If $\mathcal{R}$ satisfies all conditions of Definition 2.5.8, then it is said to be a *coherent* risk measure.

As mentioned right after Definition 2.5.6, the VaR is not convex and therefore not a coherent risk measure. However, Pflug proved in [Pfl00, Proposition 2] that the CVaR does satisfy all properties of Definition 2.5.8, which is why we have the following result.

Proposition 2.5.10 (Coherence of CVaR)

The Conditional Value-at-Risk as in Definition 2.5.7 is for any probability level $\beta \in (0,1)$ a coherent risk measure in the sense of Definition 2.5.9.

Furthermore, since CVaR is a coherent risk measure and thus satisfies properties (i) and (ii) of Definition 2.5.8, it is also continuous and subdifferentiable [SDR14, Proposition 6.6]. Another property we will need for the discretization of the probability space in Section 5.1 is that CVaR_β can be written as the minimum of a convex function involving the expected value. We define $F_\beta \colon L^1(\Xi, \mathcal{A}, \mathbb{P}) \times \mathbb{R} \to \mathbb{R}$ by

$$(\hat{z}, \alpha) \mapsto F_\beta(\hat{z}, \alpha) := \alpha + \frac{1}{1-\beta}\mathbb{E}\left((\hat{z} - \alpha)^+\right), \tag{2.5.16}$$

where $(x)^+ := \max\{0, x\}$ for all $x \in \mathbb{R}$. The following theorem stems from [RU00, Theorem 1] and shows the mentioned relationship between CVaR_β and F_β.

Proposition 2.5.11 (Characterization of CVaR)

Let $\beta \in (0,1)$ be a probability level.

(i) For fixed $\hat{z} \in L^1(\Xi, \mathcal{A}, \mathbb{P})$, the function $F_\beta(\hat{z}, \cdot) : \mathbb{R} \to \mathbb{R}$ as defined above is convex and continuously differentiable.

(ii) The CVaR_β of any loss-function $\hat{z} \in L^1(\Xi, \mathcal{A}, \mathbb{P})$ can be determined by the formula

$$\mathrm{CVaR}_\beta(\hat{z}) = \min_{\alpha \in \mathbb{R}} F_\beta(\hat{z}, \alpha).$$

(iii) The set consisting of the values of α for which the minimum in (ii) is attained, namely

$$A_\beta(\hat{z}) := \arg\min_{\alpha \in \mathbb{R}} F_\beta(\hat{z}, \alpha)$$

with $\hat{z} \in L^1(\Xi, \mathcal{A}, \mathbb{P})$, is a non-empty, closed, and bounded interval. The VaR_β of the loss is given by

$$\mathrm{VaR}_\beta(\hat{z}) = \min A_\beta(\hat{z}).$$

(iv) The statements in (ii) and (iii) immediately imply the relation

$$\mathrm{CVaR}_\beta(\hat{z}) = F_\beta\left(\hat{z}, \mathrm{VaR}_\beta(\hat{z})\right)$$

for every $\hat{z} \in L^1(\Xi, \mathcal{A}, \mathbb{P})$.

The gain of this proposition is that we can now calculate the CVaR without having calculated the VaR first. Nonetheless, we are still able to obtain the VaR from (iii) if this should be necessary. The following lemma shows that the Fenchel conjugate of CVaR can be represented as the indicator function of a convex, closed and bounded set.

Lemma 2.5.12 (Fenchel Conjugate of CVaR)

Let $\beta \in (0,1)$ be a probability level. The Fenchel conjugate of $\mathrm{CVaR}_\beta \colon L^1(\Xi, \mathcal{A}, \mathbb{P}) \to \mathbb{R}$ is given by

$$\mathrm{CVaR}_\beta^* = \delta_\Delta,$$

where

$$\Delta := \left\{ v \in L^\infty(\Xi, \mathcal{A}, \mathbb{P}) \,\middle|\, 0 \le v \le (1-\beta)^{-1} \,\mathbb{P}\text{-a.s. and } \mathbb{E}(v) = 1 \right\}.$$

Proof: Since CVaR_β is a coherent risk measure (see Proposition 2.5.10), it follows from [SDR14, Theorem 6.5] that

$$\mathrm{CVaR}_\beta(z) = \sup_{v \in \Delta} \langle v, z \rangle_{\mathcal{V}} = \sup_{v \in \mathcal{V}} \left(\langle v, z \rangle_{\mathcal{V}} - \delta_\Delta(v) \right) \tag{2.5.17}$$

for all $z \in \mathcal{V}$, where Δ is defined in [SDR14, (6.76)] by

$$\Delta := \left\{ v \in L^\infty(\Xi, \mathcal{A}, \mathbb{P}) \,\middle|\, 0 \le v \le (1-\beta)^{-1} \,\mathbb{P}\text{-a.s. and } \mathbb{E}(v) = 1 \right\}. \tag{2.5.18}$$

Now, Definition 2.2.6 and (2.5.17) imply that $\mathrm{CVaR}_\beta = \delta_\Delta^*$. Furthermore, Δ is convex, closed, and bounded. Hence, it follows from [CV20, Lemma 2.5] that δ_Δ is convex and lower semi-continuous. Thus, we can apply Theorem 2.2.7 with equality, yielding

$$\mathrm{CVaR}_\beta^* = \delta_\Delta^{**} = \delta_\Delta, \tag{2.5.19}$$

which concludes the proof. □

2.6 Partial Differential Equations

Partial differential equations (PDEs) are often used for modeling complex systems in physics, engineering, chemistry, and finance, to name just a few disciplines. Popular applications include the heat equation (for modeling heat diffusion), Maxwell's equations (for modeling electric fields), Navier-Stokes equations (for modeling the motion of fluids), or the wave equation (for modeling the propagation of mechanical or electromagnetic waves). The optimization of problems with PDEs started in the 1970s with [Lio71] and has since then become a broad field of research. In this section, which is based on both [Hin+09, Chapter 1] and [Del15], we summarize the most important results needed for this thesis, especially in Chapter 6.

A typical problem considered in optimal control is

$$\min_{y \in \mathcal{Y},\, u \in \mathcal{U}} \quad J(y,u) \quad s.t. \quad e(y,u) = 0, \tag{2.6.1}$$

where $\mathcal{Y}$ and $\mathcal{U}$ are Banach spaces, $J \colon \mathcal{Y} \times \mathcal{U} \to \mathbb{R}$ is the objective function, and the operator $e \colon \mathcal{Y} \times \mathcal{U} \to \mathcal{Z}$ with a Banach space $\mathcal{Z}$ represents a PDE or a system of PDEs. The variable $u \in \mathcal{U}$ is usually referred to as *control* variable, since it represents the parameter determining the PDE to be solved. The so-called *state* variable $y \in \mathcal{Y}$ describes the state of the considered PDE system (e.g. the distribution of heat). If, for each $u \in \mathcal{U}$, there exists a unique solution $y =: S(u)$ to $e(y,u) = 0$, we can also consider the reduced form of (2.6.1), which reads

$$\min_{u \in \mathcal{U}} \quad J\big(S(u), u\big) \quad s.t. \quad e\big(S(u), u\big) = 0. \tag{2.6.2}$$

The existence of such a unique solution can be shown, for example, if $(\bar{y}, \bar{u}) \in \mathcal{Y} \times \mathcal{U}$ is a local optimal solution to (2.6.1) and the partial derivative $e_y(\bar{y}, \bar{u})$ is a bijection. Then, the implicit function theorem [Hin+09, Theorem 1.41] yields the existence of a unique $y =: S(u)$ such that $e\big(S(u), u\big) = 0$ for all u in a neighborhood of $\bar{u}$. Furthermore, the solution mapping S is then continuously differentiable.

The theoretical framework used to examine the solution of such PDEs is based on functional analysis and so-called *Sobolev spaces*, which we introduce after the necessary preparation. Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ be an open subset, and $p \in [1, \infty)$. We consider the measure space $(\Omega, \mathcal{B}, \mu)$, where $\mathcal{B} \subset 2^\Omega$ is the Borel σ-algebra and $\mu \colon \mathcal{B} \to [0, \infty]$ the Lebesgue measure. First, similarly to Definition 2.4.2, we define the set of locally integrable functions on Ω as the factor space

$$L^p_{\text{loc}}(\Omega) := \mathcal{L}^p_{\text{loc}}(\Omega) / \sim, \tag{2.6.3}$$

where $u \sim w :\Leftrightarrow \| u - w \|_p = 0$ for all $u, w \in \mathcal{L}^p(\Omega)$ and

$$\mathcal{L}^p_{\text{loc}}(\Omega) := \left\{ u \colon \Omega \to \mathbb{R} \;\middle|\; u \text{ is } \mu\text{-measurable and } u \in \mathcal{L}^p(K) \text{ for all compact } K \subset \Omega \right\}. \tag{2.6.4}$$

One can see that $L^p(\Omega) \subset L^1_{\text{loc}}(\Omega)$ for all $p \in [1, \infty]$. Furthermore, let $C(\Omega)$ denote the space of all continuous functions from Ω to $\mathbb{R}$. For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with order $|\alpha| := \sum_{i=1}^d \alpha_i$, we define the $|\alpha|$-th order partial derivative of $u \colon \Omega \to \mathbb{R}$ at $x \in \Omega$ as

$$D^\alpha u(x) := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} u(x). \tag{2.6.5}$$

With this, we can define the space of all functions with continuous k-th order derivatives as

$$C^k(\Omega) := \{ u \in C(\Omega) \mid D^\alpha u \in C(\Omega) \text{ for } |\alpha| \leq k \}, \tag{2.6.6}$$

as well as

$$C^k(\overline{\Omega}) := \left\{ u \in C^k(\Omega) \;\middle|\; D^\alpha u \text{ has a continuous extension to } \overline{\Omega} \text{ for } |\alpha| \leq k \right\} \tag{2.6.7}$$

for all $k \in \mathbb{N}_0 \cup \{\infty\}$. If we use $\operatorname{supp}(u) := \operatorname{cl}(\{x \in \Omega \mid u(x) \neq 0\})$ to denote the support of $u \colon \Omega \to \mathbb{R}$, we can define the set

$$C_c^\infty(\Omega) := \left\{ u \in C^\infty(\overline{\Omega}) \;\middle|\; \operatorname{supp}(u) \subset \Omega \text{ is compact} \right\}, \tag{2.6.8}$$

which is dense in $L^p(\Omega)$ for all $p \in [1, \infty)$ [Hin+09, Lemma 1.4]. With this space we can now present the concept of *weak derivatives*, which generalize differentiation and are crucial for the definition of Sobolev spaces.

Definition 2.6.1 (Weak Derivative)

Let $u \in L^1_{\text{loc}}(\Omega)$ and $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ be a multi-index. If there exists a function $w \in L^1_{\text{loc}}(\Omega)$ such that

$$\int_\Omega u D^\alpha \varphi \, \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega w \varphi \, \mathrm{d}x$$

for all $\varphi \in C_c^\infty(\Omega)$, then $D^\alpha u := w$ is called the $|\alpha|$-th order weak partial derivative of u.

Note that the equation in Definition 2.6.1 determines the weak derivative $D^\alpha u \in L^1_{\text{loc}}(\Omega)$ uniquely [Hin+09, Lemma 1.5]. With these preparations we can now define the Sobolev space.

Definition 2.6.2 (Sobolev Space)

For $k \in \mathbb{N}_0$ and $p \in [1, \infty]$, we define the Sobolev space $W^{k,p}(\Omega)$ by

$$W^{k,p}(\Omega) := \left\{ u \in L^p(\Omega) \;\middle|\; u \text{ has weak derivatives } D^\alpha u \in L^p(\Omega) \text{ for all } |\alpha| \leq k \right\},$$

equipped with the norm

$$\| u \|_{W^{k,p}} := \left(\sum_{|\alpha| \leq k} \| D^\alpha u \|_p^p \right)^{1/p}, \quad \text{if } p \in [1, \infty),$$

$$\text{and} \quad \| u \|_{W^{k,\infty}} := \sum_{|\alpha| \leq k} \| D^\alpha u \|_\infty .$$

It can be shown that $W^{k,p}(\Omega)$ as defined above is indeed a Banach space [Hin+09, Theorem 1.11] for all $p \in [1, \infty]$. Moreover,

$$H^k(\Omega) := W^{k,2}(\Omega) \tag{2.6.9}$$

is a Hilbert space, which is why the space $H^1(\Omega)$ will be of special interest in this thesis. It consists of all $L^2(\Omega)$-functions with first order weak derivatives in $L^2(\Omega)$, i.e.

$$H^1(\Omega) = \left\{ y \in L^2(\Omega) \;\middle|\; D^i y \in L^2(\Omega) \text{ for all } i \in \{1, \dots, d\} \right\}. \tag{2.6.10}$$

In order to incorporate homogeneous boundary conditions into the function space, we denote by $W_0^{k,p}(\Omega)$ the closure of $C_c^\infty(\Omega)$ in $W^{k,p}(\Omega)$. If equipped with the same norm as $W^{k,p}(\Omega)$, this is a Banach space, too. Furthermore, the space

$$H_0^k(\Omega) := W_0^{k,2}(\Omega) \tag{2.6.11}$$

is a Hilbert space.

Since $W^{k,p}(\Omega) \subset L^p(\Omega)$ for all $k \in \mathbb{N}_0$ and $p \in [1, \infty]$, one can show under additional assumptions the embedding of Sobolev spaces into L^p-spaces. The following theorem stems from [Hin+09, Theorem 1.14] and is limited to one special case, which is relevant for this thesis.

Theorem 2.6.3 (Sobolev Embedding)

Let $\Omega \subset \mathbb{R}^d$ with $d \in \mathbb{N}$ be open and bounded with Lipschitz-boundary $\partial\Omega$. Then $H^1(\Omega) \hookrightarrow\hookrightarrow L^2(\Omega)$, i.e. $H^1(\Omega)$ is compactly embedded in $L^2(\Omega)$, which means that

(i) $\| y \|_{L^2(\Omega)} \leq \| y \|_{H^1(\Omega)}$ for all $y \in H^1(\Omega)$,

(ii) the embedding operator $\iota \colon H^1(\Omega) \to L^2(\Omega)$ is compact.

The requirement that Ω has a Lipschitz-boundary ensures that $\partial\Omega$ is *sufficiently regular* in the sense that it is locally the graph of a Lipschitz continuous function. For a proper definition, see [Hin+09, Definition 1.13]. The reason why we consider Sobolev spaces when it comes to the solution of PDEs is that the so-called *weak solutions* of PDEs are elements of these spaces, which

we illustrate with the following example. Let $\Omega \subset \mathbb{R}^d$ with $d \in \mathbb{N}$ be open and bounded, and $f \in L^2(\Omega)$. The *Poisson equation*

$$\begin{aligned} -\Delta y &= f \quad \text{on } \Omega, \\ y &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{2.6.12}$$

where Δ denotes the Laplace operator and $\partial\Omega$ the boundary of Ω, is an elliptic boundary value problem. A unique solution $y \in C^2(\Omega) \cap C^1(\overline{\Omega})$ can be obtained by classical methods [Eva10, Theorem 1] only for continuous right-hand sides. Therefore, we need a generalized solution concept, which is based on a variational formulation of (2.6.12). If we assume that y is *regular enough*, we can multiply (2.6.12) with a test function $v \in C_c^\infty(\Omega)$ and integrate over Ω to get

$$-\int_\Omega \Delta y v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x. \tag{2.6.13}$$

Integration by parts yields

$$\int_\Omega \nabla y \cdot \nabla v \, \mathrm{d}x - \int_{\partial\Omega} v \partial_n y \, \mathrm{d}S(x) = \int_\Omega f v \, \mathrm{d}x, \tag{2.6.14}$$

where S denotes the surface measure on $\partial\Omega$, $n \colon \partial\Omega \to \mathbb{R}^d$ is the exterior normal vector, and $\partial_n y := \nabla y \cdot n = \frac{\partial y}{\partial n}$ the normal derivative of y. Since $v \in C_c^\infty(\Omega)$ and Ω is open, we have that $\operatorname{supp}(v) \cap \partial\Omega = \emptyset$, i.e. $v(x) = 0$ for all $x \in \partial\Omega$. Therefore, (2.6.14) is equivalent to

$$\int_\Omega \nabla y \cdot \nabla v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x, \tag{2.6.15}$$

and since $C_c^\infty(\Omega)$ is dense in $H_0^1(\Omega)$ and the terms in (2.6.15) are continuous with respect to the $H_0^1(\Omega)$-norm, the equation holds for all $v \in H_0^1(\Omega)$ [Del15, p. 14]. This justifies the following definition.

Definition 2.6.4 (Weak Solution)

A function $y \in H_0^1(\Omega)$ is called a weak solution to (2.6.12) if it satisfies for all $v \in H_0^1(\Omega)$ the weak formulation

$$\int_\Omega \nabla y \cdot \nabla v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x.$$

The existence and uniqueness of a *weak solution* can be shown by the use of the Lax-Milgram lemma [Hin+09, Lemma 1.8].

Chapter 3

Optimization Problem

In this chapter, we present the risk-averse, non-smooth optimization problem that we aim at solving by the use of the algorithm developed in Chapter 4. The first section is devoted to the formulation of the problem. In the second section, we take a closer look at the PDE constraint and how it is incorporated into the objective function. Finally, an optimality condition is derived in the third section.

3.1 Problem Formulation

The problem we consider is basically a special case of the one proposed by Kouri and Surowiec in [KS18a], which thus is the basis of this section. Let

$$\mathcal{U} := L^2(\Omega) := L^2(\Omega, \mathcal{O}, \mu) \tag{3.1.1}$$

be the space of control variables where $(\Omega, \mathcal{O}, \mu)$ is a separable measure space with $\Omega \subset \mathbb{R}^d$ open and bounded with Lipschitz-boundary $\partial\Omega$, and $d \in \mathbb{N}$ (typically, $\mathcal{O}$ will be the Borel σ-algebra of Ω and μ the Lebesgue measure). Furthermore, we consider a separable probability space $(\Xi, \mathcal{A}, \mathbb{P})$ and the space of random variables (i.e. loss functions)

$$\mathcal{V} := L^2_{\mathbb{P}}(\Xi) := L^2(\Xi, \mathcal{A}, \mathbb{P}). \tag{3.1.2}$$

Both $\mathcal{U}$ and $\mathcal{V}$ are Hilbert spaces with inner products $\langle \cdot, \cdot \rangle_{\mathcal{U}}$ and $\langle \cdot, \cdot \rangle_{\mathcal{V}}$. The induced norms are denoted by $\| \cdot \|_{\mathcal{U}}$ and $\| \cdot \|_{\mathcal{V}}$, respectively.

We first state the problem and then define the involved functions.

Problem 3.1.1 (Original Problem)

$$\min_{u \in \mathcal{U}} \quad \mathrm{CVaR}_\beta(K(u)) + G(u)$$

CVaR_β denotes the Conditional Value-at-Risk for a probability level $\beta \in (0, 1)$ as in Definition 2.5.7. Remember that, according to Proposition 2.5.10, CVaR_β is coherent and therefore especially convex. Moreover, it is proper and continuous as mentioned in Section 2.5.4, and thus also lower semi-continuous. The function $K \colon \mathcal{U} \to \mathcal{V}$ involves a random field PDE solution, which is explained in more detail in Section 3.2. It is therefore the link between the constraint and the objective function. The following assumption summarizes the properties of G.

Assumption 3.1.2 (Properties of G)

The mapping $G\colon \mathcal{U} \to \overline{\mathbb{R}}$ is proper, convex, and lower semi-continuous. Furthermore, its domain is bounded, i.e. $\operatorname{dom}(G) \subset \mathbb{B}_{R_G}(\tilde{u})$ for some $\tilde{u} \in \mathcal{U}$ and a radius $R_G > 0$.

Finally, we present the saddle-point formulation of Problem 3.1.1, which is often the basis of similar analyses, as mentioned in Chapter 1. Since CVaR_β is proper, convex, and lower semi-continuous, the Fenchel biconjugate of CVaR_β (see Definition 2.2.6 (ii)) is, according to Theorem 2.2.7, equal to CVaR_β itself, i.e.

$$\mathrm{CVaR}_\beta(v) = \mathrm{CVaR}_\beta^{**}(v) = \sup_{v^* \in \mathcal{V}^*} \left(\langle v^*, v \rangle_{\mathcal{V}^*,\mathcal{V}} - \mathrm{CVaR}_\beta^*(v^*) \right) \tag{3.1.3}$$

for all $v \in \mathcal{V}$, where $\mathrm{CVaR}_\beta^*\colon \mathcal{V}^* \to \overline{\mathbb{R}}$ is the Fenchel conjugate of CVaR_β and $\langle \cdot, \cdot \rangle_{\mathcal{V}^*,\mathcal{V}}$ the duality pairing. Together with the Riesz-Fréchet representation theorem [Bre11, Theorem 5.5], we can therefore equivalently formulate Problem 3.1.1 as follows.

Problem 3.1.3 (Saddle-Point Problem)

$$\min_{u \in \mathcal{U}} \quad \max_{v \in \mathcal{V}} \quad G(u) + \langle v, K(u) \rangle_{\mathcal{V}} - \mathrm{CVaR}_\beta^*(v)$$

Note that the justification for using the maximum instead of the supremum follows from the representation of CVaR_β^* in (2.5.19) together with [CV20, Theorem 3.8] and the fact that the set Δ in (2.5.18) is non-empty, convex, bounded and closed.

3.2 PDE Constraint

The optimization problem involves a constraint in form of a partial differential equation with uncertain coefficients. This PDE can be written as

$$e(y, u) = 0 \tag{3.2.1}$$

for all controls $u \in \mathcal{U}$. Here, $y := S(u)$ denotes the weak (random field) solution of this PDE. It is an element of the space $\mathcal{Y}$, which we define after introducing the following assumption from [KS18a, Assumption 2.1].

Assumption 3.2.1 (Properties of the Solution Mapping)

(i) $y := S(u)\colon \Xi \to H^1(\Omega)$ is strongly $\mathbb{P}$-measurable for all $u \in \mathcal{U}$.

(ii) There exist a non-negative, increasing function $\rho\colon [0, \infty) \to [0, \infty)$ and a non-negative random variable $C \in L^q_{\mathbb{P}}(\Xi)$ with $q \in [1, \infty]$ satisfying

$$\| S(u)(\xi) \|_{H^1(\Omega)} \leq \rho\left(\| u \|_{\mathcal{U}} \right) C(\xi)$$

for $\mathbb{P}$-a.e. $\xi \in \Xi$ and all $u \in \mathcal{U}$.

(iii) If $(u_n)_{\mathbb{N}} \subset \mathcal{U}$ is a sequence with $u_n \rightharpoonup u \in \mathcal{U}$ as $n \to \infty$, then $S(u_n) \rightharpoonup S(u)$ in $H^1(\Omega)$ $\mathbb{P}$-a.s.

This assumption ensures that $S(u) \in \mathcal{Y}$ for all $u \in \mathcal{U}$, where $\mathcal{Y}$ is the stochastic PDE solution space

$$\mathcal{Y} := L^q\left(\Xi, \mathcal{A}, \mathbb{P}; H^1(\Omega)\right) \tag{3.2.2}$$

with $q \in [1, \infty]$ as in Assumption 3.2.1 (ii). Without loss of generality, we can assume $q \in \mathbb{N}$. Otherwise we could replace $\mathcal{Y}$ by $L^r\left(\Xi, \mathcal{A}, \mathbb{P}; H^1(\Omega)\right)$ for any integer $r < q$, since $(\Xi, \mathcal{A}, \mathbb{P})$ is a finite measure space and therefore the inclusion $L^q\left(\Xi, \mathcal{A}, \mathbb{P}; H^1(\Omega)\right) \hookrightarrow L^r\left(\Xi, \mathcal{A}, \mathbb{P}; H^1(\Omega)\right)$ holds. Now, as mentioned in [KS18a, p. 790], Assumption 3.2.1 ensures that $S(u) \in \mathcal{Y}$ for all $u \in \mathcal{U}$, thus $S\colon \mathcal{U} \to \mathcal{Y}$ is well-defined. Furthermore, the assumption implies that $S(u_n) \rightharpoonup S(u)$ in $\mathcal{Y}$ for any weakly convergent sequence $(u_n)_{n\in\mathbb{N}} \subset \mathcal{U}$ with $u_n \rightharpoonup u$ as $n \to \infty$. This is a different notion of convergence than in Assumption 3.2.1 (iii) and implies that S is weakly continuous.

In the following, we examine how the function $K\colon \mathcal{U} \to \mathcal{V}$ incorporates the solution of the PDE constraint into the objective function. For this purpose, let

$$\tilde{J}\colon H^1(\Omega) \times \Xi \to \mathbb{R} \tag{3.2.3}$$

satisfy the following assumption [KS18a, Assumption 3.1].

Assumption 3.2.2 (Properties of $\tilde{J}$)

(i) $\tilde{J}$ is a Carathéodory function, i.e. $\tilde{J}(\cdot, \xi)$ is continuous for $\mathbb{P}$-a.e. $\xi \in \Xi$ and $\tilde{J}(y, \cdot)$ is measurable for all $y \in H^1(\Omega)$.

(ii) Growth condition: If $q \in [1, \infty)$, then there exist $a \in L^2(\Xi)$ with $a \geq 0$ $\mathbb{P}$-a.s. and $c > 0$ such that

$$\left|\tilde{J}(y, \xi)\right| \leq a(\xi) + c \,\|\, y \,\|_{H^1(\Omega)}^{q/2}$$

for $\mathbb{P}$-a.e. $\xi \in \Xi$ and all $y \in \mathcal{Y}$. If $q = \infty$, then the uniform boundedness condition holds: for all $c > 0$ there exists $\gamma \in \mathcal{V}$ such that

$$\left|\tilde{J}(y, \xi)\right| \leq \gamma(\xi)$$

for $\mathbb{P}$-a.e. $\xi \in \Xi$ and all $y \in \mathcal{Y}$ with $\| y \|_{\mathcal{Y}} \leq c$.

(iii) Convexity: $\tilde{J}(\cdot, \xi)$ is convex for $\mathbb{P}$-a.e. $\xi \in \Xi$.

An example satisfying this assumption is the tracking-type functional

$$\tilde{J}(y, \xi) := \frac{1}{2} \| y - \bar{y} \|_{L^2(\Omega)}^2 \tag{3.2.4}$$

for some $\bar{y} \in L^2(\Omega)$ [KS18a, Example 3.2]. For example, if $q = 4$, i.e. $\mathcal{Y} = L^4\left(\Xi, \mathcal{A}, \mathbb{P}; H^1(\Omega)\right)$, then

$$\begin{aligned} \tilde{J}(y, \xi) &= \frac{1}{2} \| y \|_{L^2(\Omega)}^2 - \langle y, \bar{y} \rangle_{L^2(\Omega)} + \frac{1}{2} \| \bar{y} \|_{L^2(\Omega)}^2 \\ &\leq \| y \|_{L^2(\Omega)}^2 + \| \bar{y} \|_{L^2(\Omega)}^2 \\ &\leq \| y \|_{H^1(\Omega)}^2 + \| \bar{y} \|_{L^2(\Omega)}^2 \end{aligned} \tag{3.2.5}$$

for all $y \in H^1(\Omega)$ and $\xi \in \Xi$, where the first inequality follows from Cauchy-Schwarz and Young's inequality, and the second from Theorem 2.6.3 (Sobolev Embedding). In order to satisfy

Assumption 3.2.2 (ii), we can set $c := 1$ and $a(\xi) := \| \bar{y} \|^2_{L^2(\Omega)}$ for all $\xi \in \Xi$, which implies $a \in L^\infty(\Xi) \subset L^2(\Xi) = \mathcal{V}$. Obviously, Assumption 3.2.2 (i) and (iii) are satisfied for this choice of $\tilde{J}$ as well.

We can now define for any function $\tilde{J}\colon H^1(\Omega) \times \Xi \to \mathbb{R}$ the so-called *Nemytskii* or *superposition operator* $J\colon \mathcal{Y} \to \mathcal{V}$ by $J(y) := \tilde{J}(y(\cdot), \cdot)$ for all $y \in \mathcal{Y}$, i.e.

$$[J(y)](\xi) := \tilde{J}(y(\xi), \xi) \tag{3.2.6}$$

for $\mathbb{P}$-a.e. $\xi \in \Xi$. It can be shown that, under Assumption 3.2.2, J is continuous [KS18a, Theorem 3.5] and Gâteaux differentiable [KS18a, Theorem 3.9]. Furthermore, the following proposition shows that, under additional assumptions, J is even continuously Fréchet differentiable [GKT92, Theorem 7].

Proposition 3.2.3 (Continuous Differentiability of J)

Let q from the definition of $\mathcal{Y}$ in (3.2.2) satisfy $q \in (2, \infty)$. If $\tilde{J}\colon H^1(\Omega) \times \Xi \to \mathbb{R}$ is Fréchet differentiable with respect to $y \in H^1(\Omega)$ and the Fréchet derivative $\tilde{J}_y\colon H^1(\Omega) \times \Xi \to \mathbb{L}\left(H^1(\Omega), \mathbb{R}\right)$ is a Carathéodory function and satisfies the growth condition

$$\left\| \tilde{J}_y(y, \xi) \right\|_{\mathbb{L}\left(H^1(\Omega), \mathbb{R}\right)} \leq a(\xi) + c \| y \|^{q/r}_{H^1(\Omega)}$$

for some $a \in L^r(\Xi)$, $c \geq 0$, and $r := \frac{2q}{q-2}$, then the Nemytskii operator $J\colon \mathcal{Y} \to \mathcal{V}$ defined by

$$y \mapsto J(y) := \tilde{J}(y(\cdot), \cdot)$$

is continuously Fréchet differentiable.

If we revisit the example given in (3.2.4), we note that the assumptions of Proposition 3.2.3 can be easily verified with $q = 4$. Hence, in this case, J is in fact continuously Fréchet differentiable.

With this preparation we can now define the operator K, which was introduced in Problem 3.1.1. Let $K := J \circ S\colon \mathcal{U} \to \mathcal{V}$ with $J\colon \mathcal{Y} \to \mathcal{V}$ as defined in (3.2.6) and $S\colon \mathcal{U} \to \mathcal{Y}$ the PDE solution mapping. An assumption, which is needed in Section 3.3 to derive the optimality condition, is the continuous differentiability of K.

Assumption 3.2.4 (Continuous Differentiability of K)

The mapping $K := J \circ S\colon \mathcal{U} \to \mathcal{V}$ is at least once continuously Fréchet differentiable.

3.3 Existence and Optimality Condition

In this section, we show under which assumptions Problem 3.1.1 admits a solution within the domain of the objective function. Afterwards, we derive an optimality condition, similarly to the approach presented in the Master's thesis [Ang18, Section 3.2]. The difference is that now the formulations are more generic since we do neither require K to be linear nor $\mathcal{U}$ to have finite dimension.

Recall that the problem we consider is

$$\min_{u \in \mathcal{U}} \quad \mathrm{CVaR}_\beta(K(u)) + G(u). \tag{3.3.1}$$

In order to show the existence of a minimizer, we have to make sure that the conditions of Theorem 2.2.9 are satisfied. As it turns out in the following proposition, the assumptions made in Section 3.1 are sufficient.

Proposition 3.3.1 (Existence of Solution)

If the assumptions of Section 3.1 are satisfied, then Problem 3.1.1 has a solution

$$\bar{u} \in \mathrm{dom}\left(\mathrm{CVaR}_\beta \circ K + G\right).$$

Proof: We want to apply Theorem 2.2.9 and thus have to satisfy its conditions, i.e. the function $F := \mathrm{CVaR}_\beta \circ K + G \colon \mathcal{U} \to \overline{\mathbb{R}}$ needs to be proper, coercive and weakly lower semi-continuous.

Since $\mathrm{CVaR}_\beta(v) \in \mathbb{R}$ for all $v \in L^1_{\mathbb{P}}(\Xi)$ [SDR14, Section 6.2.4], and G is proper according to Assumption 3.1.2, the compound function F is proper as well.

In order to show the coercivity of F, let $(u_n)_{n\in\mathbb{N}} \subset \mathcal{U}$ be a sequence with $\| u_n \|_{\mathcal{U}} \to \infty$ as $n \to \infty$. The boundedness of $\mathrm{dom}\,(G)$ (see Assumption 3.1.2) implies that there exists an index $N \in \mathbb{N}$ such that $G(u_n) = \infty$ for all $n \geq N$. Since CVaR_β is finite-valued, this implies that $F(u_n) = \infty$ for all $n \geq N$. Therefore, by taking the limit $n \to \infty$, we get $F(u_n) \to \infty$ and hence the coercivity of F.

Finally, [KS18a, Proposition 3.8] shows that F is weakly lower semi-continuous, since Assumption 3.2.1 and Assumption 3.2.2 are satisfied. □

In order to derive the optimality condition, we can apply Theorem 2.3.4 (Fermat Principle) and deduce that, if $\bar{u} \in \mathrm{dom}\left(\mathrm{CVaR}_\beta \circ K + G\right)$ is a local minimizer, it satisfies

$$0 \in \partial_C \left(\mathrm{CVaR}_\beta \circ K + G\right)(\bar{u}). \tag{3.3.2}$$

We know from Assumption 3.2.4 that K is continuously Fréchet differentiable. Furthermore, since CVaR_β is convex and lower semi-continuous, it is also locally Lipschitz continuous on $\mathrm{int}\left(\mathrm{dom}\left(\mathrm{CVaR}_\beta\right)\right)$, according to Proposition 2.3.2, and therefore regular at $K(\bar{u})$. Application of the second part of Theorem 2.3.6 (Chain Rule) yields that $\mathrm{CVaR}_\beta \circ K$ is regular as well. On the other hand, G is also regular for the same reason, which is why we can use Theorem 2.3.5 (Sum Rule) with equality, i.e.

$$0 \in \partial_C \left(\mathrm{CVaR}_\beta \circ K\right)(\bar{u}) + \partial_C G(\bar{u}). \tag{3.3.3}$$

Consequently, we can now apply the whole Theorem 2.3.6 (Chain Rule) with equality, yielding

$$0 \in K'(\bar{u})^* \partial_C \mathrm{CVaR}_\beta\left(K(\bar{u})\right) + \partial_C G(\bar{u}). \tag{3.3.4}$$

If this condition is satisfied, there must be a $\bar{v} \in \mathcal{V}$ such that

$$\begin{cases} -K'(\bar{u})^* \bar{v} \in \partial G(\bar{u}), \\ \bar{v} \in \partial \mathrm{CVaR}_\beta\left(K(\bar{u})\right), \end{cases} \tag{3.3.5}$$

where we have intentionally replaced ∂_C by ∂ due to the convexity and lower semi-continuity of CVaR_β and G; see [CV20, Theorem 13.8]. Now we can apply Lemma 2.2.8 to the second condition in (3.3.5) to get the equivalent formulation

$$\begin{cases} -K'(\bar{u})^* \bar{v} \in \partial G(\bar{u}), \\ K(\bar{u}) \in \partial \mathrm{CVaR}_\beta^*(\bar{v}), \end{cases} \tag{3.3.6}$$

where CVaR_β^* is the Fenchel conjugate of CVaR_β as in Definition 2.2.6. Altogether, we have proven the following optimality condition.

Proposition 3.3.2 (Optimality Condition)

If $\bar{u} \in \mathrm{dom}\left(\mathrm{CVaR}_\beta \circ K + G\right)$ is a local minimizer of Problem 3.1.1, then there exists $\bar{v} \in \mathcal{V}$ such that $\overline{w} := (\bar{u}, \bar{v})$ satisfies

$$0 \in H(\overline{w}) := \begin{pmatrix} \partial G(\bar{u}) + K'(\bar{u})^* \bar{v} \\ \partial \mathrm{CVaR}_\beta^*(\bar{v}) - K(\bar{u}) \end{pmatrix}. \tag{OC}$$

Chapter 4

Algorithm

In this chapter we propose a stochastic algorithm that solves Problem 3.1.1. In the first section we describe the motivation, define the necessary notation and present the algorithm in a primal-dual and an implicit form. The former is used for implementation whereas the latter will be the basis of the convergence analysis. In the second section we formally introduce a randomization technique which will accelerate the numerical computation. Based on this randomized algorithm, we prove almost sure weak convergence under some additional assumptions in Section 4.3. Finally, we investigate in the fourth section how these assumptions can be satisfied in the case of scalar and deterministic step sizes.

4.1 Stochastic Primal-Dual Proximal Splitting Method

In order to motivate the structure of the algorithm, we derive an equivalent formulation of the optimality condition (OC), which involves proximal operators. We do this by applying Lemma 2.2.15 to (3.3.6) and therefore have to prove its assumptions, i.e. G and CVaR_β^* need to be proper, convex, and lower semi-continuous. However, G has this properties by Assumption 3.1.2 and CVaR_β^* is the Fenchel Conjugate of a proper function and as such convex and lower semi-continuous [BC17, Proposition 13.13]. Moreover, since CVaR_β is convex and lower semi-continuous, CVaR_β^* is proper [BC17, Theorem 13.37]. This justifies the application of Lemma 2.2.15 to (3.3.6), yielding

$$\begin{cases} \bar{u} = \mathrm{prox}_G^T\left(\bar{u} - TK'(\bar{u})^*\bar{v}\right) \\ \bar{v} = \mathrm{prox}_{\mathrm{CVaR}_\beta^*}^\Sigma\left(\bar{v} + \Sigma K\left(\bar{u}\right)\right) \end{cases} \tag{4.1.1}$$

with self-adjoint, strongly monotone operators $T \in \mathbb{L}\left(\mathcal{U}, \mathcal{U}\right)$ and $\Sigma \in \mathbb{L}\left(\mathcal{V}, \mathcal{V}\right)$. This suggests the fixed-point iteration

$$\begin{cases} v_{k+1} = \mathrm{prox}_{\mathrm{CVaR}_\beta^*}^{\Sigma_k}\left(v_k + \Sigma_k K\left(u_k\right)\right) \\ u_{k+1} = \mathrm{prox}_G^{T_k}\left(u_k - T_k K'(u_k)^* v_{k+1}\right) \end{cases} \tag{4.1.2}$$

where we have used u_k instead of u_{k+1} in the first line in order to decouple the dual from the primal update. This allows us to compute v_{k+1} without knowing u_{k+1} in advance, thus we can first update the dual and then the primal variable. This order was chosen deliberately since, in this way, we can reduce the number of computationally demanding PDE solutions by a factor of up to two. This becomes clearer after introducing the following concept.

In order to accelerate the algorithm, we will make use of what we call *component-wise gradient freezing* or *CGF*. This term stems from the discrete case in Section 6.2, where in each iteration

only some components of the gradient of K are updated (and therefore the PDEs only need to be updated for a subset of samples). In the more general framework, we need to consider projections onto subsets instead of components of a discrete gradient. Therefore, we define for every element[1] $A \in \mathcal{A}$ (i.e. A is a subset of Ξ where $(\Xi, \mathcal{A}, \mathbb{P})$ is the probability space) the function $\Pi_A \colon \mathcal{V} \to \mathcal{V}$ as

$$\Pi_A(v) := \chi_A(\cdot) v(\cdot) \qquad \text{for all } v \in \mathcal{V}, \tag{4.1.3}$$

$$\text{i.e.} \quad \Pi_A(v)(\xi) = \chi_A(\xi) v(\xi) \qquad \text{for all } v \in \mathcal{V}, \xi \in \Xi, \tag{4.1.4}$$

where $\chi_A \colon \Xi \to \{0, 1\}$ is the characteristic function of A. We have that $\Pi_A \in \mathbb{L}(\mathcal{V}, \mathcal{V})$ for every $A \in \mathcal{A}$ and especially $\Pi_A = \mathrm{Id}_{\mathcal{V}}$ if $A = \Xi$. Furthermore, given a sequence $(A_k)_{k \in \mathbb{N}} \subset \mathcal{A}$ of random events and a sequence $(u_k)_{k \in \mathbb{N}_0} \subset \mathcal{U}$ of primal iterates, we recursively define for every $k \in \mathbb{N}$ the functions $\hat{K}_k \colon \mathcal{U} \to \mathcal{V}$ and $\hat{K}_k'(\cdot)^* \colon \mathcal{U} \to \mathbb{L}(\mathcal{V}, \mathcal{U})$ by

$$\hat{K}_k(u) := \Pi_{A_k}\big(K(u)\big) + \big(\mathrm{Id}_{\mathcal{V}} - \Pi_{A_k}\big)\Big(\hat{K}_{k-1}(u_{k-1})\Big), \tag{4.1.5}$$

$$\text{and} \quad \hat{K}_k'(u)^* := K'(u)^* \circ \Pi_{A_k} + \hat{K}_{k-1}'(u_{k-1})^* \circ \big(\mathrm{Id}_{\mathcal{V}} - \Pi_{A_k}\big), \tag{4.1.6}$$

for every $u \in \mathcal{U}$, where $\hat{K}_0 := K$ and $\hat{K}_0'(\cdot)^* := K'(\cdot)^*$. Note that the special case without CGF (i.e. $A_k = \Xi$ for all $k \in \mathbb{N}$) implies that $\hat{K}_k \equiv K$ and $\hat{K}_k'(\cdot)^* \equiv K'(\cdot)^*$ for all $k \in \mathbb{N}$.

So as not to rely on some predefined, deterministic choice of the sequence $(A_k)_{k \in \mathbb{N}} \subset \mathcal{A}$, we want to allow for a random selection of the set A_k in each iteration $k \in \mathbb{N}$. This randomization is formalized in Section 4.2, whereas examples of the concrete choice of $(A_k)_{k \in \mathbb{N}} \subset \mathcal{A}$ are given in Section 6.2.

With this preparation, we can now replace K and K' in the fixed-point iterations (4.1.2) with $\hat{K}_k$ and $\hat{K}_k'$, respectively. Additionally, we add a so-called *over-relaxation step* in the dual variable, which corresponds to the case $\theta = 1$ in [CP11a, Section 3.1]. The reason for this step lies in the desired structure of the implicit formulation in (4.1.7). As we will see later in Lemma 4.4.8, this allows us to make the preconditioning operator M_k self-adjoint by the choice of Σ_k and T_k. Altogether, we can formulate Algorithm 4.1 with the following inputs:

- a stopping criterion tolerance of $\varepsilon > 0$;
- starting vectors $u_0 \in \mathcal{U}$ and $v_0 \in \mathcal{V}$;
- initial step size operators $T_0 \in \mathbb{L}(\mathcal{U}, \mathcal{U})$ and $\Sigma_0 \in \mathbb{L}(\mathcal{V}, \mathcal{V})$;
- an initial index set $A_0 = \Xi$.

The choice of $A_0 = \Xi$ guarantees that the recursive definition of $\hat{K}_k$ and $\hat{K}_k'$ (see (4.1.5) and (4.1.6), respectively) makes sense for the case $k = 0$. It also implies that, in the first iteration, the gradients are not frozen at all. This seems to be a reasonable start, since u_1 and v_1 will not be influenced by the randomization this way.

After reading Algorithm 4.1, it should be clear that, by updating the dual variable first, we save the computation of many PDEs, because A_k (which determines the subset of samples for which the PDEs need to be updated) changes right after the computation of u_k (which determines the PDEs themselves). Therefore, the PDE solution computed by calculating $\hat{K}_k(u_k)$ in Line 6 can be reused for the calculation of $\hat{K}_k'(u_k)^*$ in Line 8. Note that we have not yet specified how the step size operators are determined in practice. However, we show in Section 4.4 how they could be

[1] Due to the discrete case in Section 6.2, we also refer to $A \in \mathcal{A}$ as the *index set*.

chosen in the case of scalar and deterministic step sizes to guarantee convergence. In Section 6.1, we present a concrete choice of these step sizes in practice.

Algorithm 4.1 (Primal-Dual Formulation)

1 Initialize $k := 0$.
2 **repeat**
3 **if** $k \geq 1$ **then**
4 Randomly select $A_k \in \mathcal{A}$.
5 **end**
6 $v_{k+1} := \operatorname{prox}^{\Sigma_k}_{\mathrm{CVaR}^*_\beta}\left(v_k + \Sigma_k \hat{K}_k\left(u_k\right)\right)$
7 $\bar{v}_{k+1} := 2v_{k+1} - v_k$
8 $u_{k+1} := \operatorname{prox}^{T_k}_{G}\left(u_k - T_k \hat{K}'_k\left(u_k\right)^* \bar{v}_{k+1}\right)$
9 Determine step size operators $T_{k+1} \in \mathbb{L}\left(\mathcal{U}, \mathcal{U}\right)$ and $\Sigma_{k+1} \in \mathbb{L}\left(\mathcal{V}, \mathcal{V}\right)$.
10 Update $k \leftarrow k + 1$.
11 **until** $\| u_k - u_{k-1} \|_{\mathcal{U}} < \varepsilon$ **and** $\| v_k - v_{k-1} \|_{\mathcal{V}} < \varepsilon$

Due to the properties of CVaR_β and the assumptions from Chapter 3, we can already prove that the sequence of variables generated by Algorithm 4.1 is bounded.

Lemma 4.1.1 (Boundedness of Primal and Dual Variables)

Let the sequence $\left(\left(u_k, v_k\right)\right)_{k \in \mathbb{N}_0} \subset \mathcal{U} \times \mathcal{V}$ be generated by Algorithm 4.1. For all $\bar{u} \in \mathcal{U}$ and $\bar{v} \in \mathcal{V}$ there exist radii $R_{\mathcal{U}}, R_{\mathcal{V}} > 0$ such that

$$u_k \in \mathbb{B}_{R_{\mathcal{U}}}\left(\bar{u}\right) \quad \text{and} \quad v_k \in \mathbb{B}_{R_{\mathcal{V}}}\left(\bar{v}\right)$$

for all $k \in \mathbb{N}$.

Proof: Let $\bar{u} \in \mathcal{U}$ and $\bar{v} \in \mathcal{V}$. Since Assumption 3.1.2 states that the domain of G is bounded, we know that $\operatorname{dom}\left(G\right) \subset \mathbb{B}_{R_{\mathcal{U}}}\left(\bar{u}\right)$ for a radius $R_{\mathcal{U}} > 0$. Furthermore, it follows immediately from Definition 2.2.14 that $\operatorname{prox}^{T_k}_{G}\left(\mathcal{U}\right) \subset \operatorname{dom}\left(G\right)$ for all $k \in \mathbb{N}_0$. Thus, together with Line 8 of Algorithm 4.1, we conclude that $u_k \in \mathbb{B}_{R_{\mathcal{U}}}\left(\bar{u}\right)$ for all $k \in \mathbb{N}$.

Analogously, we can prove that there exist $R_{\mathcal{V}} > 0$ such that $v_k \in \mathbb{B}_{R_{\mathcal{V}}}\left(\bar{v}\right)$ for all $k \in \mathbb{N}$. The only condition we need to show for this is that the domain of CVaR^*_β (see Line 6 of Algorithm 4.1) is bounded as well. This property follows directly from the representation $\mathrm{CVaR}^*_\beta = \delta_\Delta$ with the bounded set Δ defined in Lemma 2.5.12, which implies that $\operatorname{dom}\left(\mathrm{CVaR}^*_\beta\right) = \operatorname{dom}\left(\delta_\Delta\right) = \Delta$ is indeed bounded. □

Note that, if the starting vectors u_0 and v_0 are chosen adequately, the assertion even holds for all $k \in \mathbb{N}_0$.

The proof of convergence of Algorithm 4.1 in Section 4.3 will be based on the proximal point method [CV20, Section 8.1]. Therefore, we need to equivalently reformulate the above algorithm to an implicit form. The goal is to write Lines 6 to 8 of Algorithm 4.1 like

$$0 \in W_k H\left(w_{k+1}\right) + D_k\left(w_{k+1}\right) + M_k\left(w_{k+1} - w_k\right), \tag{4.1.7}$$

for all $k \in \mathbb{N}_0$, where

- $w_k := (u_k, v_k) \in \mathcal{W} := \mathcal{U} \times \mathcal{V}$,
- $W_k \in \mathbb{L}(\mathcal{W}, \mathcal{W})$ is the step size operator,
- $H \colon \mathcal{W} \rightrightarrows \mathcal{W}$ stems from the optimality condition (OC),
- $D_k \colon \mathcal{W} \to \mathcal{W}$ is the so-called *discrepancy function*, and
- $M_k \in \mathbb{L}(\mathcal{W}, \mathcal{W})$ is the *preconditioning operator*.

We fix $k \in \mathbb{N}_0$ and start with Line 6 of Algorithm 4.1. This equation is, by Definition 2.2.14 of the proximal operator and equation (2.2.11), equivalent to

$$v_{k+1} = \left(\mathrm{Id}_{\mathcal{V}} + \Sigma_k \partial \mathrm{CVaR}_\beta^*\right)^{-1}\left(v_k + \Sigma_k \hat{K}_k\left(u_k\right)\right), \tag{4.1.8}$$

which can be rearranged to

$$\Sigma_k \partial \mathrm{CVaR}_\beta^*\left(v_{k+1}\right) - \Sigma_k \hat{K}_k\left(u_k\right) + v_{k+1} - v_k = 0. \tag{4.1.9}$$

We now add and subtract the term $\Sigma_k\left[K\left(u_{k+1}\right) - K\left(u_k\right) - K'\left(u_k\right)\left(u_{k+1} - u_k\right)\right]$ to get

$$\begin{aligned} \Sigma_k \overbrace{\left[\partial \mathrm{CVaR}_\beta^*\left(v_{k+1}\right) - K\left(u_{k+1}\right)\right]}^{\rightsquigarrow H} + \overbrace{\Sigma_k\left[K\left(u_{k+1}\right) - K\left(u_k\right) - K'\left(u_k\right)\left(u_{k+1} - u_k\right)\right]}^{\rightsquigarrow D_k^{LIN}} \\ + \underbrace{\Sigma_k\left[K\left(u_k\right) - \hat{K}_k\left(u_k\right)\right]}_{\rightsquigarrow D_k^{CGF}} + \underbrace{\Sigma_k K'\left(u_k\right)\left(u_{k+1} - u_k\right) + v_{k+1} - v_k}_{\rightsquigarrow M_k} = 0, \end{aligned} \tag{4.1.10}$$

The squiggly arrow annotations at the curly brackets indicate into which function definition the respective term will enter. Since these functions will be defined later, this annotations can be ignored for the moment.

We now consider Lines 7 to 8 of Algorithm 4.1. Analogously to (4.1.8), we have

$$u_{k+1} = \left(\mathrm{Id}_{\mathcal{U}} + T_k \partial G\right)^{-1}\left(u_k - T_k \hat{K}_k'\left(u_k\right)^*\left(2v_{k+1} - v_k\right)\right), \tag{4.1.11}$$

which is equivalent to

$$T_k \partial G\left(u_{k+1}\right) + T_k \hat{K}_k'\left(u_k\right)^*\left(2v_{k+1} - v_k\right) + u_{k+1} - u_k = 0. \tag{4.1.12}$$

Addition and subtraction of $T_k\left[K'\left(u_{k+1}\right)^* v_{k+1} + K'\left(u_k\right)^*\left(2v_{k+1} - v_k\right)\right]$ yields

$$T_k \overbrace{\left[\partial G\left(u_{k+1}\right) + K'\left(u_{k+1}\right)^* v_{k+1}\right]}^{\rightsquigarrow H} + \overbrace{T_k\left[K'\left(u_k\right)^* - K'\left(u_{k+1}\right)^*\right] v_{k+1}}^{\rightsquigarrow D_k^{LIN}}$$
$$+ \underbrace{T_k\left[\hat{K}_k'\left(u_k\right)^* - K'\left(u_k\right)^*\right]\left(2v_{k+1} - v_k\right)}_{\rightsquigarrow D_k^{CGF}} + \underbrace{T_k K'\left(u_k\right)^*\left(v_{k+1} - v_k\right) + u_{k+1} - u_k}_{\rightsquigarrow M_k} = 0. \quad (4.1.13)$$

Now we can use equations (4.1.10) and (4.1.13) to define the functions of the desired inclusion in (4.1.7):

- The step size operator on the product space $\mathcal{W}$ is simply defined as

$$W_k := \begin{pmatrix} T_k & 0 \\ 0 & \Sigma_k \end{pmatrix} \in \mathbb{L}\left(\mathcal{W}, \mathcal{W}\right), \quad (4.1.14)$$

  where $T_k \in \mathbb{L}\left(\mathcal{U}, \mathcal{U}\right)$ and $\Sigma_k \in \mathbb{L}\left(\mathcal{V}, \mathcal{V}\right)$ are the step size operators of Algorithm 4.1.

- We already know the function H from the optimality condition in (OC):

$$H(w) = \begin{pmatrix} \partial G(u) + K'(u)^* v \\ \partial \mathrm{CVaR}_\beta^*(v) - K(u) \end{pmatrix} \quad \text{for all } w := (u, v) \in \mathcal{W}. \quad (4.1.15)$$

- The discrepancy function $D_k := D_k^{LIN} + D_k^{CGF}$ is divided into two parts. For all $w := (u, v) \in \mathcal{W}$, the discrepancy from linearization is

$$D_k^{LIN}(w) := \begin{pmatrix} T_k\left[K'\left(u_k\right)^* - K'\left(u\right)^*\right] v \\ \Sigma_k\left[K\left(u\right) - K\left(u_k\right) - K'\left(u_k\right)\left(u - u_k\right)\right] \end{pmatrix}, \quad (4.1.16)$$

  and the discrepancy from CGF is

$$D_k^{CGF}(w) := \begin{pmatrix} T_k\left[\hat{K}_k'\left(u_k\right)^* - K'\left(u_k\right)^*\right]\left(2v - v_k\right) \\ \Sigma_k\left[K\left(u_k\right) - \hat{K}_k\left(u_k\right)\right] \end{pmatrix}. \quad (4.1.17)$$

  This is indeed a reasonable definition since, on the one hand, if K is linear, we have that $K' \equiv K$ and therefore $D_k^{LIN}(w) = 0$ for all $w \in \mathcal{W}$ and all $k \in \mathbb{N}$. On the other hand, if we do not employ component-wise gradient freezing (i.e. $A_k = \Xi$ for all $k \in \mathbb{N}$), the definitions (4.1.5) and (4.1.6) imply $\hat{K}_k \equiv K$ and $\hat{K}_k'(u)^* = K'(u)^*$ for all $u \in \mathcal{U}$ and all $k \in \mathbb{N}$. Therefore, in this case we have $D_k^{CGF}(w) = 0$ for all $w \in \mathcal{W}$.

- The preconditioning operator is given by

$$M_k := \begin{pmatrix} \mathrm{Id}_{\mathcal{U}} & T_k K'\left(u_k\right)^* \\ \Sigma_k K'\left(u_k\right) & \mathrm{Id}_{\mathcal{V}} \end{pmatrix}. \quad (4.1.18)$$

Therefore, the following implicit formulation is, by construction, equivalent to Algorithm 4.1, where Lines 6 to 8 have been replaced by (4.1.7).

Algorithm 4.2 (Implicit Formulation)

1 Initialize $k := 0$.
2 **repeat**
3 **if** $k \geq 1$ **then**
4 Randomly select $A_k \in \mathcal{A}$.
5 **end**
6 Find $w_{k+1} \in \mathcal{W}$ satisfying $0 \in W_k H(w_{k+1}) + D_k(w_{k+1}) + M_k(w_{k+1} - w_k)$.
7 Determine step size operator $W_{k+1} \in \mathbb{L}(\mathcal{W}, \mathcal{W})$.
8 Update $k \leftarrow k + 1$.
9 **until** $\| u_k - u_{k-1} \|_{\mathcal{U}} < \varepsilon$ **and** $\| v_k - v_{k-1} \|_{\mathcal{V}} < \varepsilon$

4.2 Randomization

Recall the subdifferential inclusion (4.1.7)

$$0 \in W_k H(w_{k+1}) + D_k(w_{k+1}) + M_k(w_{k+1} - w_k) \quad \text{for all } k \in \mathbb{N}_0, \tag{4.2.1}$$

for a sequence $(w_k)_{k \in \mathbb{N}_0} \subset \mathcal{W}$. In order to simplify the notation, we define for all $k \in \mathbb{N}_0$ the mapping $\tilde{H}_k \colon \mathcal{W} \rightrightarrows \mathcal{W}$ as

$$\tilde{H}_k(w) := W_k H(w) + D_k(w) \tag{4.2.2}$$

for all $w \in \mathcal{W}$. Thus, (4.2.1) is equivalent to

$$0 \in \tilde{H}_k(w_{k+1}) + M_k(w_{k+1} - w_k) \quad \text{for all } k \in \mathbb{N}_0. \tag{4.2.3}$$

Since the set $A_k \in \mathcal{A}$ is selected randomly at each iteration $k \in \mathbb{N}$ of Algorithm 4.2, the inclusion (4.2.3) has to hold P-almost surely. Therefore, we consider the randomized inclusion

$$0 \in \tilde{H}_k(w_{k+1}) + M_k(w_{k+1} - w_k) \quad \text{P-a.s.} \quad \text{for all } k \in \mathbb{N}_0, \tag{RI}$$

where the probability measure P has not yet been defined. To make up for this, let $(\Theta, \mathcal{F}, \mathsf{P})$ be a probability space where $\mathsf{E}(\cdot \,|\, \mathcal{B})$ denotes the conditional expectation given $\mathcal{B} \subset \mathcal{F}$ (see Definition 2.5.4). From now on, we use A_k to denote a random variable on $(\Theta, \mathcal{F}, \mathsf{P})$ with values in $\mathcal{A}$. Since A_k is involved in the definition (4.1.17) of D_k^{CGF}, D_k is a random variable as well. Furthermore, we assume that the step size operator W_k is a random variable because we want to keep the possibility to make it dependent on previous iterations. The definition (4.1.18) directly implies that M_k and consequently w_{k+1} are random variables as well. In order to properly describe the information available in each iteration, we will use a filtration as introduced in Definition 2.5.5. The following definition summarizes the conditions under which we have a filtration $\mathfrak{F} := (\mathcal{F}_k)_{k \in \mathbb{N}_0}$ such that $\mathcal{F}_k$ models all the information available before commencing iteration k of Algorithm 4.2 (i.e. *before* the random selection of $A_k \in \mathcal{A}$).

Definition 4.2.1 (Compatible Filtration)

Let $\mathfrak{F} := (\mathcal{F}_k)_{k\in\mathbb{N}_0}$ be a filtration with $\mathcal{F}_0 := \mathcal{F}_1 := \{\Theta, \emptyset\}$, and $(w_k)_{k\in\mathbb{N}_0} \subset \mathcal{R}(\mathcal{F}, \mathcal{W})$ a solution to (RI). We call $\mathfrak{F}$ a *compatible filtration* if, for every $k \in \mathbb{N}$ with $k \geq 2$, $\mathcal{F}_k$ is the smallest σ-algebra such that the following random variables are $\mathcal{F}_k$-measurable for all $i \in \{2, \dots, k\}$:

$$
\begin{aligned}
A_{i-1} &: \Theta \to \mathcal{A}, \\
D_{i-1} &: \Theta \to \{f : \mathcal{W} \to \mathcal{W}\}, \\
W_i &: \Theta \to \mathbb{L}(\mathcal{W}, \mathcal{W}), \\
M_i &: \Theta \to \mathbb{L}(\mathcal{W}, \mathcal{W}), \\
w_i &: \Theta \to \mathcal{W}.
\end{aligned}
$$

We have omitted the indices $i \in \{0, 1\}$ here, because the first random selection of $A_k \in \mathcal{A}$ takes place in iteration $k = 1$. Thus, every *random variable* used before is actually not random at all, since it depends only on the input of the algorithm. Also note that we slightly abuse the notation and do not explicitly mention the dependence on the probability space, for example the mapping $(M_k(\cdot))\,(w_{k+1}(\cdot) - w_k(\cdot)) : \Theta \to \mathcal{W}$ is just denoted by $M_k\,(w_{k+1} - w_k)$.

4.3 Weak Convergence

In this section, we present a convergence proof of Algorithm 4.2 based on the ideas in [Val18] and [CMV19]. The first subsection deals with an abstract convergence result that is not specific to this particular algorithm. However, it clearly has a structure we will be able to exploit later. Since the so-called *stochastic quasi-Fejér monotonicity* is crucial for the proof, we present a sufficient condition for this property in the second subsection. Finally, in the third subsection, we derive estimates that are tailored to the algorithm such that we have a more tractable formulation of the abstract result from the first subsection.

4.3.1 Abstract Proof

For the proof of convergence, we want to adopt the idea in [VP17] of *testing* the subdifferential inclusion (RI) by a linear operator, as described in [Val18]. It requires us to use the weighted norm we introduced in Definition 2.2.12. The following assumption summarizes the required properties which guarantee that the weighted norm is indeed a norm.

Assumption 4.3.1

Let $\mathfrak{F} := (\mathcal{F}_k)_{k\in\mathbb{N}_0}$ be a filtration and $(R_k)_{k\in\mathbb{N}_0}$ a sequence of random variables such that $R_k \in \mathcal{R}(\mathcal{F}_k, \mathbb{L}(\mathcal{W}, \mathcal{W}))$ for all $k \in \mathbb{N}_0$. We assume that there exist $\Theta' \in \mathcal{F}^1$ and $\varepsilon \neq 0$ such that, for all $\theta \in \Theta'$ and all $k \in \mathbb{N}_0$, the following holds:

(i) $R_k(\theta)$ is self-adjoint;

(ii) $R_k(\theta) \geq \varepsilon^2 \mathrm{Id}_{\mathcal{W}}$.

Note that Assumption 4.3.1 (ii) especially implies the positive definiteness of $R_k(\theta)$. In the following, we will focus on real-valued, non-negative random variables, and therefore define

$$l_+(\mathfrak{F}) := \left\{(x_k)_{k\in\mathbb{N}_0} \,\middle|\, x_k \in \mathcal{R}(\mathcal{F}_k, [0,\infty)) \text{ for all } k \in \mathbb{N}_0\right\} \tag{4.3.1}$$

and

$$l_+^p(\mathfrak{F}) := \left\{(x_k)_{k\in\mathbb{N}_0} \in l_+(\mathfrak{F}) \,\middle|\, \sum_{k\in\mathbb{N}_0} x_k^p < \infty \quad \text{P-a.s.}\right\} \tag{4.3.2}$$

for all $p \in (0,\infty)$ and a filtration $\mathfrak{F} = (\mathcal{F}_k)_{k\in\mathbb{N}_0}$. Now, we first present a lemma that shows under which conditions a sequence of non-negative random variables converges P-a.s. It stems from [CP15, Lemma 2.2] and was first proven in [RS71, Theorem 1].

Lemma 4.3.2 (Robbins-Siegmund)

Let $\mathfrak{F} := (\mathcal{F}_k)_{k\in\mathbb{N}_0}$ be a filtration. Let further $(\alpha_k)_{k\in\mathbb{N}_0}, (\vartheta_k)_{k\in\mathbb{N}_0} \in l_+(\mathfrak{F})$ and $(\eta_k)_{k\in\mathbb{N}_0}, (\chi_k)_{k\in\mathbb{N}_0} \in l_+^1(\mathfrak{F})$ such that

$$\mathsf{E}(\alpha_{k+1} \mid \mathcal{F}_k) + \vartheta_k \leq (1+\chi_k)\alpha_k + \eta_k \quad \text{P-a.s.}$$

for all $k \in \mathbb{N}_0$. Then $(\vartheta_k)_{k\in\mathbb{N}_0} \in l_+^1(\mathfrak{F})$ and there exists an $\alpha \in \mathcal{R}(\mathcal{F}, [0,\infty))$ such that

$$\alpha_k \longrightarrow \alpha \quad \text{P-a.s.}$$

From now on, let $\mathfrak{W}\left((x_k)_{k\in\mathbb{N}_0}\right)$ denote the set of all weak accumulation points of the sequence $(x_k)_{k\in\mathbb{N}_0} \subset \mathcal{W}$. The following proposition is an extension of parts of [CP15, Proposition 2.3] that enables us to use the weighted, stochastic norm $\|\cdot\|_{R_k}$ with $R_k \in \mathcal{R}(\mathcal{F}_k, \mathbb{L}(\mathcal{W},\mathcal{W}))$. Moreover, the right-hand side of the inequality in Proposition 4.3.3 (ii) additionally contains the term $\lambda_k(z)$, which will be helpful when estimating the discrepancy resulting from CGF. The proof is almost identical to [CP15] but does not require the subset of $\mathcal{W}$ to be closed.

Proposition 4.3.3 (Weak Convergence)

Let $\mathcal{Z} \subset \mathcal{W}$ be non-empty and $(w_k)_{k\in\mathbb{N}_0} \subset \mathcal{R}(\mathcal{F},\mathcal{W})$ a sequence of random variables. Let $(R_k)_{k\in\mathbb{N}_0}$ satisfy Assumption 4.3.1. We define the filtration $\mathfrak{F} := (\mathcal{F}_k)_{k\in\mathbb{N}_0} \subset \mathcal{F}$ such that w_k and R_k are $\mathcal{F}_k$-measurable for all $k \in \mathbb{N}_0$. We further assume that

(i) $\exists C \in \mathbb{R}$ such that $\|R_k\|_{\mathbb{L}(\mathcal{W},\mathcal{W})} \leq C^2$ P-a.s. for all $k \in \mathbb{N}$;

(ii) for every $z \in \mathcal{Z}$ there exist sequences $(\Delta_k(z))_{k\in\mathbb{N}_0} \in l_+(\mathfrak{F})$ and $(\lambda_k(z))_{k\in\mathbb{N}_0} \in l_+^1(\mathfrak{F})$ such that the so-called *stochastic quasi-Fejér monotonicity*

$$\frac{1}{2}\mathsf{E}\left(\|w_{k+1}-z\|_{R_{k+1}}^2 \,\middle|\, \mathcal{F}_k\right) + \Delta_k(z) \leq \frac{1}{2}\|w_k - z\|_{R_k}^2 + \lambda_k(z)$$

holds P-a.s. for all $k \in \mathbb{N}_0$;

(iii) $\mathfrak{W}\left((w_k)_{k\in\mathbb{N}_0}\right) \subset \mathcal{Z}$ P-a.s., i.e. there exists $\overline{\Theta} \in \mathcal{F}^1$ such that $\mathfrak{W}\left((w_k(\theta))_{k\in\mathbb{N}_0}\right) \subset \mathcal{Z}$ for all $\theta \in \overline{\Theta}$;

(iv) there exists $\overline{\Theta}' \in \mathcal{F}^1$ such that, for all $\theta \in \overline{\Theta}'$, there is a linear operator $R_\infty(\theta) \in \mathbb{L}(\mathcal{W}, \mathcal{W})$ satisfying

$$R_{n_k}(\theta)w \to R_\infty(\theta)w \quad \text{as} \quad k \to \infty$$

for all $w \in \mathcal{W}$ and every weakly convergent subsequence $\left(w_{n_k}(\theta)\right)_{k\in\mathbb{N}_0}$.

Then there exists $\overline{w} \in \mathcal{R}(\mathcal{F}, \mathcal{Z})$ such that $w_k \rightharpoonup \overline{w}$ P-a.s. as $k \to \infty$.

Proof: We divide the proof into three steps.

Step 1 P-a.s.-boundedness of $(w_k)_{k\in\mathbb{N}_0}$.

Let $z \in \mathcal{Z}$ be fixed. If we employ Lemma 4.3.2 with $\chi_k(\theta) = 0$ for all $\theta \in \Theta$, $\vartheta_k = \Delta_k(z)$, $\eta_k = \lambda_k(z)$, and $\alpha_k = \frac{1}{2} \| w_k - z \|^2_{R_k}$, we deduce from condition (ii) that there exists $\alpha \in \mathcal{R}(\mathcal{F}, [0, \infty))$ such that

$$\frac{1}{2} \| w_k - z \|^2_{R_k} \to \alpha \quad \text{P-a.s.} \tag{4.3.3}$$

Thus, the sequence $\left(\| w_k - z \|^2_{R_k} \right)_{k\in\mathbb{N}_0}$ is bounded P-a.s. and so is $(w_k)_{k\in\mathbb{N}_0}$ due to Assumption 4.3.1 (ii), i.e. there exists $\overline{\Theta}'' \in \mathcal{F}^1$ such that $(w_k(\theta))_{k\in\mathbb{N}_0}$ is bounded for all $\theta \in \overline{\Theta}''$.

Step 2 Construction of $\tilde{\Theta}$.

Analogously to the proof of [CP15, Proposition 2.3 (ii)], one can easily show that we can omit the squaring and the multiplication by $\frac{1}{2}$ in (4.3.3) and still get P-a.s.-convergence, i.e.

$$\forall z \in \mathcal{Z} \quad \exists \Theta_z \in \mathcal{F}^1,\ \alpha_z \in \mathcal{R}(\mathcal{F}, [0, \infty)) \quad \text{s.t.}$$
$$\| w_k(\theta) - z \|_{R_k(\theta)} \to \alpha_z(\theta) \quad \text{for all } \theta \in \Theta_z. \tag{4.3.4}$$

Since $\mathcal{W}$ is separable, we can find a countable subset $\mathcal{Y} \subset \mathcal{Z}$ such that $\mathrm{cl}(\mathcal{Y}) \supset \mathcal{Z}$. Now let

$$\tilde{\Theta} := \Theta' \cap \left(\bigcap_{z\in\mathcal{Y}} \Theta_z \right), \tag{4.3.5}$$

where Θ' stems from Assumption 4.3.1 and ensures that $\| \cdot \|_{R_k(\theta)}$ is indeed a norm for every $k \in \mathbb{N}_0$ and $\theta \in \tilde{\Theta}$. It follows from the countability of $\mathcal{Y}$ that

$$\begin{aligned} \mathrm{P}\left(\tilde{\Theta}\right) &= 1 - \mathrm{P}\left(\tilde{\Theta}^c\right) = 1 - \mathrm{P}\left(\Theta'^c \cup \left(\bigcup_{z\in\mathcal{Y}} \Theta_z^c \right) \right) \\ &\geq 1 - \mathrm{P}\left(\Theta'^c\right) - \sum_{z\in\mathcal{Y}} \mathrm{P}\left(\Theta_z^c\right) = 1, \end{aligned} \tag{4.3.6}$$

which implies that $\mathrm{P}\left(\tilde{\Theta}\right) = 1$, i.e. $\tilde{\Theta} \in \mathcal{F}^1$.

Let now $z \in \mathcal{Z}$ be fixed. The goal is to show convergence of $\| w_k(\theta) - z \|_{R_k(\theta)}$, where θ stems from the more general $\tilde{\Theta}$ instead of a specific Θ_z as in (4.3.4). Since

$\mathrm{cl}(\mathcal{Y}) \supset \mathcal{Z}$, there exists a sequence $(z_n)_{n\in\mathbb{N}_0} \subset \mathcal{Y}$ such that $z_n \to z$. If we apply (4.3.4) to every element of this sequence, we conclude that there exists for every $n \in \mathbb{N}_0$ a random variable $\alpha_n \in \mathcal{R}(\mathcal{F}, [0,\infty))$ such that

$$\| w_k(\theta) - z_n \|_{R_k(\theta)} \to \alpha_n(\theta) \tag{4.3.7}$$

for all $\theta \in \tilde{\Theta}$ as $k \to \infty$. Since $\| \cdot \|_{R_k(\theta)}$ is a norm for all $k \in \mathbb{N}_0$ and $\theta \in \tilde{\Theta}$, we can apply the usual triangle inequality to get

$$\begin{aligned} -\| z_n - z \|_{R_k(\theta)} &\leq \| w_k(\theta) - z \|_{R_k(\theta)} - \| w_k(\theta) - z_n \|_{R_k(\theta)} \\ &\leq \| z_n - z \|_{R_k(\theta)} \end{aligned} \tag{4.3.8}$$

for all $n \in \mathbb{N}_0$. According to condition (i), we have

$$\| z_n - z \|_{R_k(\theta)} \leq \sqrt{\| R_k(\theta) \|_{\mathbb{L}(\mathcal{W},\mathcal{W})}} \| z_n - z \|_{\mathcal{W}} \leq C \| z_n - z \|_{\mathcal{W}} \tag{4.3.9}$$

for all $n \in \mathbb{N}_0$, $k \in \mathbb{N}$, and $\theta \in \tilde{\Theta}$. Therefore, (4.3.8) implies

$$\begin{aligned} -C \| z_n - z \|_{\mathcal{W}} &\leq \| w_k(\theta) - z \|_{R_k(\theta)} - \| w_k(\theta) - z_n \|_{R_k(\theta)} \\ &\leq C \| z_n - z \|_{\mathcal{W}} . \end{aligned} \tag{4.3.10}$$

Applying $\liminf_{k\to\infty}$ and $\limsup_{k\to\infty}$ on the middle part of the inequality yields

$$\begin{aligned} -C \| z_n - z \|_{\mathcal{W}} &\leq \liminf_{k\to\infty} \left(\| w_k(\theta) - z \|_{R_k(\theta)} - \| w_k(\theta) - z_n \|_{R_k(\theta)} \right) \\ &\overset{(4.3.7)}{=} \liminf_{k\to\infty} \left(\| w_k(\theta) - z \|_{R_k(\theta)} \right) - \alpha_n(\theta) \\ &\leq \limsup_{k\to\infty} \left(\| w_k(\theta) - z \|_{R_k(\theta)} \right) - \alpha_n(\theta) \\ &\overset{(4.3.7)}{=} \limsup_{k\to\infty} \left(\| w_k(\theta) - z \|_{R_k(\theta)} - \| w_k(\theta) - z_n \|_{R_k(\theta)} \right) \\ &\leq C \| z_n - z \|_{\mathcal{W}} \end{aligned} \tag{4.3.11}$$

for all $n \in \mathbb{N}_0$. Hence, taking the limit $n \to \infty$, we obtain that

$$\lim_{k\to\infty} \| w_k(\theta) - z \|_{R_k(\theta)} = \lim_{n\to\infty} \alpha_n(\theta) \tag{4.3.12}$$

for all $\theta \in \tilde{\Theta}$, i.e. the sequence $\left(\| w_k(\theta) - z \|_{R_k(\theta)} \right)_{k\in\mathbb{N}_0}$ converges. We also have that $\lim_{n\to\infty} \alpha_n(\theta) < \infty$ because otherwise, (4.3.12) and (4.3.11) would imply that $C \| z_n - z \|_{\mathcal{W}} = \infty$ for all $n \in \mathbb{N}_0$. The improvement compared to (4.3.4) is the changed order of quantifiers:

$$\begin{aligned} &\exists \tilde{\Theta} \in \mathcal{F}^1 \quad \forall z \in \mathcal{Z} \quad \exists \alpha_z \in \mathcal{R}(\mathcal{F}, [0,\infty)) \quad \text{s.t.} \\ &\qquad\qquad \| w_k(\theta) - z \|_{R_k(\theta)} \to \alpha_z(\theta) \quad \text{for all } \theta \in \tilde{\Theta}. \end{aligned} \tag{4.3.13}$$

Step 3 Convergence of $(w_k)_{k\in\mathbb{N}_0}$.

Let $\overline{\Theta}$ and $\overline{\Theta}'$ be defined as in conditions (iii) and (iv) of this proposition, $\overline{\Theta}''$ as defined in **Step 1**, and $\tilde{\Theta}$ as constructed in **Step 2**. Let

$$\theta \in \hat{\Theta} := \overline{\Theta} \cap \overline{\Theta}' \cap \overline{\Theta}'' \cap \tilde{\Theta} \tag{4.3.14}$$

be fixed. According to [BC17, Lemma 2.46], $(w_k(\theta))_{k\in\mathbb{N}_0}$ converges weakly if and only if it is bounded and possesses at most one weak accumulation point. Since the boundedness of $(w_k(\theta))_{k\in\mathbb{N}_0}$ was already shown in **Step 1**, it suffices to show that all weak accumulation points coincide. Therefore, let $\overline{w}(\theta), \overline{w}'(\theta) \in \mathfrak{W}\left((w_k(\theta))_{k\in\mathbb{N}_0}\right)$ be two weak accumulation points of $(w_k(\theta))_{k\in\mathbb{N}_0}$, i.e.

$$w_{n_k}(\theta) \rightharpoonup \overline{w} := \overline{w}(\theta) \tag{4.3.15}$$

$$\text{and} \quad w_{l_k}(\theta) \rightharpoonup \overline{w}' := \overline{w}'(\theta) \tag{4.3.16}$$

as $k \to \infty$. Because of condition (iii), we know that $\overline{w}, \overline{w}' \in \mathcal{Z}$. Therefore, we can apply (4.3.13) to see that $\left(\| w_k(\theta) - \overline{w} \|_{R_k(\theta)}\right)_{k\in\mathbb{N}_0}$ and $\left(\left\| w_k(\theta) - \overline{w}' \right\|_{R_k(\theta)}\right)_{k\in\mathbb{N}_0}$ converge. Since

$$\begin{aligned}\left\langle w_k(\theta), \overline{w} - \overline{w}' \right\rangle_{R_k(\theta)} - \frac{1}{2} \| \overline{w} \|^2_{R_k(\theta)} + \frac{1}{2} \left\| \overline{w}' \right\|^2_{R_k(\theta)} \\ = \frac{1}{2} \left\| w_k(\theta) - \overline{w}' \right\|^2_{R_k(\theta)} - \frac{1}{2} \| w_k(\theta) - \overline{w} \|^2_{R_k(\theta)},\end{aligned} \tag{4.3.17}$$

the left-hand side must also converge, say

$$\left\langle w_k(\theta), \overline{w} - \overline{w}' \right\rangle_{R_k(\theta)} - \frac{1}{2} \| \overline{w} \|^2_{R_k(\theta)} + \frac{1}{2} \left\| \overline{w}' \right\|^2_{R_k(\theta)} \to \rho \in \mathbb{R} \tag{4.3.18}$$

as $k \to \infty$. The weak convergence of the subsequence $w_{n_k}(\theta) \rightharpoonup \overline{w}$ together with the strong convergence of the operator in condition (iv), i.e. $R_{n_k}(\theta)w \to R_\infty(\theta)w$ for all $w \in \mathcal{W}$, imply that

$$\begin{aligned}\left\langle w_{n_k}(\theta), \overline{w} - \overline{w}' \right\rangle_{R_{n_k}(\theta)} - \frac{1}{2} \| \overline{w} \|^2_{R_{n_k}(\theta)} + \frac{1}{2} \left\| \overline{w}' \right\|^2_{R_{n_k}(\theta)} \\ \to \left\langle \overline{w}, \overline{w} - \overline{w}' \right\rangle_{R_\infty(\theta)} - \frac{1}{2} \| \overline{w} \|^2_{R_\infty(\theta)} + \frac{1}{2} \left\| \overline{w}' \right\|^2_{R_\infty(\theta)}\end{aligned} \tag{4.3.19}$$

as $k \to \infty$. Since we have shown in (4.3.18) that the whole sequence converges, it follows that

$$\rho = \left\langle \overline{w}, \overline{w} - \overline{w}' \right\rangle_{R_\infty(\theta)} - \frac{1}{2} \| \overline{w} \|^2_{R_\infty(\theta)} + \frac{1}{2} \left\| \overline{w}' \right\|^2_{R_\infty(\theta)}. \tag{4.3.20}$$

Analogously, by considering the subsequence $w_{l_k}(\theta) \rightharpoonup \overline{w}'$, we get

$$\rho = \left\langle \overline{w}', \overline{w} - \overline{w}' \right\rangle_{R_\infty(\theta)} - \frac{1}{2} \| \overline{w} \|^2_{R_\infty(\theta)} + \frac{1}{2} \left\| \overline{w}' \right\|^2_{R_\infty(\theta)}. \tag{4.3.21}$$

Subtracting (4.3.21) from (4.3.20) yields

$$0 = \left\langle \overline{w} - \overline{w}', \overline{w} - \overline{w}' \right\rangle_{R_\infty(\theta)} = \left\| \overline{w} - \overline{w}' \right\|^2_{R_\infty(\theta)} \tag{4.3.22}$$

and therefore $\overline{w} = \overline{w}'$, since Assumption 4.3.1 (ii) implies the positive definiteness of $R_\infty(\theta)$. The measurability of $\overline{w} \colon \Theta \to \mathcal{Z}$ follows from [Pet38, Corollary 1.13] and since $\mathsf{P}\left(\hat{\Theta}\right) = 1$, we conclude that $w_k \rightharpoonup \overline{w} \in \mathcal{R}(\mathcal{F}, \mathcal{Z})$ P-a.s. as $k \to \infty$. □

4.3.2 Stochastic Quasi-Fejér Monotonicity

The aforementioned testing approach adapted from [VP17] requires us to multiply the inclusion (RI) by a testing operator $Z_k \in \mathbb{L}(\mathcal{W}, \mathcal{W})$ for every $k \in \mathbb{N}_0$, which is, in this case, also a random variable. The resulting product $Z_k M_k \in \mathcal{R}(\mathcal{F}_k, \mathbb{L}(\mathcal{W}, \mathcal{W}))$ will take the place of the operator R_k in Proposition 4.3.3. Therefore, a sufficient condition for the stochastic quasi-Fejér monotonicity

$$\frac{1}{2}\mathsf{E}\left(\| w_{k+1} - z \|^2_{R_{k+1}} \,\middle|\, \mathcal{F}_k\right) + \Delta_k(z) \leq \frac{1}{2}\| w_k - z \|^2_{R_k} + \lambda_k(z)$$

in Proposition 4.3.3 (ii) will be helpful when it comes to verifying the assumptions of the proposition. The following lemma is a result of applying the conditional expectation to [Val18, Theorem 2.1] and adding the term λ_k. Consequently, the presented proof is almost identical to the original one.

Lemma 4.3.4

Let (RI) be solvable for a sequence $(w_k)_{k\in\mathbb{N}_0} \subset \mathcal{R}(\mathcal{F}, \mathcal{W})$. Let further $\mathfrak{F}$ be a compatible filtration as in Definition 4.2.1 and $(Z_k)_{k\in\mathbb{N}_0} \subset \mathbb{L}(\mathcal{W}, \mathcal{W})$ a sequence of testing operators. If, for all $k \in \mathbb{N}_0$, $Z_k M_k$ is self-adjoint P-a.s. and the condition

$$\begin{aligned} &\mathsf{E}\left(\left\langle \tilde{H}_k(w_{k+1}), w_{k+1} - z \right\rangle_{Z_k} \,\middle|\, \mathcal{F}_k\right) + \lambda_k \\ &\quad \geq \frac{1}{2}\mathsf{E}\left(\| w_{k+1} - z \|^2_{Z_{k+1}M_{k+1} - Z_k M_k} \,\middle|\, \mathcal{F}_k\right) - \frac{1}{2}\mathsf{E}\left(\| w_{k+1} - w_k \|^2_{Z_k M_k} \,\middle|\, \mathcal{F}_k\right) + \Delta_k \end{aligned} \tag{ECI}$$

is satisfied P-a.s. for some $z \in \mathcal{W}$, $(\Delta_k)_{k\in\mathbb{N}_0} \in l_+(\mathfrak{F})$, and $(\lambda_k)_{k\in\mathbb{N}_0} \in l^1_+(\mathfrak{F})$, then the so-called *stochastic quasi-Fejér monotonicity*

$$\frac{1}{2}\mathsf{E}\left(\| w_{k+1} - z \|^2_{Z_{k+1}M_{k+1}} \,\middle|\, \mathcal{F}_k\right) + \Delta_k \leq \frac{1}{2}\| w_k - z \|^2_{Z_k M_k} + \lambda_k \tag{SQF}$$

holds P-a.s.

Proof: Let $k \in \mathbb{N}_0$ be fixed. Inserting (RI) into (ECI) yields

$$\begin{aligned} &-\mathsf{E}\left(\langle w_{k+1} - w_k, w_{k+1} - z \rangle_{Z_k M_k} \,\middle|\, \mathcal{F}_k\right) + \lambda_k \\ &\quad \geq \frac{1}{2}\mathsf{E}\left(\| w_{k+1} - z \|^2_{Z_{k+1}M_{k+1} - Z_k M_k} \,\middle|\, \mathcal{F}_k\right) - \frac{1}{2}\mathsf{E}\left(\| w_{k+1} - w_k \|^2_{Z_k M_k} \,\middle|\, \mathcal{F}_k\right) + \Delta_k \end{aligned} \tag{4.3.23}$$

P-a.s. Together with the three-point version of Pythagoras' identity

$$\begin{aligned}&\langle w_{k+1} - w_k, w_{k+1} - z\rangle_{Z_k M_k} \\ &\qquad = \frac{1}{2} \| w_{k+1} - w_k \|^2_{Z_k M_k} - \frac{1}{2} \| w_k - z \|^2_{Z_k M_k} + \frac{1}{2} \| w_{k+1} - z \|^2_{Z_k M_k}, \end{aligned} \tag{4.3.24}$$

we conclude that (4.3.23) is equivalent to

$$\begin{aligned}&-\frac{1}{2}\mathsf{E}\left(\| w_{k+1} - w_k \|^2_{Z_k M_k} \,\middle|\, \mathcal{F}_k\right) + \frac{1}{2} \| w_k - z \|^2_{Z_k M_k} - \frac{1}{2}\mathsf{E}\left(\| w_{k+1} - z \|^2_{Z_k M_k} \,\middle|\, \mathcal{F}_k\right) + \lambda_k \\ &\qquad \geq \frac{1}{2}\mathsf{E}\left(\| w_{k+1} - z \|^2_{Z_{k+1} M_{k+1}} \,\middle|\, \mathcal{F}_k\right) - \frac{1}{2}\mathsf{E}\left(\| w_{k+1} - z \|^2_{Z_k M_k} \,\middle|\, \mathcal{F}_k\right) \\ &\qquad\quad - \frac{1}{2}\mathsf{E}\left(\| w_{k+1} - w_k \|^2_{Z_k M_k} \,\middle|\, \mathcal{F}_k\right) + \Delta_k \end{aligned} \tag{4.3.25}$$

P-a.s. After rearranging, we can easily see that this is equivalent to (SQF). □

In order to better understand the motivation for using the testing approach, we apply, under the assumptions of Lemma 4.3.4, the expectation to (SQF), which yields

$$\frac{1}{2}\mathsf{E}\left(\| w_{k+1} - z \|^2_{Z_{k+1} M_{k+1}}\right) + \mathsf{E}\left(\Delta_k\right) \leq \frac{1}{2}\mathsf{E}\left(\| w_k - z \|^2_{Z_k M_k}\right) + \mathsf{E}\left(\lambda_k\right) \tag{4.3.26}$$

P-a.s. for all $k \in \mathbb{N}_0$. If we sum over all $k \in \{0, \dots, N-1\}$ with some $N \in \mathbb{N}$, we get

$$\frac{1}{2}\mathsf{E}\left(\| w_N - z \|^2_{Z_N M_N}\right) \leq \frac{1}{2}\mathsf{E}\left(\| w_0 - z \|^2_{Z_0 M_0}\right) + \sum_{k=0}^{N-1} \mathsf{E}\left(\lambda_k\right), \tag{4.3.27}$$

where we have used $\Delta_k \geq 0$ due to the assumption that $(\Delta_k)_{k\in\mathbb{N}_0} \in l_+(\mathfrak{F})$. Furthermore, since $(\lambda_k)_{k\in\mathbb{N}_0} \in l^1_+(\mathfrak{F})$, we know that there exists a constant $C \in [0, \infty)$ such that

$$\sum_{k=1}^{\infty} \lambda_k = C \quad \text{P-a.s.} \tag{4.3.28}$$

Applying the expectation together with the theorem of monotone convergence [Kle13, Theorem 4.20] implies that

$$\sum_{k=1}^{N-1} \mathsf{E}\left(\lambda_k\right) \nearrow C \quad \text{P-a.s.} \tag{4.3.29}$$

as $N \to \infty$. Hence, it follows from (4.3.27) that

$$\frac{1}{2}\mathsf{E}\left(\| w_N - z \|^2_{Z_N M_N}\right) \leq \frac{1}{2}\mathsf{E}\left(\| w_0 - z \|^2_{Z_0 M_0}\right) + C \tag{4.3.30}$$

P-a.s. We can see that the operator $Z_N M_N$ forms a local metric which measures how close the iterates w_N are to a solution z. Therefore, the goal is to choose the step size and testing operators such that $Z_N M_N$ grows fast as $N \to \infty$. In fact, we have the following result showing under which circumstances we obtain a convergence rate in the sense of mean convergence as in [Kle13, Definition 6.8]. It is adapted from [Val18, Proposition 2.4].

Proposition 4.3.5 (Convergence Rate in Expectation)

Under the assumptions of Lemma 4.3.4, assume that $Z_N M_N \geq \mu(N)\,\mathrm{Id}_{\mathcal{W}}$ P-a.s. for all $N \in \mathbb{N}$ and an increasing function $\mu\colon \mathbb{N} \to [0,\infty)$ with $\lim_{N\to\infty} \mu(N) = \infty$. Then

$$\mathsf{E}\left(\| w_N - z \|_{\mathcal{W}}^2\right) \to 0$$

at the rate $O(1/\mu(N))$ as $N \to \infty$.

Proof: Let $N \in \mathbb{N}$. The inequality (4.3.30) together with the definition of the weighted inner product (2.2.5) implies

$$\mathsf{E}\Big(\langle Z_N M_N (w_N - z), w_N - z\rangle_{\mathcal{W}} \Big) \leq \mathsf{E}\left(\| w_0 - z \|_{Z_0 M_0}^2\right) + 2C =: C', \tag{4.3.31}$$

and since $Z_N M_N \geq \mu(N)\,\mathrm{Id}_{\mathcal{W}}$ P-a.s., we have

$$\mathsf{E}\left(\left\| w^N - z \right\|_{\mathcal{W}}^2\right) \leq \frac{C'}{\mu(N)}. \tag{4.3.32}$$

Taking the limit $N \to \infty$ concludes the proof. □

4.3.3 Convergence to Optimal Solution

The following proposition shows under which specific condition on Δ_k, together with some other assumptions, the iterates of (RI) not only converge weakly to *some* limit, but to an optimal solution of Problem 3.1.1. It is basically a reformulation of [CMV19, Proposition 2.2].

Theorem 4.3.6 (Weak Convergence to Optimal Solution)

Let (RI) be solvable for a sequence $(w_k)_{k\in\mathbb{N}_0} \subset \mathcal{R}(\mathcal{F},\mathcal{W})$. Let further $\mathfrak{F}$ be a compatible filtration as in Definition 4.2.1, $\mathcal{Z} \subset H^{-1}(0)$ a non-empty subset, and $(Z_k)_{k\in\mathbb{N}_0} \subset \mathbb{L}(\mathcal{W},\mathcal{W})$ a sequence of testing operators. We assume that

(i) $R_k := Z_k M_k$ satisfies Assumption 4.3.1 for all $k \in \mathbb{N}_0$;

(ii) $\exists C \in \mathbb{R}$ such that $\| Z_k M_k \|_{\mathbb{L}(\mathcal{W},\mathcal{W})} \leq C^2$ P-a.s. for all $k \in \mathbb{N}$;

(iii) $\exists \hat{\delta} > 0$ such that, for all $z \in \mathcal{Z}$, there exist sequences $(\Delta_k(z))_{k\in\mathbb{N}_0} \in l_+(\mathfrak{F})$ and $(\lambda_k(z))_{k\in\mathbb{N}_0} \in l_+^1(\mathfrak{F})$ satisfying (ECI) with

$$\Delta_k(z) \geq \frac{\hat{\delta}}{2}\mathsf{E}\left(\| w_{k+1} - w_k \|_{Z_k M_k}^2 \,\middle|\, \mathcal{F}_k\right) \quad \text{P-a.s.}$$

for all $k \in \mathbb{N}_0$;

(iv) if $Z_k M_k (w_{k+1} - w_k) \to 0$ P-a.s. as $k \to \infty$, then every weak accumulation point of $(w_k)_{k\in\mathbb{N}_0}$ is P-a.s. in $\mathcal{Z}$, i.e.

$$\mathfrak{W}\left((w_k(\theta))_{k\in\mathbb{N}_0}\right) \subset \mathcal{Z}$$

for all $\theta \in \overline{\Theta}$ and some $\overline{\Theta} \in \mathcal{F}^1$;

(v) there exists $\overline{\Theta}' \in \mathcal{F}^1$ such that, for all $\theta \in \overline{\Theta}'$, there is a linear operator $R_\infty(\theta) \in \mathbb{L}(\mathcal{W}, \mathcal{W})$ satisfying

$$Z_{n_k} M_{n_k}(\theta) w \to R_\infty(\theta) w \quad \text{as} \quad k \to \infty$$

for all $w \in \mathcal{W}$ and every weakly convergent subsequence $\left(w_{n_k}(\theta)\right)_{k \in \mathbb{N}}$.

Then $w_k \rightharpoonup \overline{w}$ P-a.s. as $k \to \infty$ for some $\overline{w} \in \mathcal{R}(\mathcal{F}, \mathcal{Z})$.

Proof: We want to use Proposition 4.3.3 with $\mathcal{Z} = H^{-1}(0)$ and $R_k := Z_k M_k$ for all $k \in \mathbb{N}_0$. Therefore, we need to prove its assumptions.

First of all, condition (i) guarantees that Assumption 4.3.1 is true and Proposition 4.3.3 (i) is equivalent to Theorem 4.3.6 (ii). Condition (iii) and Lemma 4.3.4 imply that (SQF) holds for all $z \in \mathcal{Z}$, which verifies Proposition 4.3.3 (ii). Since Proposition 4.3.3 (iv) is equivalent to Theorem 4.3.6 (v), it only remains to show that $\mathfrak{W}\left((w_k)_{k \in \mathbb{N}_0}\right) \subset \mathcal{Z}$ P-a.s. But we have already seen that (SQF) holds for all $z \in \mathcal{Z}$, so we can employ Lemma 4.3.2 with $\chi_k \equiv 0$, $\vartheta_k = \Delta_k(z)$, $\eta_k = \lambda_k(z)$ and $\alpha_k = \frac{1}{2} \| w_k - z \|^2_{Z_k M_k}$, yielding $(\Delta_k)_{k \in \mathbb{N}_0} \in l^1_+(\mathfrak{F})$, i.e.

$$\sum_{k \in \mathbb{N}_0} \mathsf{E}\left(\frac{\hat{\delta}}{2} \| w_{k+1} - w_k \|^2_{Z_k M_k} \middle| \mathcal{F}_k \right) \leq \sum_{k \in \mathbb{N}_0} \Delta_k(z) < \infty \quad \text{P-a.s.}, \tag{4.3.33}$$

where the first part of the inequality stems from condition (iii). If we now apply the expectation together with [Kle13, Theorem 5.3 (vi)], we see that

$$\sum_{k \in \mathbb{N}_0} \mathsf{E}\left(\frac{\hat{\delta}}{2} \| w_{k+1} - w_k \|^2_{Z_k M_k} \right) < \infty. \tag{4.3.34}$$

It follows from [Kle13, Theorem 6.12 (i)] that

$$\frac{\hat{\delta}}{2} \| w_{k+1} - w_k \|^2_{Z_k M_k} \to 0 \quad \text{P-a.s.} \tag{4.3.35}$$

as $k \to \infty$. Since $Z_k M_k$ is P-a.s. self-adjoint and positive (semi-)definite for every $k \in \mathbb{N}_0$, there exists a self-adjoint and positive semi-definite $Q_k \in \mathcal{R}(\mathcal{F}_k, \mathbb{L}(\mathcal{W}, \mathcal{W}))$ such that $Q_k^2 := Q_k \circ Q_k = Z_k M_k$ [Rud91, Theorem 12.33]. It follows from (4.3.35) that

$$\| Q_k (w_{k+1} - w_k) \|^2_{\mathcal{W}} \to 0 \quad \text{P-a.s.}, \tag{4.3.36}$$

and together with the Cauchy–Schwarz inequality we get

$$\begin{aligned} \| Z_k M_k (w_{k+1} - w_k) \|^2_{\mathcal{W}} &= \left\langle Q_k^2 \circ Q_k (w_{k+1} - w_k), Q_k (w_{k+1} - w_k) \right\rangle_{\mathcal{W}} \\ &\leq \left\| Q_k^2 \right\|_{\mathbb{L}(\mathcal{W}, \mathcal{W})} \| Q_k (w_{k+1} - w_k) \|^2_{\mathcal{W}} \\ &\leq C^2 \| Q_k (w_{k+1} - w_k) \|^2_{\mathcal{W}} \to 0 \quad \text{P-a.s.} \end{aligned} \tag{4.3.37}$$

as $k \to \infty$, where we have used condition (ii) for the last inequality. Therefore, there exists $\overline{\Theta} \in \mathcal{F}^1$ such that $Z_k M_k(\theta) (w_{k+1}(\theta) - w_k(\theta)) \to 0$ as $k \to \infty$ for all $\theta \in \overline{\Theta}$. This verifies the assumption in (iv), thus the inclusion $\mathfrak{W}\left((w_k)_{k \in \mathbb{N}_0}\right) \subset \mathcal{Z}$ holds P-a.s. □

4.4 Scalar and Deterministic Step Sizes

In this section, we examine how the conditions of Theorem 4.3.6 can be satisfied in the case of scalar and deterministic step sizes. The goal is to develop assumptions that are easier to verify than those in the previous section. After stating some of these assumptions in the first subsection, we show that they enable us to prove that the central inequality (ECI) holds. In the third subsection, we summarize the results into a theorem proving the almost sure weak (and under additional assumptions *global*) convergence of Algorithm 4.1. Since this requires the primal iterates to stay inside a neighborhood where the fundamental assumptions are satisfied, we conclude with a technical result that could help proving this condition.

4.4.1 Fundamental Assumptions

In the following, we need a few fundamental assumptions, which are almost identical to [CMV19, Section 3.1]. First of all, we already require by Assumption 3.2.4 that K is continuously Fréchet differentiable. This can be used analogously to [CV20, Lemma 2.11] to show that K is locally Lipschitz continuous. In fact, since $K' \colon \mathcal{U} \to \mathbb{L}(\mathcal{U}, \mathcal{V})$ is continuous, we can find for a given $\bar{u} \in \mathcal{U}$ a constant $\delta > 0$ such that

$$\| K'(u) - K'(\bar{u}) \|_{\mathbb{L}(\mathcal{U},\mathcal{V})} \leq 1 \tag{4.4.1}$$

for all $u \in \mathbb{B}_\delta(\bar{u})$. This inequality implies

$$\| K'(u) \|_{\mathbb{L}(\mathcal{U},\mathcal{V})} \leq 1 + \| K'(\bar{u}) \|_{\mathbb{L}(\mathcal{U},\mathcal{V})} \tag{4.4.2}$$

for all $u \in \mathbb{B}_\delta(\bar{u})$, and together with the mean value theorem [DM13, Theorem 3.2.7] we get

$$\begin{aligned} \| K(u) - K(u') \|_{\mathcal{V}} &\leq \sup_{t \in [0,1]} \| K'(u' + t(u - u')) \|_{\mathbb{L}(\mathcal{U},\mathcal{V})} \| u - u' \|_{\mathcal{U}} \\ &\leq \left(1 + \| K'(\bar{u}) \|_{\mathbb{L}(\mathcal{U},\mathcal{V})}\right) \| u - u' \|_{\mathcal{U}} \end{aligned} \tag{4.4.3}$$

for all $u, u' \in \mathbb{B}_\delta(\bar{u})$, since $u' + t(u - u') \in \mathbb{B}_\delta(\bar{u})$ as well. This justifies the following definition.

Definition 4.4.1 (Locally Lipschitz K)

For a given $\bar{u} \in \mathcal{U}$, let $\mathcal{U}_K^1$ be a neighborhood of $\bar{u}$ such that

$$L := 1 + \| K'(\bar{u}) \|_{\mathbb{L}(\mathcal{U},\mathcal{V})}$$

is the Lipschitz constant of K on $\mathcal{U}_K^1$, i.e.

$$\| K(u) - K(u') \|_{\mathcal{V}} \leq L \| u - u' \|_{\mathcal{U}}$$

for all $u, u' \in \mathcal{U}_K^1$.

In addition, we need the following local property of the Fréchet derivative K'.

Assumption 4.4.2 (Locally Lipschitz K')

For a given $\bar{u} \in \mathcal{U}$, there exist a constant $L' \geq 0$ and a neighborhood $\mathcal{U}_K^2$ of $\bar{u}$ such that

$$\| K'(u) - K'(u') \|_{\mathbb{L}(\mathcal{U},\mathcal{V})} \leq L' \| u - u' \|_{\mathcal{U}}$$

for all $u, u' \in \mathcal{U}_K^2$.

Analogously to the reasoning preceding Definition 4.4.1, one can easily see that Assumption 4.4.2 is satisfied for every twice continuously differentiable $K \colon \mathcal{U} \to \mathcal{V}$. Before stating the next assumption, we derive a property which will be useful later. It follows from Assumption 4.4.2 together with the mean value theorem [DM13, Theorem 3.2.6]. The latter implies that

$$K(\bar{u}) = K(u) + K'(u)(\bar{u} - u) + \int_0^1 \Big(K'\big(u + t(\bar{u} - u)\big) - K'(u) \Big)(\bar{u} - u) \, \mathrm{d}t \tag{4.4.4}$$

for all $u, \bar{u} \in \mathcal{U}$. Applying the Cauchy–Schwarz inequality as well as Assumption 4.4.2 yields

$$\begin{aligned} &\langle K(\bar{u}) - K(u) - K'(u)(\bar{u} - u), v \rangle_{\mathcal{V}} \\ &\qquad \leq \int_0^1 \| K'(u + t(\bar{u} - u)) - K'(u) \|_{\mathbb{L}(\mathcal{U},\mathcal{V})} \| \bar{u} - u \|_{\mathcal{U}} \, \mathrm{d}t \, \| v \|_{\mathcal{V}} \\ &\qquad \leq \frac{L'}{2} \| \bar{u} - u \|_{\mathcal{U}}^2 \| v \|_{\mathcal{V}} \end{aligned} \tag{4.4.5}$$

for all $u, \bar{u} \in \mathcal{U}_K^2$ and $v \in \mathcal{V}$. If we now choose $v := K(\bar{u}) - K(u) - K'(u)(\bar{u} - u)$, we get the desired inequality

$$\| K(\bar{u}) - K(u) - K'(u)(\bar{u} - u) \|_{\mathcal{V}} \leq \frac{L'}{2} \| \bar{u} - u \|_{\mathcal{U}}^2 . \tag{4.4.6}$$

Since we know from the beginning of Section 4.1 that CVaR_β^* is convex, we also know that $\partial \mathrm{CVaR}_\beta^*$ is a monotone operator. Similarly, ∂G is monotone due to the postulated convexity of G. However, G can also be strongly convex and hence ∂G strongly monotone, according to Lemma 2.2.11. This case is addressed in the following assumption, where we use a local version of strong monotonicity.

Assumption 4.4.3 (Monotone ∂G)

For given $\bar{u} \in \mathcal{U}$ and $\bar{v} \in \mathcal{V}$, there exist a constant $\gamma_G \geq 0$ and a neighborhood $\mathcal{U}_G$ of $\bar{u}$ such that ∂G is strongly monotone with factor γ_G at $\bar{u}$ for $\bar{z} := -K'(\bar{u})^* \bar{v}$, i.e.

$$\langle z - \bar{z}, u - \bar{u} \rangle_{\mathcal{U}} \geq \gamma_G \| u - \bar{u} \|_{\mathcal{U}}^2 \quad \text{for all } u \in \mathcal{U}_G, \ z \in \partial G(u).$$

Here, strong monotonicity with factor $\gamma_G = 0$ just means *monotonicity*, which is compliant with Definition 2.2.10.

Since CVaR_β^* is not strongly convex and therefore $\partial \mathrm{CVaR}_\beta^*$ not strongly monotone, we do not need to assume anything regarding $\partial \mathrm{CVaR}_\beta^*$ here. However, if G is strongly monotone, we might be able to accelerate the algorithm as in [CMV19] by using a factor $\tilde{\gamma}_G$, which is introduced in the following assumption.

Assumption 4.4.4

For given $\bar{u} \in \mathcal{U}$, $\bar{v} \in \mathcal{V}$, and $\gamma_G \geq \tilde{\gamma}_G \geq 0$, there exist a constant $\gamma_{\mathcal{U}} > 0$ and a neighborhood $\mathcal{U}_K^3$ of $\bar{u}$ such that

$$(\gamma_G - \tilde{\gamma}_G) \| u - \bar{u} \|_{\mathcal{U}}^2 + \left\langle \left(K'(u) - K'(\bar{u})\right)(u - \bar{u}), \bar{v} \right\rangle_{\mathcal{V}} \geq \gamma_{\mathcal{U}} \| u - \bar{u} \|_{\mathcal{U}}^2$$

for all $u \in \mathcal{U}_G \cap \mathcal{U}_K^3$.

Note that this assumption trivially holds if K is linear. In general, Assumption 4.4.3 and Assumption 4.4.4 imply that

$$\langle z - \bar{z}, u - \bar{u} \rangle_{\mathcal{U}} + \left\langle K'(u)^* \bar{v} - K'(\bar{u})^* \bar{v}, u - \bar{u} \right\rangle_{\mathcal{U}} \geq \gamma_{\mathcal{U}} \| u - \bar{u} \|_{\mathcal{U}}^2 \tag{4.4.7}$$

for all $u \in \mathcal{U}_G \cap \mathcal{U}_K^3$ and $z \in \partial G(u)$. According to Definition 2.2.10, this is equivalent to (a local version of) the strong monotonicity of the set-valued mapping $\partial G + K'(\cdot)^* \bar{v} \colon \mathcal{U} \rightrightarrows \mathcal{U}$. On the other hand, if the mapping $\mathcal{U} \ni u \mapsto G(u) + \langle \bar{v}, K(u) \rangle_{\mathcal{V}}$ is (locally) strongly convex, Lemma 2.2.11 also implies the local strong monotonicity of $\partial G + K'(\cdot)^* \bar{v}$. Hence, taking a look at the saddle-point formulation in Problem 3.1.3, one can see that Assumption 4.4.4 acts as a local convexity assumption similar to sufficient second-order conditions; see [CMV19, p. 8].

Furthermore, it is a simplification of the so-called *three-point condition*, which is a combination of a smoothness estimate of K as well as a second-order growth condition. It is introduced in the following lemma, which is a simplified version of [CMV19, Proposition 3.2].

Lemma 4.4.5 (Three-Point Condition on K)

Suppose Assumption 4.4.2 (Locally Lipschitz K), Assumption 4.4.3 (Monotone ∂G), and Assumption 4.4.4 hold. Then, the following three-point condition is true for all $u, u' \in \mathcal{U}_K^2 \cap \mathcal{U}_K^3$ and $\xi > 0$:

$$\left\langle \left(K'(u') - K'(\bar{u})\right)^* \bar{v}, u - \bar{u} \right\rangle_{\mathcal{U}} + (\gamma_G - \tilde{\gamma}_G) \| u - \bar{u} \|_{\mathcal{U}}^2 \geq (\gamma_{\mathcal{U}} - \xi) \| u - \bar{u} \|_{\mathcal{U}}^2 - \frac{(L')^2}{4\xi} \| u - u' \|_{\mathcal{U}}^2 \| \bar{v} \|_{\mathcal{V}}^2 .$$

Proof: Let $u, u' \in \mathcal{U}_K^2 \cap \mathcal{U}_K^3$ be arbitrary and denote by Λ the left-hand side of the inequality to be proven. Applying Assumption 4.4.4 as well as the Cauchy-Schwarz inequality yields

$$\begin{aligned} \Lambda &:= \left\langle \left(K'(u') - K'(\bar{u})\right)^* \bar{v}, u - \bar{u} \right\rangle_{\mathcal{U}} + (\gamma_G - \tilde{\gamma}_G) \| u - \bar{u} \|_{\mathcal{U}}^2 \\ &\geq \gamma_{\mathcal{U}} \| u - \bar{u} \|_{\mathcal{U}}^2 - \left\langle \left(K'(u) - K'(u')\right)(u - \bar{u}), \bar{v} \right\rangle_{\mathcal{V}} \\ &\geq \gamma_{\mathcal{U}} \| u - \bar{u} \|_{\mathcal{U}}^2 - \| K'(u) - K'(u') \|_{\mathbb{L}(\mathcal{U},\mathcal{V})} \| u - \bar{u} \|_{\mathcal{U}} \| \bar{v} \|_{\mathcal{V}} . \end{aligned} \tag{4.4.8}$$

If we now use Assumption 4.4.2 and the scaled version of Young's inequality, which is also

referred to as *Peter-Paul inequality*, we get

$$\begin{aligned}\Lambda &\geq \gamma_{\mathcal{U}} \| u - \bar{u} \|_{\mathcal{U}}^2 - L' \| u - u' \|_{\mathcal{U}} \| u - \bar{u} \|_{\mathcal{U}} \| \bar{v} \|_{\mathcal{V}} \\ &\geq (\gamma_{\mathcal{U}} - \xi) \| u - \bar{u} \|_{\mathcal{U}}^2 - \frac{(L')^2}{4\xi} \| u - u' \|_{\mathcal{U}}^2 \| \bar{v} \|_{\mathcal{V}}^2\end{aligned} \tag{4.4.9}$$

for any $\xi > 0$, which concludes the proof. □

In a final step, we want to combine all the defined neighborhoods into one definition.

Assumption 4.4.6 (Desired Ball)

For given $\bar{u} \in \mathcal{U}$ and $\bar{v} \in \mathcal{V}$, let $\mathcal{U}_K^1$, $\mathcal{U}_K^2$, and $\mathcal{U}_G$ be the neighborhoods as defined in Definition 4.4.1, Assumption 4.4.2, and Assumption 4.4.3, respectively. For γ_G from Assumption 4.4.3 and an arbitrary $\tilde{\gamma}_G \in [0, \gamma_G]$, let further $\mathcal{U}_K^3$ be the neighborhood as in Assumption 4.4.4. Let $\rho_{\mathcal{U}}(\bar{u}) \in [0, \infty]$ be a radius such that $\mathbb{B}_{\rho_{\mathcal{U}}(\bar{u})}(\bar{u}) \subset \mathcal{U}_K^1 \cap \mathcal{U}_K^2 \cap \mathcal{U}_K^3 \cap \mathcal{U}_G$. We define the *desired ball* $\mathcal{B}(\bar{u}, \bar{v}) \subset \mathcal{U}$ as

$$\mathcal{B}(\bar{u}, \bar{v}) := \mathbb{B}_{\rho_{\mathcal{U}}(\bar{u})}(\bar{u})$$

and assume that it is non-empty.

Note that, since $\mathcal{B}(\bar{u}, \bar{v})$ is closed and convex, it is also weakly closed [CV20, Lemma 1.10].

4.4.2 Satisfaction of Central Inequality

In this subsection we formulate estimates which guarantee that the inequality (ECI) holds in the case of scalar and deterministic step sizes. For this purpose, we develop explicit bounds on the step size and testing operators, which also require that the iterates stay in a neighborhood of the critical point. First, we assume that the step size operator defined in (4.1.14) takes the form

$$W_k := \begin{pmatrix} \tau_k \mathrm{Id}_{\mathcal{U}} & 0 \\ 0 & \sigma \mathrm{Id}_{\mathcal{V}} \end{pmatrix} \in \mathbb{L}(\mathcal{W}, \mathcal{W}) \tag{4.4.10}$$

for all $k \in \mathbb{N}_0$ with $\tau_k, \sigma \in (0, \infty)$, i.e. $T_k := \tau_k \mathrm{Id}_{\mathcal{U}}$ and $\Sigma_k := \sigma \mathrm{Id}_{\mathcal{V}}$. Consequently, the previously defined testing operator becomes

$$Z_k := \begin{pmatrix} \varphi_k \mathrm{Id}_{\mathcal{U}} & 0 \\ 0 & \psi \mathrm{Id}_{\mathcal{V}} \end{pmatrix} \in \mathbb{L}(\mathcal{W}, \mathcal{W}) \tag{4.4.11}$$

for all $k \in \mathbb{N}_0$ with $\varphi_k, \psi \in (0, \infty)$. Note that σ and ψ are intentionally written without index k because they are not supposed to change. Multiplying the preconditioning operator M_k by Z_k yields

$$Z_k M_k = \begin{pmatrix} \varphi_k \mathrm{Id}_{\mathcal{U}} & \varphi_k \tau_k K'(u_k)^* \\ \psi \sigma K'(u_k) & \psi \mathrm{Id}_{\mathcal{V}} \end{pmatrix} \in \mathcal{R}(\mathcal{F}_k, \mathbb{L}(\mathcal{W}, \mathcal{W})) \tag{4.4.12}$$

for all $k \in \mathbb{N}_0$. The following assumption specifies some relationships between the step size and testing operators.

Assumption 4.4.7 (Step-Size-Testing-Relation)

Let $\tilde{\gamma}_G \geq 0$, $\kappa \in (0,1)$, and $\eta > 0$. We require that the following relationships hold for all $k \in \mathbb{N}_0$:

(i) $\varphi_k \tau_k = \psi\sigma = \eta$

(ii) $\varphi_{k+1} = \varphi_k \left(1 + 2\tilde{\gamma}_G \tau_k\right)$

(iii) $(1-\kappa)\,\mathrm{Id}_{\mathcal{V}} \geq \sigma\tau_k K'(u_k) \circ K'(u_k)^*$ P-a.s.

An example of possible step size choices satisfying this assumption is presented in Section 6.1. Note that a sufficient condition to satisfy Assumption 4.4.7 (iii) is

$$\sigma\tau_k \,\|K'(u_k)\|^2_{\mathbb{L}(\mathcal{U},\mathcal{V})} < 1 \quad \text{P-a.s.} \tag{4.4.13}$$

In this case, we can find a $\kappa \in (0,1)$ such that

$$\|K'(u_k)^*\|^2_{\mathbb{L}(\mathcal{V},\mathcal{U})} = \|K'(u_k)\|^2_{\mathbb{L}(\mathcal{U},\mathcal{V})} \leq \frac{1-\kappa}{\sigma\tau_k} \quad \text{P-a.s.,} \tag{4.4.14}$$

and since $\|K'(u_k)^* v\|^2_{\mathcal{U}} \leq \|K'(u_k)^*\|^2_{\mathbb{L}(\mathcal{V},\mathcal{U})} \|v\|^2_{\mathcal{V}}$ for all $v \in \mathcal{V}$, this implies that

$$\|K'(u_k)^* v\|^2_{\mathcal{U}} \leq \frac{1-\kappa}{\sigma\tau_k} \|v\|^2_{\mathcal{V}} \quad \text{P-a.s.} \tag{4.4.15}$$

for all $v \in \mathcal{V}$. Now, the definition of the norm $\|\cdot\|_{\mathcal{U}}$ and multiplication by $\sigma\tau_k$ yield

$$\sigma\tau_k \left\langle \left(K'(u_k) \circ K'(u_k)^*\right) v, v \right\rangle_{\mathcal{U}} \leq (1-\kappa) \langle v, v \rangle_{\mathcal{U}} \quad \text{P-a.s.} \tag{4.4.16}$$

for all $v \in \mathcal{V}$, which means that $(1-\kappa)\,\mathrm{Id}_{\mathcal{V}} - \sigma\tau_k K'(u_k) \circ K'(u_k)^*$ is P-a.s. positive semi-definite.

We can now formulate the following lemma, which also gives us a justification for using $\|\cdot\|_{Z_k M_k}$ to denote at least a semi-norm. It is based on [CMV19, Lemma 3.4].

Lemma 4.4.8

Let $k \in \mathbb{N}_0$ and suppose Assumption 4.4.7 holds. Then $Z_k M_k$ is self-adjoint P-a.s. and satisfies

$$Z_k M_k \geq \begin{pmatrix} \delta\varphi_k \mathrm{Id}_{\mathcal{U}} & 0 \\ 0 & \frac{\kappa-\delta}{1-\delta}\psi \mathrm{Id}_{\mathcal{V}} \end{pmatrix} \quad \text{P-a.s.}$$

for any $\delta \in (0, \kappa]$.

Proof: Let $k \in \mathbb{N}_0$. Applying Assumption 4.4.7 (i) to (4.4.12), we can easily see that $Z_k M_k$ is self-adjoint P-a.s. Proving the desired inequality is equivalent to showing that

$$Z_k M_k - \begin{pmatrix} \delta\varphi_k \mathrm{Id}_{\mathcal{U}} & 0 \\ 0 & \frac{\kappa-\delta}{1-\delta}\psi \mathrm{Id}_{\mathcal{V}} \end{pmatrix} = \begin{pmatrix} (1-\delta)\varphi_k \mathrm{Id}_{\mathcal{U}} & \varphi_k\tau_k K'(u_k)^* \\ \psi\sigma K'(u_k) & \frac{1-\kappa}{1-\delta}\psi \mathrm{Id}_{\mathcal{V}} \end{pmatrix} \tag{4.4.17}$$

is P-a.s. positive semi-definite. Therefore, we consider for every $w = (u, v) \in \mathcal{U} \times \mathcal{V}$ the

inner product

$$\begin{aligned} Q &:= \left\langle \begin{pmatrix} (1-\delta)\varphi_k \mathrm{Id}_{\mathcal{U}} & \varphi_k\tau_k K'(u_k)^* \\ \psi\sigma K'(u_k) & \frac{1-\kappa}{1-\delta}\psi \mathrm{Id}_{\mathcal{V}} \end{pmatrix} w, w \right\rangle_{\mathcal{W}} \\ &= (1-\delta)\varphi_k \| u \|_{\mathcal{U}}^2 - 2\varphi_k\tau_k \left\langle K'(u_k)^* v, -u \right\rangle_{\mathcal{U}} + \frac{1-\kappa}{1-\delta}\psi \| v \|_{\mathcal{V}}^2 \quad \text{P-a.s.,} \end{aligned} \tag{4.4.18}$$

where we have used the identity $\psi\sigma = \varphi_k\tau_k$ from Assumption 4.4.7 (i). For the middle term, we use the Cauchy–Schwarz as well as the scaled version of Young's inequality to estimate

$$\begin{aligned} 2\varphi_k\tau_k \left\langle K'(u_k)^* v, -u \right\rangle_{\mathcal{U}} &\leq 2\varphi_k\tau_k \left\| K'(u_k)^* v \right\|_{\mathcal{U}} \| u \|_{\mathcal{U}} \\ &\leq \varphi_k(1-\delta) \| u \|_{\mathcal{U}}^2 + \frac{\varphi_k\tau_k^2}{1-\delta} \left\| K'(u_k)^* v \right\|_{\mathcal{U}}^2 \quad \text{P-a.s.} \end{aligned} \tag{4.4.19}$$

with $\delta \in (0, \kappa]$. Plugging this into the previous equation yields

$$\begin{aligned} Q &\geq \frac{1-\kappa}{1-\delta}\psi \| v \|_{\mathcal{V}}^2 - \frac{\varphi_k\tau_k^2}{1-\delta} \left\| K'(u_k)^* v \right\|_{\mathcal{U}}^2 \\ &= (1-\delta)^{-1} \Big(\langle (1-\kappa)\psi v, v \rangle_{\mathcal{V}} - \left\langle \varphi_k\tau_k^2 \left(K'(u_k) \circ K'(u_k)^* \right) v, v \right\rangle_{\mathcal{V}} \Big) \quad \text{P-a.s.} \end{aligned} \tag{4.4.20}$$

Now, using the identity $\eta = \varphi_k\tau_k$ as well as Assumption 4.4.7 (iii) multiplied by ψ yields $Q \geq 0$ P-a.s., which implies that (4.4.17) is P-a.s. positive semi-definite. □

The following technical result gives us a first estimate of a bound for some of the terms of the sum occurring in (ECI). It is based on [CMV19, Lemma 3.6].

Lemma 4.4.9

Let $(\bar{u}, \bar{v}) = \bar{w} \in \mathcal{W}$, and $k \in \mathbb{N}_0$ fixed. Suppose that, for some radius $\rho_{\mathcal{U}}(\bar{u}) \geq 0$, Assumption 4.4.6 (Desired Ball) is satisfied (i.e. $\mathcal{B}(\bar{u}, \bar{v})$ is non-empty) and that the inclusion $u_k, u_{k+1} \in \mathcal{B}(\bar{u}, \bar{v})$ is P-a.s. true. Furthermore, assume that $\mathfrak{F}$ is a compatible filtration as in Definition 4.2.1, the inequality $\| v_{k+1} - \bar{v} \|_{\mathcal{V}} < \gamma_{\mathcal{U}} \frac{2}{3L'}$ holds P-a.s., and let $\xi \in \left(0, \gamma_{\mathcal{U}} - \frac{3L'}{2} \| v_{k+1} - \bar{v} \|_{\mathcal{V}}\right]$. If Assumption 4.4.7 holds, then

$$\begin{aligned} &\mathsf{E}\left(\left\langle \tilde{H}_k(w_{k+1}), w_{k+1} - \bar{w} \right\rangle_{Z_k} \middle| \mathcal{F}_k \right) - \frac{1}{2}\mathsf{E}\left(\| w_{k+1} - \bar{w} \|_{Z_{k+1}M_{k+1} - Z_kM_k}^2 \middle| \mathcal{F}_k \right) \\ &\qquad \geq -\eta\mathsf{E}\left(\left(\frac{(L')^2}{4\xi} \| \bar{v} \|_{\mathcal{V}}^2 + \frac{3L'}{2} \| v_{k+1} - \bar{v} \|_{\mathcal{V}} \right) \| u_{k+1} - u_k \|_{\mathcal{U}}^2 \middle| \mathcal{F}_k \right) \\ &\qquad\quad - \eta\mathsf{E}\left(\left\| \left(\hat{K}_k'(u_k)^* - K'(u_k)^* \right) (2v_{k+1} - v_k) \right\|_{\mathcal{U}} \| u_{k+1} - \bar{u} \|_{\mathcal{U}} \middle| \mathcal{F}_k \right) \\ &\qquad\quad - \eta\mathsf{E}\left(\left\| \hat{K}_k(u_k) - K(u_k) \right\|_{\mathcal{V}} \| v_{k+1} - \bar{v} \|_{\mathcal{V}} \middle| \mathcal{F}_k \right) \end{aligned}$$

holds P-a.s.

Proof: First, we define

$$Q := \left\langle \tilde{H}_k\left(w_{k+1}\right), w_{k+1} - \overline{w} \right\rangle_{Z_k} - \frac{1}{2} \parallel w_{k+1} - \overline{w} \parallel^2_{Z_{k+1}M_{k+1} - Z_k M_k} . \tag{4.4.21}$$

Unless stated otherwise, all equations in this proof are meant to hold P-a.s. It follows from (4.4.12) and Assumption 4.4.7 that

$$Z_{k+1}M_{k+1} - Z_k M_k = \begin{pmatrix} \left(\varphi_{k+1} - \varphi_k\right) \mathrm{Id}_{\mathcal{U}} & \eta K'\left(u_{k+1}\right)^* - \eta K'\left(u_k\right)^* \\ \eta K'\left(u_{k+1}\right) - \eta K'\left(u_k\right) & 0 \end{pmatrix}. \tag{4.4.22}$$

Since Assumption 4.4.7 (i) and (ii) imply $\varphi_{k+1} - \varphi_k = 2\eta\tilde{\gamma}_G$, we can expand

$$\begin{aligned} &\frac{1}{2} \parallel w_{k+1} - \overline{w} \parallel^2_{Z_{k+1}M_{k+1} - Z_k M_k} \\ &\qquad = \eta\tilde{\gamma}_G \parallel u_{k+1} - \overline{u} \parallel^2_{\mathcal{U}} + \eta \left\langle \left(K'\left(u_{k+1}\right) - K'\left(u_k\right)\right)\left(u_{k+1} - \overline{u}\right), v_{k+1} - \overline{v} \right\rangle_{\mathcal{V}} . \end{aligned} \tag{4.4.23}$$

Together with the definition of $\tilde{H}_k$ in (4.2.2) as well as the discrepancy function D_k, we then have

$$\begin{aligned} Q = &\left\langle H\left(w_{k+1}\right), w_{k+1} - \overline{w} \right\rangle_{Z_k W_k} + \left\langle D_k^{LIN}\left(w_{k+1}\right), w_{k+1} - \overline{w} \right\rangle_{Z_k} \\ &+ \left\langle D_k^{CGF}\left(w_{k+1}\right), w_{k+1} - \overline{w} \right\rangle_{Z_k} - \eta\tilde{\gamma}_G \parallel u_{k+1} - \overline{u} \parallel^2_{\mathcal{U}} \\ &- \eta \left\langle \left(K'\left(u_{k+1}\right) - K'\left(u_k\right)\right)\left(u_{k+1} - \overline{u}\right), v_{k+1} - \overline{v} \right\rangle_{\mathcal{V}} . \end{aligned} \tag{4.4.24}$$

The first term of the sum on the right-hand side can be expanded as follows, using the definition of H in (4.1.15):

$$\begin{aligned} \left\langle H\left(w_{k+1}\right), w_{k+1} - \overline{w} \right\rangle_{Z_k W_k} = &\eta \left\langle \partial G\left(u_{k+1}\right) + K'\left(u_{k+1}\right)^* v_{k+1}, u_{k+1} - \overline{u} \right\rangle_{\mathcal{U}} \\ &+ \eta \left\langle \partial \mathrm{CVaR}^*_\beta\left(v_{k+1}\right) - K\left(u_{k+1}\right), v_{k+1} - \overline{v} \right\rangle_{\mathcal{V}} . \end{aligned} \tag{4.4.25}$$

Now let $z_1 := -K'\left(\overline{u}\right)^* \overline{v} \in \partial G\left(\overline{u}\right)$ and $z_2 := K\left(\overline{u}\right) \in \partial \mathrm{CVaR}^*_\beta\left(\overline{v}\right)$. Addition and subtraction of these terms yields

$$\begin{aligned} &\left\langle H\left(w_{k+1}\right), w_{k+1} - \overline{w} \right\rangle_{Z_k W_k} \\ &\qquad = \eta \left\langle \partial G\left(u_{k+1}\right) - z_1, u_{k+1} - \overline{u} \right\rangle_{\mathcal{U}} + \eta \left\langle K'\left(u_{k+1}\right)^* v_{k+1} + z_1, u_{k+1} - \overline{u} \right\rangle_{\mathcal{U}} \\ &\qquad + \eta \left\langle \partial \mathrm{CVaR}^*_\beta\left(v_{k+1}\right) - z_2, v_{k+1} - \overline{v} \right\rangle_{\mathcal{V}} + \eta \left\langle z_2 - K\left(u_{k+1}\right), v_{k+1} - \overline{v} \right\rangle_{\mathcal{V}} . \end{aligned} \tag{4.4.26}$$

The monotonicity of $\partial \mathrm{CVaR}^*_\beta$ as well as the (strong) monotonicity of ∂G with factor γ_G (see Assumption 4.4.3) imply

$$\begin{aligned} &\left\langle H\left(w_{k+1}\right), w_{k+1} - \overline{w} \right\rangle_{Z_k W_k} \\ &\qquad \geq \eta\gamma_G \parallel u_{k+1} - \overline{u} \parallel^2_{\mathcal{U}} + \eta \left\langle K'\left(u_{k+1}\right)^* v_{k+1} - K'\left(\overline{u}\right)^* \overline{v}, u_{k+1} - \overline{u} \right\rangle_{\mathcal{U}} \\ &\qquad + \eta \left\langle K\left(\overline{u}\right) - K\left(u_{k+1}\right), v_{k+1} - \overline{v} \right\rangle_{\mathcal{V}} . \end{aligned} \tag{4.4.27}$$

We now recall the definitions of D_k^{LIN} and D_k^{CGF} in (4.1.16) and (4.1.17), respectively, which imply

$$\begin{aligned}\left\langle D_k^{LIN}\left(w_{k+1}\right), w_{k+1}-\overline{w}\right\rangle_{Z_k} &= \eta\left\langle \left(K'\left(u_k\right)^* - K'\left(u_{k+1}\right)^*\right)v_{k+1}, u_{k+1}-\overline{u}\right\rangle_{\mathcal{U}} \\ &\quad + \eta\left\langle K\left(u_{k+1}\right) - K\left(u_k\right) - K'\left(u_k\right)\left(u_{k+1}-u_k\right), v_{k+1}-\overline{v}\right\rangle_{\mathcal{V}}\end{aligned} \tag{4.4.28}$$

and

$$\begin{aligned}\left\langle D_k^{CGF}\left(w_{k+1}\right), w_{k+1}-\overline{w}\right\rangle_{Z_k} &= \eta\left\langle \left(\hat{K}_k'\left(u_k\right)^* - K'\left(u_k\right)^*\right)\left(2v_{k+1}-v_k\right), u_{k+1}-\overline{u}\right\rangle_{\mathcal{U}} \\ &\quad + \eta\left\langle K\left(u_k\right) - \hat{K}_k\left(u_k\right), v_{k+1}-\overline{v}\right\rangle_{\mathcal{V}}.\end{aligned} \tag{4.4.29}$$

Combining (4.4.24), (4.4.27), (4.4.28), and (4.4.29) yields

$$\begin{aligned}\eta^{-1}Q \geq\ & \left(\gamma_G - \tilde{\gamma}_G\right)\left\| u_{k+1}-\overline{u}\right\|_{\mathcal{U}}^2 \\ &+ \left\langle K'\left(u_{k+1}\right)^* v_{k+1} - K'\left(\overline{u}\right)^*\overline{v}, u_{k+1}-\overline{u}\right\rangle_{\mathcal{U}} \\ &+ \left\langle K\left(\overline{u}\right) - K\left(u_{k+1}\right), v_{k+1}-\overline{v}\right\rangle_{\mathcal{V}} \\ &+ \left\langle \left(K'\left(u_k\right)^* - K'\left(u_{k+1}\right)^*\right)v_{k+1}, u_{k+1}-\overline{u}\right\rangle_{\mathcal{U}} \\ &+ \left\langle K\left(u_{k+1}\right) - K\left(u_k\right) - K'\left(u_k\right)\left(u_{k+1}-u_k\right), v_{k+1}-\overline{v}\right\rangle_{\mathcal{V}} \\ &- \left\langle \left(K'\left(u_{k+1}\right) - K'\left(u_k\right)\right)\left(u_{k+1}-\overline{u}\right), v_{k+1}-\overline{v}\right\rangle_{\mathcal{V}} \\ &+ \left\langle \left(\hat{K}_k'\left(u_k\right)^* - K'\left(u_k\right)^*\right)\left(2v_{k+1}-v_k\right), u_{k+1}-\overline{u}\right\rangle_{\mathcal{U}} \\ &+ \left\langle K\left(u_k\right) - \hat{K}_k\left(u_k\right), v_{k+1}-\overline{v}\right\rangle_{\mathcal{V}}.\end{aligned} \tag{4.4.30}$$

Rearranging after addition and subtraction of the terms $\langle K'\left(u_{k+1}\right)\left(u_{k+1}-\overline{u}\right), v_{k+1}-\overline{v}\rangle_{\mathcal{V}}$ and $\langle K'\left(u_k\right)\left(u_{k+1}-\overline{u}\right), \overline{v}\rangle_{\mathcal{V}}$ yields

$$\begin{aligned}\eta^{-1}Q \geq\ & \left(\gamma_G - \tilde{\gamma}_G\right)\left\| u_{k+1}-\overline{u}\right\|_{\mathcal{U}}^2 + \left\langle \left(K'\left(u_k\right) - K'\left(\overline{u}\right)\right)^*\overline{v}, u_{k+1}-\overline{u}\right\rangle_{\mathcal{U}} \\ &+ \left\langle K\left(\overline{u}\right) - K\left(u_{k+1}\right) - K'\left(u_{k+1}\right)\left(\overline{u}-u_{k+1}\right), v_{k+1}-\overline{v}\right\rangle_{\mathcal{V}} \\ &+ \left\langle K\left(u_{k+1}\right) - K\left(u_k\right) - K'\left(u_k\right)\left(u_{k+1}-u_k\right), v_{k+1}-\overline{v}\right\rangle_{\mathcal{V}} \\ &+ 2\left\langle \left(K'\left(u_k\right) - K'\left(u_{k+1}\right)\right)\left(u_{k+1}-\overline{u}\right), v_{k+1}-\overline{v}\right\rangle_{\mathcal{V}} \\ &+ R,\end{aligned} \tag{4.4.31}$$

where

$$\begin{aligned}R :=\ & \left\langle \left(\hat{K}_k'\left(u_k\right)^* - K'\left(u_k\right)^*\right)\left(2v_{k+1}-v_k\right), u_{k+1}-\overline{u}\right\rangle_{\mathcal{U}} \\ &+ \left\langle K\left(u_k\right) - \hat{K}_k\left(u_k\right), v_{k+1}-\overline{v}\right\rangle_{\mathcal{V}}.\end{aligned} \tag{4.4.32}$$

In order to further estimate (4.4.31), we now consider each line of the right-hand side separately:

1. On the first line, we can apply Lemma 4.4.5, i.e.

$$\begin{aligned}(\gamma_G - \tilde{\gamma}_G) \| u_{k+1} - \bar{u} \|_{\mathcal{U}}^2 + \left\langle \left(K'(u_k) - K'(\bar{u})\right)^* \bar{v}, u_{k+1} - \bar{u} \right\rangle_{\mathcal{U}} \\ \geq (\gamma_{\mathcal{U}} - \xi) \| u_{k+1} - \bar{u} \|_{\mathcal{U}}^2 - \frac{(L')^2}{4\xi} \| u_{k+1} - u_k \|_{\mathcal{U}}^2 \| \bar{v} \|_{\mathcal{V}}^2 \end{aligned} \tag{4.4.33}$$

   for all $\xi > 0$.

2. For the second line, we use the Cauchy-Schwarz inequality and (4.4.6) to estimate

$$\begin{aligned}&\langle K(\bar{u}) - K(u_{k+1}) - K'(u_{k+1})(\bar{u} - u_{k+1}), v_{k+1} - \bar{v} \rangle_{\mathcal{V}} \\ &\qquad \geq - \| v_{k+1} - \bar{v} \|_{\mathcal{V}} \| K(\bar{u}) - K(u_{k+1}) - K'(u_{k+1})(\bar{u} - u_{k+1}) \|_{\mathcal{V}} \\ &\qquad \geq - \frac{L'}{2} \| v_{k+1} - \bar{v} \|_{\mathcal{V}} \| u_{k+1} - \bar{u} \|_{\mathcal{U}}^2 . \end{aligned} \tag{4.4.34}$$

3. Using (4.4.5), the third line can be estimated as follows:

$$\begin{aligned}\langle K(u_{k+1}) - K(u_k) - K'(u_k)(u_{k+1} - u_k), v_{k+1} - \bar{v} \rangle_{\mathcal{V}} \\ \geq - \frac{L'}{2} \| v_{k+1} - \bar{v} \|_{\mathcal{V}} \| u_{k+1} - u_k \|_{\mathcal{U}}^2 . \end{aligned} \tag{4.4.35}$$

4. For the fourth line, we need the Cauchy-Schwarz inequality, Assumption 4.4.2, and Young's inequality, to estimate

$$\begin{aligned}&2 \left\langle \left(K'(u_k) - K'(u_{k+1})\right)(u_{k+1} - \bar{u}), v_{k+1} - \bar{v} \right\rangle_{\mathcal{V}} \\ &\qquad \geq -2L' \| u_k - u_{k+1} \|_{\mathcal{U}} \| u_{k+1} - \bar{u} \|_{\mathcal{U}} \| v_{k+1} - \bar{v} \|_{\mathcal{V}} \\ &\qquad \geq -L' \| v_{k+1} - \bar{v} \|_{\mathcal{V}} \| u_{k+1} - u_k \|_{\mathcal{U}}^2 - L' \| v_{k+1} - \bar{v} \|_{\mathcal{V}} \| u_{k+1} - \bar{u} \|_{\mathcal{U}}^2 . \end{aligned} \tag{4.4.36}$$

Putting all these estimates (4.4.31), (4.4.33), (4.4.34), (4.4.35), and (4.4.36) together, we get

$$\begin{aligned}\eta^{-1} Q \geq (\gamma_{\mathcal{U}} - \xi) \| u_{k+1} - \bar{u} \|_{\mathcal{U}}^2 - \frac{(L')^2}{4\xi} \| u_{k+1} - u_k \|_{\mathcal{U}}^2 \| \bar{v} \|_{\mathcal{V}}^2 \\ - \frac{L'}{2} \| v_{k+1} - \bar{v} \|_{\mathcal{V}} \| u_{k+1} - \bar{u} \|_{\mathcal{U}}^2 - \frac{L'}{2} \| v_{k+1} - \bar{v} \|_{\mathcal{V}} \| u_{k+1} - u_k \|_{\mathcal{U}}^2 \\ - L' \| v_{k+1} - \bar{v} \|_{\mathcal{V}} \| u_{k+1} - u_k \|_{\mathcal{U}}^2 - L' \| v_{k+1} - \bar{v} \|_{\mathcal{V}} \| u_{k+1} - \bar{u} \|_{\mathcal{U}}^2 + R, \end{aligned} \tag{4.4.37}$$

which can be rearranged to

$$\begin{aligned}\eta^{-1} Q \geq \left(\gamma_{\mathcal{U}} - \xi - \frac{3L'}{2} \| v_{k+1} - \bar{v} \|_{\mathcal{V}} \right) \| u_{k+1} - \bar{u} \|_{\mathcal{U}}^2 \\ - \left(\frac{(L')^2}{4\xi} \| \bar{v} \|_{\mathcal{V}}^2 + \frac{3L'}{2} \| v_{k+1} - \bar{v} \|_{\mathcal{V}} \right) \| u_{k+1} - u_k \|_{\mathcal{U}}^2 + R. \end{aligned} \tag{4.4.38}$$

Since we assumed that $\xi \leq \gamma_{\mathcal{U}} - \frac{3L'}{2} \| v_{k+1} - \bar{v} \|_{\mathcal{V}}$, the inequality (4.4.38) implies

$$\eta^{-1} Q \geq -\left(\frac{(L')^2}{4\xi} \| \bar{v} \|_{\mathcal{V}}^2 + \frac{3L'}{2} \| v_{k+1} - \bar{v} \|_{\mathcal{V}} \right) \| u_{k+1} - u_k \|_{\mathcal{U}}^2 + R. \tag{4.4.39}$$

Now we consider the definition of R in (4.4.32) again and use the Cauchy-Schwarz inequality to estimate

$$\begin{aligned} R \geq &- \left\| \left(\hat{K}_k' (u_k)^* - K' (u_k)^* \right) (2v_{k+1} - v_k) \right\|_{\mathcal{U}} \| u_{k+1} - \bar{u} \|_{\mathcal{U}} \\ &- \left\| \hat{K}_k (u_k) - K (u_k) \right\|_{\mathcal{V}} \| v_{k+1} - \bar{v} \|_{\mathcal{V}} . \end{aligned} \tag{4.4.40}$$

Plugging this into (4.4.39) yields

$$\begin{aligned} \eta^{-1} Q \geq &- \left(\frac{(L')^2}{4\xi} \| \bar{v} \|_{\mathcal{V}}^2 + \frac{3L'}{2} \| v_{k+1} - \bar{v} \|_{\mathcal{V}} \right) \| u_{k+1} - u_k \|_{\mathcal{U}}^2 \\ &- \left\| \left(\hat{K}_k' (u_k)^* - K' (u_k)^* \right) (2v_{k+1} - v_k) \right\|_{\mathcal{U}} \| u_{k+1} - \bar{u} \|_{\mathcal{U}} \\ &- \left\| \hat{K}_k (u_k) - K (u_k) \right\|_{\mathcal{V}} \| v_{k+1} - \bar{v} \|_{\mathcal{V}} . \end{aligned} \tag{4.4.41}$$

Multiplication by η and application $\mathsf{E}(\cdot \mid \mathcal{F}_k)$ on both sides concludes the proof. □

In the following, we want to use the inequality proven in Lemma 4.4.9 to show that (ECI) holds as well. Since the functions $\hat{K}_k$ and $\hat{K}_k'$ defined in (4.1.5) and (4.1.6) are involved in the former, we need an additional assumption to guarantee that the discrepancy resulting from CGF will somehow decrease.

Assumption 4.4.10 (Reduction of CGF)

There exists a constant $M > 0$ such that the sequence $(A_k)_{k \in \mathbb{N}} \subset \mathcal{R}(\mathcal{F}, \mathcal{A})$ satisfies

$$\mathsf{E}\left(\operatorname*{ess\,sup}_{\xi \in \Xi} \chi_{A_k^{\mathrm{c}}}(\xi) \,\middle|\, \mathcal{F}_k \right) \leq M k^{-3} \quad \text{P-a.s.}$$

for all $k \in \mathbb{N}$, where $A_k^{\mathrm{c}}(\theta)$ denotes the complement of $A_k(\theta) \in \mathcal{A}$ in Ξ for all $\theta \in \Theta$.

In order to better understand the meaning of this assumption, we recall the definition of the essential supremum to see that

$$\operatorname*{ess\,sup}_{\xi \in \Xi} \chi_{A_k^{\mathrm{c}}}(\xi) = \inf_{\substack{N \subset \Xi \\ \mathbb{P}(N) = 0}} \left(\sup_{\xi \in \Xi \setminus N} \chi_{A_k^{\mathrm{c}}}(\xi) \right) = \begin{cases} 0, & \text{if } \mathbb{P}\left(A_k^{\mathrm{c}}\right) = 0, \\ 1, & \text{if } \mathbb{P}\left(A_k^{\mathrm{c}}\right) \neq 0, \end{cases} \quad \text{P-a.s.} \tag{4.4.42}$$

for all $k \in \mathbb{N}$. Thus, the assumption restricts the expectation of the set A_k^{c} having measure > 0 in iteration k, given the information of all preceding iterations. If we would additionally assume that the sequence $(A_k)_{k \in \mathbb{N}}$ is independent, then the σ-algebra generated by $\chi_{A_k^{\mathrm{c}}}$ and $\mathcal{F}_k$ would be independent, too. This follows from the assumption that $\mathfrak{F}$ is a compatible filtration as in

Definition 4.2.1. Thus, we could apply [Kle13, Theorem 8.14 (vi)], yielding

$$\mathsf{E}\left(\operatorname*{ess\,sup}_{\xi\in\Xi} \chi_{A_k^c}(\xi) \,\middle|\, \mathcal{F}_k\right) = \mathsf{E}\left(\operatorname*{ess\,sup}_{\xi\in\Xi} \chi_{A_k^c}(\xi)\right) \quad \mathsf{P}\text{-a.s.} \tag{4.4.43}$$

The improvement here is that we have the expectation instead of the conditional expectation on the right-hand side. Together with (4.4.42) it follows that

$$\mathsf{E}\left(\operatorname*{ess\,sup}_{\xi\in\Xi} \chi_{A_k^c}(\xi) \,\middle|\, \mathcal{F}_k\right) = \mathsf{P}\Big(\left\{\theta\in\Theta \,\middle|\, \mathbb{P}\left(A_k^c(\theta)\right)\neq 0\right\}\Big) = \mathsf{P}\Big(\mathbb{P}\left(A_k^c\right)\neq 0\Big). \tag{4.4.44}$$

Note that $\mathbb{P}$ refers to the probability measure of $\mathcal{V} = L^2(\Xi, \mathcal{A}, \mathbb{P})$, whereas P is the probability measure of $(\Theta, \mathcal{F}, \mathsf{P})$, which stems from the randomization of the algorithm (see Section 4.2). In conclusion, in order to satisfy Assumption 4.4.10, it suffices to assume independence of $(A_k)_{k\in\mathbb{N}}$ as well as

$$\mathsf{P}\Big(\mathbb{P}\left(A_k^c\right)\neq 0\Big) \leq Mk^{-3} \quad \mathsf{P}\text{-a.s.} \tag{4.4.45}$$

for some constant $M > 0$. In a discrete setting as in Chapter 6, the left-hand side would correspond to the probability that at least one component of the gradient is frozen in iteration k. An example of a possible choice of $(A_k)_{k\in\mathbb{N}} \subset \mathcal{A}$ is given in Section 6.2.

With Assumption 4.4.10 we can now prove the following result, which gives us upper bounds on the conditional expectation of the CGF-error.

Lemma 4.4.11 (Upper Bound of the CGF-Error)

For given $(\bar{u}, \bar{v}) = \bar{w} \in \mathcal{W}$, suppose that Assumption 4.4.6 (Desired Ball) is satisfied for some radius $\rho_{\mathcal{U}}(\bar{u}) \geq 0$. Let the sequence $((u_k, v_k))_{k\in\mathbb{N}_0} = (w_k)_{k\in\mathbb{N}_0} \subset \mathcal{R}(\mathcal{F}, \mathcal{W})$ be such that $(u_k, v_k) \in \mathbb{B}_{R_{\mathcal{U}}}(\bar{u}) \cap \mathcal{B}(\bar{u}, \bar{v}) \times \mathbb{B}_{R_{\mathcal{V}}}(\bar{v})$ P-a.s. for all $k \in \mathbb{N}_0$ and some radii $R_{\mathcal{U}}, R_{\mathcal{V}} > 0$. Furthermore, we assume that the sequence $(A_k)_{k\in\mathbb{N}} \subset \mathcal{R}(\mathcal{F}, \mathcal{A})$ satisfies Assumption 4.4.10, and that $\mathfrak{F}$ is a compatible filtration as in Definition 4.2.1. Then the following inequalities hold P-a.s. for all $k \in \mathbb{N}$:

(i) $\mathsf{E}\left(\left\|\hat{K}_k(u_k) - K(u_k)\right\|_{\mathcal{V}} \,\middle|\, \mathcal{F}_k\right) \leq 2LR_{\mathcal{U}}Mk^{-2}$,

(ii) $\mathsf{E}\left(\left\|\left(\hat{K}_k'(u_k)^* - K'(u_k)^*\right)(2v_{k+1} - v_k)\right\|_{\mathcal{U}} \,\middle|\, \mathcal{F}_k\right) \leq 2L'R_{\mathcal{U}}\left(3R_{\mathcal{V}} + \|\bar{v}\|_{\mathcal{V}}\right)Mk^{-2}$.

Proof: All equations are meant to hold P-a.s. First note that

$$\Pi_{A_j^c} \circ \cdots \circ \Pi_{A_k^c} = \Pi_{\cap_{i=j}^k A_i^c} \tag{4.4.46}$$

for all $k \in \mathbb{N}$ and $j \in \{1, \ldots, k\}$. We show by mathematical induction that the following equality holds for all $k \in \mathbb{N}$:

$$\hat{K}_k'(u_k)^* - K'(u_k)^* = \sum_{j=1}^{k}\left(K'(u_{j-1})^* - K'(u_j)^*\right) \circ \Pi_{\cap_{i=j}^k A_i^c}. \tag{4.4.47}$$

- **Base case ($k = 1$):** Using the definition $\hat{K}_0'(\cdot)^* := K'(\cdot)^*$ and (4.1.6) we have

$$\begin{aligned}\hat{K}_1'(u_1)^* - K'(u_1)^* &= K'(u_1)^* \circ \Pi_{A_1} + \hat{K}_0'(u_0)^* \circ \left(\mathrm{Id} - \Pi_{A_1}\right) - K'(u_1)^* \\ &= \left(K'(u_0)^* - K'(u_1)^*\right) \circ \Pi_{A_1^c}.\end{aligned} \tag{4.4.48}$$

- **Inductive step ($k \rightsquigarrow k+1$):** Using (4.1.6) again yields

$$\begin{aligned}\hat{K}_{k+1}'(u_{k+1})^* - K'(u_{k+1})^* &= K'(u_{k+1}) \circ \Pi_{A_{k+1}} + \hat{K}_k'(u_k)^* \circ \Pi_{A_{k+1}^c} - K'(u_{k+1})^* \\ &= \hat{K}_k'(u_k)^* \circ \Pi_{A_{k+1}^c} - K'(u_{k+1}) \circ \Pi_{A_{k+1}^c},\end{aligned} \tag{4.4.49}$$

  which is, by the induction hypothesis for k, equal to

$$\begin{aligned}&\left[\sum_{j=1}^{k}\left(K'(u_{j-1})^* - K'(u_j)^*\right) \circ \Pi_{\cap_{i=j}^k A_i^c} + K'(u_k)^*\right] \circ \Pi_{A_{k+1}^c} - K'(u_{k+1}) \circ \Pi_{A_{k+1}^c} \\ &\qquad = \sum_{j=1}^{k+1}\left(K'(u_{j-1})^* - K'(u_j)^*\right) \circ \Pi_{\cap_{i=j}^{k+1} A_i^c}.\end{aligned} \tag{4.4.50}$$

Analogously to the proof of (4.4.47), we can use the definition of $\hat{K}$ in (4.1.5) to show that

$$\hat{K}_k(u_k) - K(u_k) = \sum_{j=1}^{k} \Pi_{\cap_{i=j}^k A_i^c}\left(K(u_{j-1}) - K(u_j)\right) \tag{4.4.51}$$

for all $k \in \mathbb{N}$. Now, we divide the rest of the proof into two parts, one for each inequality to be proven.

(i) Let $k \in \mathbb{N}$ and $\tilde{v}_j := K(u_{j-1}) - K(u_j)$ for all $j \in \{1, \ldots, k\}$. From (4.4.51) we know that

$$\hat{K}_k(u_k) - K(u_k) = \sum_{j=1}^{k} \Pi_{\cap_{i=j}^k A_i^c}(\tilde{v}_j). \tag{4.4.52}$$

Therefore, we can estimate

$$\begin{aligned}\mathsf{E}\left(\left\|\hat{K}_k(u_k) - K(u_k)\right\|_{\mathcal{V}} \middle| \mathcal{F}_k\right) &\leq \sum_{j=1}^{k} \mathsf{E}\left(\left\|\Pi_{\cap_{i=j}^k A_i^c}(\tilde{v}_j)\right\|_{\mathcal{V}} \middle| \mathcal{F}_k\right) \\ &= \sum_{j=1}^{k} \mathsf{E}\left(\left\|\prod_{i=j}^{k} \chi_{A_i^c}(\cdot)\tilde{v}_j(\cdot)\right\|_{\mathcal{V}} \middle| \mathcal{F}_k\right),\end{aligned} \tag{4.4.53}$$

where we have used equation (4.4.46) as well as the definition of Π_A in (4.1.3). The definition of the norm of $\mathcal{V} = L^2(\Xi, \mathcal{A}, \mathbb{P})$ and the fact that $\chi_{A_i^c}^2 \equiv \chi_{A_i^c}$ imply

$$\left\|\prod_{i=j}^{k} \chi_{A_i^c}(\cdot)\tilde{v}_j(\cdot)\right\|_{\mathcal{V}} = \left(\int_{\Xi} \prod_{i=j}^{k} \chi_{A_i^c}(\xi)\tilde{v}_j(\xi)^2 \mathrm{dP}(\xi)\right)^{1/2} \tag{4.4.54}$$

for all $j \in \{1, \dots, k\}$. Hölder's inequality (with $p = 1$ and $q = \infty$) yields

$$\begin{aligned} \int_{\Xi} \prod_{i=j}^{k} \chi_{A_i^c}(\xi) \tilde{v}_j(\xi)^2 \mathrm{dP}(\xi) &\leq \left(\int_{\Xi} \prod_{i=j}^{k-1} \chi_{A_i^c}(\xi) \tilde{v}_j(\xi)^2 \mathrm{dP}(\xi) \right) \left(\operatorname*{ess\,sup}_{\xi \in \Xi} \chi_{A_k^c}(\xi) \right) \\ &\leq \left\| \tilde{v}_j \right\|_{\mathcal{V}}^2 \operatorname*{ess\,sup}_{\xi \in \Xi} \chi_{A_k^c}(\xi), \end{aligned} \tag{4.4.55}$$

where we have used in the last inequality that $\chi_{A_i^c}(\xi) \leq 1$ for all $i \in \{j, \dots, k-1\}$ and all $\xi \in \Xi$. Due to the local Lipschitz continuity of K (see Definition 4.4.1) and the fact that $u_j \in \mathbb{B}_{R_{\mathcal{U}}}(\bar{u})$ P-a.s. for all $j \in \mathbb{N}_0$, we can estimate

$$\left\| \tilde{v}_j \right\|_{\mathcal{V}} = \left\| K\left(u_{j-1}\right) - K\left(u_j\right) \right\|_{\mathcal{V}} \leq L \left\| u_{j-1} - u_j \right\|_{\mathcal{U}} \leq 2LR_{\mathcal{U}} \tag{4.4.56}$$

for all $j \in \{1, \dots, k\}$. If we combine (4.4.53), (4.4.54), (4.4.55), and (4.4.56), we get

$$\begin{aligned} \mathsf{E}\left(\left\| \hat{K}_k\left(u_k\right) - K\left(u_k\right) \right\|_{\mathcal{V}} \middle| \mathcal{F}_k \right) &\leq 2LR_{\mathcal{U}} k \mathsf{E}\left(\operatorname*{ess\,sup}_{\xi \in \Xi} \chi_{A_k^c}(\xi) \middle| \mathcal{F}_k \right) \\ &\leq 2LR_{\mathcal{U}} M k^{-2}, \end{aligned} \tag{4.4.57}$$

where we have used Assumption 4.4.10 for the last inequality.

(ii) The proof of the second inequality is similar to the one above. Let $k \in \mathbb{N}$ and $\tilde{v} := 2v_{k+1} - v_k$. From (4.4.47) we know that

$$\hat{K}_k'\left(u_k\right)^* - K'\left(u_k\right)^* = \sum_{j=1}^{k} \left(K'\left(u_{j-1}\right)^* - K'\left(u_j\right)^* \right) \circ \Pi_{\cap_{i=j}^{k} A_i^c}. \tag{4.4.58}$$

Therefore, we have

$$\begin{aligned} &\mathsf{E}\left(\left\| \left(\hat{K}_k'\left(u_k\right)^* - K'\left(u_k\right)^* \right)(\tilde{v}) \right\|_{\mathcal{U}} \middle| \mathcal{F}_k \right) \\ &\quad \leq \sum_{j=1}^{k} \mathsf{E}\left(\left\| K'\left(u_{j-1}\right)^* - K'\left(u_j\right)^* \right\|_{\mathbb{L}(\mathcal{V},\mathcal{U})} \left\| \Pi_{\cap_{i=j}^{k} A_i^c}(\tilde{v}) \right\|_{\mathcal{V}} \middle| \mathcal{F}_k \right) \\ &\quad \leq \sum_{j=1}^{k} \left\| K'\left(u_{j-1}\right)^* - K'\left(u_j\right)^* \right\|_{\mathbb{L}(\mathcal{V},\mathcal{U})} \mathsf{E}\left(\left\| \Pi_{\cap_{i=j}^{k} A_i^c}(\tilde{v}) \right\|_{\mathcal{V}} \middle| \mathcal{F}_k \right), \end{aligned} \tag{4.4.59}$$

where we have used the $\mathcal{F}_k$-measurability of $\left\| K'\left(u_{j-1}\right)^* - K'\left(u_j\right)^* \right\|_{\mathbb{L}(\mathcal{V},\mathcal{U})} : \Theta \to \mathbb{R}$ for all $j \in \{1, \dots, k\}$ together with [Kle13, Theorem 8.14 (iii)]. Due to the local Lipschitz continuity of K' (see Assumption 4.4.2) and the fact that $u_j \in \mathbb{B}_{R_{\mathcal{U}}}(\bar{u})$ P-a.s. for all $j \in \mathbb{N}_0$, we can estimate

$$\left\| K'\left(u_{j-1}\right)^* - K'\left(u_j\right)^* \right\|_{\mathbb{L}(\mathcal{V},\mathcal{U})} = \left\| K'\left(u_{j-1}\right) - K'\left(u_j\right) \right\|_{\mathbb{L}(\mathcal{U},\mathcal{V})} \leq 2L' R_{\mathcal{U}} \tag{4.4.60}$$

for all $j \in \{1, \dots, k\}$. For the term of (4.4.59) involving the conditional expectation, we use Hölder's inequality analogously to (4.4.55) to get

$$\left\| \Pi_{\cap_{i=j}^{k} A_i^c}(\tilde{v}) \right\|_{\mathcal{V}} \leq \| \tilde{v} \|_{\mathcal{V}} \operatorname*{ess\,sup}_{\xi \in \Xi} \chi_{A_k^c}(\xi). \tag{4.4.61}$$

Since $v_j \in \mathbb{B}_{R_\mathcal{V}}(\bar{v})$ P-a.s. for all $j \in \mathbb{N}_0$, we know that

$$\begin{aligned} \| \tilde{v} \|_\mathcal{V} = \| 2v_{k+1} - v_k \|_\mathcal{V} &\leq 2 \| v_{k+1} - \bar{v} \|_\mathcal{V} + \| v_k - \bar{v} \|_\mathcal{V} + \| \bar{v} \|_\mathcal{V} \\ &\leq 3R_\mathcal{V} + \| \bar{v} \|_\mathcal{V} . \end{aligned} \tag{4.4.62}$$

If we combine this with (4.4.59), (4.4.60), and (4.4.61), and apply Assumption 4.4.10, we get

$$\begin{aligned} &\mathsf{E}\left(\left\| \left(\hat{K}_k'(u_k)^* - K'(u_k)^* \right)(2v_{k+1} - v_k) \right\|_\mathcal{U} \middle| \mathcal{F}_k \right) \\ &\qquad \leq \sum_{j=1}^{k} 2L' R_\mathcal{U} \left(3R_\mathcal{V} + \| \bar{v} \|_\mathcal{V} \right) \mathsf{E}\left(\operatorname*{ess\,sup}_{\xi \in \Xi} \chi_{A_k^\xi}(\xi) \middle| \mathcal{F}_k \right) \\ &\qquad \leq 2L' R_\mathcal{U} \left(3R_\mathcal{V} + \| \bar{v} \|_\mathcal{V} \right) M k^{-2}, \end{aligned} \tag{4.4.63}$$

which concludes the proof. □

Now we can prove under which additional conditions the inequality (ECI) is satisfied.

Proposition 4.4.12 (Satisfaction of Central Inequality)

For given $(\bar{u}, \bar{v}) = \bar{w} \in \mathcal{W}$, suppose that Assumption 4.4.6 (Desired Ball) is satisfied for some radius $\rho_\mathcal{U}(\bar{u}) \geq 0$. Let the sequence $((u_k, v_k))_{k \in \mathbb{N}_0} = (w_k)_{k \in \mathbb{N}_0} \subset \mathcal{R}(\mathcal{F}, \mathcal{W})$ be such that $(u_k, v_k) \in \mathbb{B}_{R_\mathcal{U}}(\bar{u}) \cap \mathcal{B}(\bar{u}, \bar{v}) \times \mathbb{B}_{R_\mathcal{V}}(\bar{v})$ P-a.s. for all $k \in \mathbb{N}_0$ and some radii $R_\mathcal{U}, R_\mathcal{V} > 0$. Furthermore, we assume that the sequence $(A_k)_{k \in \mathbb{N}} \subset \mathcal{R}(\mathcal{F}, \mathcal{A})$ satisfies Assumption 4.4.10, $\mathfrak{F}$ is a compatible filtration as in Definition 4.2.1, and Assumption 4.4.7 (Step-Size-Testing-Relation) holds. If $\| v_{k+1} - \bar{v} \|_\mathcal{V} < \gamma_\mathcal{U} \frac{2}{3L'}$ P-a.s. for all $k \in \mathbb{N}_0$ and there exists $\hat{\delta} \in (0,1)$ such that

$$\varphi_k \geq \frac{2\eta}{\left(1 - \hat{\delta}\right)\delta} \left(\frac{(L')^2}{4\xi_{k+1}} \| \bar{v} \|_\mathcal{V}^2 + \gamma_\mathcal{U} - \xi_{k+1} \right) \quad \text{P-a.s.}$$

for some $\delta \in (0, \kappa]$, $\xi_{k+1} \in \mathcal{R}\left(\mathcal{F}_{k+1}, \left(0, \gamma_\mathcal{U} - \frac{3L'}{2} \| v_{k+1} - \bar{v} \|_\mathcal{V}\right]\right)$ and all $k \in \mathbb{N}_0$, then there exists $(\Delta_k)_{k \in \mathbb{N}_0} \in l_+(\mathfrak{F})$ with

$$\Delta_k \geq \frac{\hat{\delta}}{2} \mathsf{E}\left(\| w_{k+1} - w_k \|_{Z_k M_k}^2 \middle| \mathcal{F}_k \right) \quad \text{P-a.s.}$$

for all $k \in \mathbb{N}_0$ such that (ECI) is satisfied for $z = \bar{w}$ and

$$\begin{aligned} \lambda_k := \eta \mathsf{E}&\left(\left\| \left(\hat{K}_k'(u_k)^* - K'(u_k)^* \right)(2v_{k+1} - v_k) \right\|_\mathcal{U} \| u_{k+1} - \bar{u} \|_\mathcal{U} \middle| \mathcal{F}_k \right) \\ &+ \eta \mathsf{E}\left(\left\| \hat{K}_k(u_k) - K(u_k) \right\|_\mathcal{V} \| v_{k+1} - \bar{v} \|_\mathcal{V} \middle| \mathcal{F}_k \right). \end{aligned}$$

Moreover, we have $(\lambda_k)_{k \in \mathbb{N}_0} \in l_+^1(\mathfrak{F})$, i.e. $\sum_{k \in \mathbb{N}_0} \lambda_k < \infty$ P-a.s.

Proof: First, we want to show that (ECI) is satisfied for all $k \in \mathbb{N}_0$ with $z = \overline{w}$, λ_k as defined above, and some sequence $(\Delta_k)_{k\in\mathbb{N}_0} \in l_+(\mathfrak{F})$, i.e.

$$\begin{aligned}&\mathsf{E}\left(\left\langle \tilde{H}_k(w_{k+1}), w_{k+1}-z\right\rangle_{Z_k} \middle| \mathcal{F}_k\right) + \lambda_k \\ &\qquad \geq \frac{1}{2}\mathsf{E}\left(\| w_{k+1}-z \|^2_{Z_{k+1}M_{k+1}-Z_kM_k} \middle| \mathcal{F}_k\right) - \frac{1}{2}\mathsf{E}\left(\| w_{k+1}-w_k \|^2_{Z_kM_k} \middle| \mathcal{F}_k\right) + \Delta_k \quad \text{(ECI)}\end{aligned}$$

holds P-a.s. Let now $k \in \mathbb{N}_0$ be fixed. If we plug $z = \overline{w}$ into (ECI) and rearrange such that the left-hand side is identical to the one of Lemma 4.4.9, we get

$$\begin{aligned}&\mathsf{E}\left(\left\langle \tilde{H}_k(w_{k+1}), w_{k+1}-\overline{w}\right\rangle_{Z_k} \middle| \mathcal{F}_k\right) - \frac{1}{2}\mathsf{E}\left(\| w_{k+1}-\overline{w} \|^2_{Z_{k+1}M_{k+1}-Z_kM_k} \middle| \mathcal{F}_k\right) \\ &\qquad \geq -\frac{1}{2}\mathsf{E}\left(\| w_{k+1}-w_k \|^2_{Z_kM_k} \middle| \mathcal{F}_k\right) + \Delta_k - \lambda_k \quad \text{P-a.s.} \quad (4.4.64)\end{aligned}$$

Therefore, if we apply Lemma 4.4.9, we only need to show that we can find a $\Delta_k \in \mathcal{R}(\mathcal{F}_k, [0,\infty))$ satisfying

$$\begin{aligned}&-\eta\mathsf{E}\left(\left(\frac{(L')^2}{4\xi_{k+1}}\| \bar{v} \|^2_{\mathcal{V}} + \frac{3L'}{2}\| v_{k+1}-\bar{v} \|_{\mathcal{V}}\right)\| u_{k+1}-u_k \|^2_{\mathcal{U}} \middle| \mathcal{F}_k\right) \\ &\qquad \geq -\frac{1}{2}\mathsf{E}\left(\| w_{k+1}-w_k \|^2_{Z_kM_k} \middle| \mathcal{F}_k\right) + \Delta_k \quad \text{P-a.s.} \quad (4.4.65)\end{aligned}$$

as well as

$$\Delta_k \geq \frac{\hat{\delta}}{2}\mathsf{E}\left(\| w_{k+1}-w_k \|^2_{Z_kM_k} \middle| \mathcal{F}_k\right) \quad \text{P-a.s.} \quad (4.4.66)$$

This is true, if

$$\begin{aligned}&-\eta\rho_{k+1}\mathsf{E}\left(\| u_{k+1}-u_k \|^2_{\mathcal{U}} \middle| \mathcal{F}_k\right) + \frac{1}{2}\mathsf{E}\left(\| w_{k+1}-w_k \|^2_{Z_kM_k} \middle| \mathcal{F}_k\right) \\ &\qquad \geq \frac{\hat{\delta}}{2}\mathsf{E}\left(\| w_{k+1}-w_k \|^2_{Z_kM_k} \middle| \mathcal{F}_k\right) \quad (4.4.67)\end{aligned}$$

holds P-a.s., where $\rho_{k+1} := \frac{(L')^2}{4\xi_{k+1}}\| \bar{v} \|^2_{\mathcal{V}} + \frac{3L'}{2}\| v_{k+1}-\bar{v} \|_{\mathcal{V}}$. We know from Lemma 4.4.8 that

$$\| w_{k+1}-w_k \|^2_{Z_kM_k} \geq \delta\varphi_k \| u_{k+1}-u_k \|^2_{\mathcal{U}} + \frac{\kappa-\delta}{1-\delta}\psi \| v_{k+1}-v_k \|^2_{\mathcal{V}} \quad \text{P-a.s.} \quad (4.4.68)$$

for any $\delta \in (0,\kappa]$, where κ stems from Assumption 4.4.7. Therefore, in order to show (4.4.67), it suffices to prove that

$$\begin{aligned}&\frac{1-\hat{\delta}}{2}\delta\varphi_k\mathsf{E}\left(\| u_{k+1}-u_k \|^2_{\mathcal{U}} \middle| \mathcal{F}_k\right) + \frac{\left(1-\hat{\delta}\right)(\kappa-\delta)}{2(1-\delta)}\psi\mathsf{E}\left(\| v_{k+1}-v_k \|^2_{\mathcal{V}} \middle| \mathcal{F}_k\right) \\ &\qquad \geq \eta\rho_{k+1}\mathsf{E}\left(\| u_{k+1}-u_k \|^2_{\mathcal{U}} \middle| \mathcal{F}_k\right) \quad \text{P-a.s.} \quad (4.4.69)\end{aligned}$$

Since the second term of the sum on the left-hand side is non-negative and the conditional

expectation is monotone [Kle13, Theorem 8.14 (ii)], this is true if

$$\frac{1-\hat{\delta}}{2}\delta\varphi_k \geq \eta\rho_{k+1} = \eta\frac{(L')^2}{4\xi_{k+1}} \| \bar{v} \|_{\mathcal{V}}^2 + \eta\frac{3L'}{2} \| v_{k+1} - \bar{v} \|_{\mathcal{V}} \quad \text{P-a.s.} \tag{4.4.70}$$

Now, the assumption $\xi_{k+1} \leq \gamma_{\mathcal{U}} - \frac{3L'}{2} \| v_{k+1} - \bar{v} \|_{\mathcal{V}}$ P-a.s. implies that it suffices to show

$$\frac{1-\hat{\delta}}{2}\delta\varphi_k \geq \eta\left(\frac{(L')^2}{4\xi_{k+1}} \| \bar{v} \|_{\mathcal{V}}^2 + \gamma_{\mathcal{U}} - \xi_{k+1}\right) \quad \text{P-a.s.,} \tag{4.4.71}$$

which is equivalent to the assumed inequality

$$\varphi_k \geq \frac{2\eta}{\left(1-\hat{\delta}\right)\delta}\left(\frac{(L')^2}{4\xi_{k+1}} \| \bar{v} \|_{\mathcal{V}}^2 + \gamma_{\mathcal{U}} - \xi_{k+1}\right) \quad \text{P-a.s.} \tag{4.4.72}$$

It remains to prove that $(\lambda_k)_{k\in\mathbb{N}_0} \in l_+^1(\mathfrak{F})$, which is also a condition of (ECI). Recall that

$$\begin{aligned}\lambda_k := \eta\mathsf{E}\left(\left\| \left(\hat{K}_k'(u_k)^* - K'(u_k)^*\right)(2v_{k+1} - v_k)\right\|_{\mathcal{U}} \| u_{k+1} - \bar{u} \|_{\mathcal{U}} \,\middle|\, \mathcal{F}_k\right) \\ + \eta\mathsf{E}\left(\left\| \hat{K}_k(u_k) - K(u_k)\right\|_{\mathcal{V}} \| v_{k+1} - \bar{v} \|_{\mathcal{V}} \,\middle|\, \mathcal{F}_k\right)\end{aligned} \tag{4.4.73}$$

for all $k \in \mathbb{N}_0$. Since $(u_k, v_k) \in \mathbb{B}_{R_{\mathcal{U}}}(\bar{u}) \times \mathbb{B}_{R_{\mathcal{V}}}(\bar{v})$ P-a.s. for all $k \in \mathbb{N}_0$, we have

$$\begin{aligned}\lambda_k \leq \eta R_{\mathcal{U}}\mathsf{E}\left(\left\| \left(\hat{K}_k'(u_k)^* - K'(u_k)^*\right)(2v_{k+1} - v_k)\right\|_{\mathcal{U}} \,\middle|\, \mathcal{F}_k\right) \\ + \eta R_{\mathcal{V}}\mathsf{E}\left(\left\| \hat{K}_k(u_k) - K(u_k)\right\|_{\mathcal{V}} \,\middle|\, \mathcal{F}_k\right)\end{aligned} \tag{4.4.74}$$

P-a.s. Therefore, in order to show that $(\lambda_k)_{k\in\mathbb{N}_0} \in l_+^1(\mathfrak{F})$, it suffices to show that

$$\begin{aligned}\sum_{k=1}^{\infty} R_{\mathcal{U}}\mathsf{E}\left(\left\| \left(\hat{K}_k'(u_k)^* - K'(u_k)^*\right)(2v_{k+1} - v_k)\right\|_{\mathcal{U}} \,\middle|\, \mathcal{F}_k\right) \\ + \sum_{k=1}^{\infty} R_{\mathcal{V}}\mathsf{E}\left(\left\| \hat{K}_k(u_k) - K(u_k)\right\|_{\mathcal{V}} \,\middle|\, \mathcal{F}_k\right) < \infty\end{aligned} \tag{4.4.75}$$

P-a.s., where $k = 0$ is intentionally disregarded because the corresponding term of the sum is zero anyway. According to Lemma 4.4.11, we know that

$$\begin{aligned}\sum_{k=1}^{\infty} R_{\mathcal{U}}\mathsf{E}\left(\left\| \left(\hat{K}_k'(u_k)^* - K'(u_k)^*\right)(2v_{k+1} - v_k)\right\|_{\mathcal{U}} \,\middle|\, \mathcal{F}_k\right) \\ + \sum_{k=1}^{\infty} R_{\mathcal{V}}\mathsf{E}\left(\left\| \hat{K}_k(u_k) - K(u_k)\right\|_{\mathcal{V}} \,\middle|\, \mathcal{F}_k\right) \\ \leq 2R_{\mathcal{U}}M\left(L'R_{\mathcal{U}}\left(3R_{\mathcal{V}} + \| \bar{v} \|_{\mathcal{V}}\right) + LR_{\mathcal{V}}\right)\sum_{k=1}^{\infty} k^{-2} < \infty\end{aligned} \tag{4.4.76}$$

P-a.s., since $\sum_{k=1}^{\infty} k^{-2} = \frac{\pi^2}{6}$ [Dan12]. □

4.4.3 Convergence

In this subsection we combine all results of the previous subsections of Section 4.4 to formulate a theorem proving the weak convergence of the iterates of Algorithm 4.1 to a critical point. We will use Theorem 4.3.6, which requires (ECI) to hold for all z in some subset $\mathcal{Z} \subset H^{-1}(0)$. Therefore, in order to being able to use Proposition 4.4.12 to show that (ECI) is satisfied, we need to make the following assumption.

Assumption 4.4.13 (Existence of $\mathcal{Z}$)

Let the sequence $((u_k, v_k))_{k\in\mathbb{N}_0} \subset \mathcal{R}(\mathcal{F}, \mathcal{U}\times\mathcal{V})$ be generated by Algorithm 4.1. Let further $(\bar{u}, \bar{v}) \in H^{-1}(0)$ be fixed. It follows from Lemma 4.1.1 that there exist radii $R_{\mathcal{U}}, R_{\mathcal{V}} > 0$ such that $(u_k, v_k) \in \mathbb{B}_{R_{\mathcal{U}}}(\bar{u}) \times \mathbb{B}_{R_{\mathcal{V}}}(\bar{v})$ P-a.s. for all $k \in \mathbb{N}$. We assume that there exist a non-empty, bounded subset $\mathcal{Z} \subset H^{-1}(0)$ and constants[2] $L, L', \gamma_G \geq 0$, and $\gamma_{\mathcal{U}} > 0$ such that

(i) for all $(u, v) \in \mathcal{Z}$ there exists a radius $\rho_{\mathcal{U}}(u) \in [0, \infty]$ such that Assumption 4.4.6 (Desired Ball) is satisfied for the given constants $L, L', \gamma_G \geq 0$ and $\gamma_{\mathcal{U}} > 0$;

(ii) $\mathcal{Z} \subset \mathcal{B} \times \mathcal{V}$ where $\mathcal{B} := \bigcap_{(u,v)\in\mathcal{Z}} \mathcal{B}(u, v)$ and $\mathcal{B}(u, v) = \mathbb{B}_{\rho_{\mathcal{U}}(u)}(u)$ is the desired ball from Assumption 4.4.6;

(iii) $\mathcal{Z} \supset H^{-1}(0) \cap \mathcal{D}$ where $\mathcal{D} := \mathbb{B}_{R_{\mathcal{U}}}(\bar{u}) \times \mathbb{B}_{R_{\mathcal{V}}}(\bar{v})$.

Basically, Assumption 4.4.13 (i) and (ii) guarantee that every element of $\mathcal{Z}$ lies in a ball where the assumptions from Section 4.4.1 hold for the same constants $L, L', \gamma_G \geq 0$, and $\gamma_{\mathcal{U}} > 0$. Part (iii) of Assumption 4.4.13 makes sure that $\mathcal{Z}$ contains at least all the critical points which can possibly be reached by the algorithm. Note that, in the special case where the assumptions from Section 4.4.1 hold globally, the desired balls may have infinite radius. Therefore, Assumption 4.4.13 (i) and (ii) are satisfied for all $\mathcal{Z} \subset H^{-1}(0)$, and a simple choice satisfying Assumption 4.4.13 (iii) could be $\mathcal{Z} := H^{-1}(0)$.

Now we can extend Proposition 4.4.12 to show that (ECI) is satisfied not only at one arbitrary point in $\mathcal{W}$, but at every point of a subset $\mathcal{Z} \subset H^{-1}(0)$. This is what we need to formulate the following theorem, which is based on [CMV19, Theorem 4.1]. It shows P-a.s. weak convergence of the sequence generated by Algorithm 4.1 to a saddle point in $\mathcal{Z}$.

Theorem 4.4.14 (Weak Convergence of the Algorithm)

Let the sequence $(w_k)_{k\in\mathbb{N}_0} = ((u_k, v_k))_{k\in\mathbb{N}_0} \subset \mathcal{R}(\mathcal{F}, \mathcal{W})$ be generated by Algorithm 4.1 and assume that the following conditions are satisfied:

(i) $\mathcal{Z} \subset H^{-1}(0)$ satisfies Assumption 4.4.13 with constants $L, L', \gamma_G \geq 0$, $\gamma_{\mathcal{U}} > 0$, radii $R_{\mathcal{U}}, R_{\mathcal{V}} > 0$, and the sets $\mathcal{B}$ and $\mathcal{D}$ as defined in Assumption 4.4.13 (ii) and (iii), respectively;

(ii) $u_k \in \mathcal{B}$ P-a.s. for all $k \in \mathbb{N}_0$;

(iii) Assumption 4.4.7 (Step-Size-Testing-Relation) holds for some $\tilde{\gamma}_G \in [0, \gamma_G]$, $\kappa \in (0, 1)$, and $\eta > 0$;

[2] These constants stem from Definition 4.4.1, Assumption 4.4.2, Assumption 4.4.3, and Assumption 4.4.4, which are inherent in Assumption 4.4.6.

(iv) the sequence $(A_k)_{k\in\mathbb{N}} \subset \mathcal{R}(\mathcal{F}, \mathcal{A})$ satisfies Assumption 4.4.10 (Reduction of CGF) with a constant $M > 0$;

(v) $\mathfrak{F}$ is a compatible filtration as in Definition 4.2.1;

(vi) $\| v_{k+1} - \bar{v} \|_{\mathcal{V}} < \gamma_{\mathcal{U}} \frac{2}{3L'}$ P-a.s. for all $k \in \mathbb{N}_0$ and all $(\bar{u}, \bar{v}) \in \mathcal{Z}$;

(vii) there exist $\delta \in (0, \kappa)$, $\hat{\delta} \in (0, 1)$, and $\xi_{k+1} \in \mathcal{R}(\mathcal{F}_{k+1}, (0, \infty))$ such that, for all $k \in \mathbb{N}_0$ and $(\bar{u}, \bar{v}) \in \mathcal{Z}$, we have

$$\xi_{k+1} \in \left(0, \gamma_{\mathcal{U}} - \frac{3L'}{2} \| v_{k+1} - \bar{v} \|_{\mathcal{V}}\right] \quad \text{P-a.s.}$$

and

$$\varphi_k \geq \frac{2\eta}{\left(1 - \hat{\delta}\right)\delta} \left(\frac{(L')^2}{4\xi_{k+1}} \| \bar{v} \|_{\mathcal{V}}^2 + \gamma_{\mathcal{U}} - \xi_{k+1} \right) \quad \text{P-a.s.}$$

(viii) the primal step sizes are bounded away from 0, i.e. there exists $\underline{\tau} \in (0, \infty)$ such that $\tau_k \geq \underline{\tau}$ for all $k \in \mathbb{N}_0$;

(ix) the mapping

$$(u, v) \mapsto \begin{pmatrix} -K'(u)^* v \\ K(u) \end{pmatrix}$$

is P-a.s. weak-to-strong continuous in $\mathcal{D}$, i.e. its graph is P-a.s. sequentially closed in the Cartesian product of the weak and the strong topological space of $\mathcal{W}$ in the sense of [BC17, Section 2.4];

(x) there exists a linear operator $\overline{K'} \in \mathbb{L}(\mathcal{U}, \mathcal{V})$ such that, for every P-a.s. weakly convergent subsequence $(w_{n_k})_{k\in\mathbb{N}_0}$, we have $K'(u_{n_k}) \to \overline{K'}$ P-a.s. as $k \to \infty$.

Then $w_k \rightharpoonup \hat{w}$ P-a.s. as $k \to \infty$ for some $\hat{w} \in \mathcal{R}(\mathcal{F}, \mathcal{Z})$.

Proof: We want to use Theorem 4.3.6 and therefore need to satisfy its assumptions:

(i) We know from condition (iii) that Assumption 4.4.7 holds, and according to Assumption 4.4.7 (ii), $(\varphi_k)_{k\in\mathbb{N}_0}$ is monotonically increasing. Thus, if we apply Lemma 4.4.8 with $\delta \in (0, \kappa)$, we conclude that $Z_k M_k$ is P-a.s. self-adjoint and there exists $\varepsilon \neq 0$ such that $Z_k M_k \geq \varepsilon^2 \mathrm{Id}_{\mathcal{W}}$ P-a.s. for all $k \in \mathbb{N}_0$. Therefore, Assumption 4.3.1 is satisfied as well.

(ii) Assumption 4.4.7 (i) and condition (viii) imply that the sequence $(\varphi_k)_{k\in\mathbb{N}_0}$ is bounded from above. Moreover, we know from condition (i) that $(u_k, v_k) \in \mathcal{D}$ for all $k \in \mathbb{N}$. Therefore, since $K' \colon \mathcal{U} \to \mathbb{L}(\mathcal{U}, \mathcal{V})$ is continuous, we also know that there exists a $C' \in (0, \infty)$ such that $\| K'(u_k) \|_{\mathbb{L}(\mathcal{U},\mathcal{V})} \leq C'$ P-a.s. for all $k \in \mathbb{N}$. Altogether, considering the definition of $Z_k M_k$ in (4.4.12), we can easily see that there exists a $C \in \mathbb{R}$ such that $\| Z_k M_k \|_{\mathbb{L}(\mathcal{W},\mathcal{W})} \leq C^2$ P-a.s. for all $k \in \mathbb{N}$.

(iii) In order to prove assumption (iii) of Theorem 4.3.6, we apply Proposition 4.4.12 to every element $(\bar{u}, \bar{v}) = z \in \mathcal{Z}$. Its assumptions are satisfied by conditions (i)-(vii) of

this theorem. Therefore, for every $z \in \mathcal{Z}$, we have sequences $(\Delta_k(z))_{k\in\mathbb{N}_0} \in l_+(\mathfrak{F})$ and $(\lambda_k(z))_{k\in\mathbb{N}_0} \in l_+^1(\mathfrak{F})$ with

$$\Delta_k(z) \geq \frac{\hat{\delta}}{2}\mathsf{E}\left(\| w_{k+1} - w_k \|_{Z_k M_k}^2 \,\middle|\, \mathcal{F}_k\right) \quad \text{P-a.s.}$$

for all $k \in \mathbb{N}_0$ such that (ECI) is satisfied.

(iv) For the proof of Theorem 4.3.6 (iv), we assume that $Z_k M_k (w_{k+1} - w_k) \to 0$ P-a.s. as $k \to \infty$. We now need to show that every weak accumulation point of $(w_k)_{k\in\mathbb{N}_0}$ is P-a.s. in $\mathcal{Z}$. Therefore, let $(\bar{u}, \bar{v}) = \bar{w} \in \mathcal{W}$ be a weak accumulation point such that $(u_{n_k+1}, v_{n_k+1}) = w_{n_k+1} \rightharpoonup \bar{w}$ P-a.s. as $k \to \infty$. Since the sequence $(w_k)_{k\in\mathbb{N}_0}$ was generated by Algorithm 4.1, we know that (RI) holds, i.e.

$$0 \in W_{n_k} H\left(w_{n_k+1}\right) + D_{n_k}\left(w_{n_k+1}\right) + M_{n_k}\left(w_{n_k+1} - w_{n_k}\right) \quad \text{P-a.s.} \tag{4.4.77}$$

for all $k \in \mathbb{N}_0$, where we have used the definition of $\tilde{H}$ in (4.2.2). Now let

$$S(w) := \begin{pmatrix} \partial G(u) \\ \partial \mathrm{CVaR}_\beta^*(v) \end{pmatrix} \tag{4.4.78}$$

for all $(u, v) = w \in \mathcal{W}$ and

$$x_{n_k+1} := W_{n_k}\begin{pmatrix} -K'\left(u_{n_k+1}\right)^* v_{n_k+1} \\ K\left(u_{n_k+1}\right) \end{pmatrix} - D_{n_k}\left(w_{n_k+1}\right) - M_{n_k}\left(w_{n_k+1} - w_{n_k}\right). \tag{4.4.79}$$

The inclusion (4.4.77) is then equivalent to

$$x_{n_k+1} \in W_{n_k} S\left(w_{n_k+1}\right) \quad \text{P-a.s.} \tag{4.4.80}$$

We have already seen that $(\varphi_k)_{k\in\mathbb{N}_0}$ is monotonically increasing. Consequently, the sequence of primal step sizes $(\tau_k)_{k\in\mathbb{N}_0}$ is monotonically decreasing, and since it is bounded from below according to condition (viii), it must converge to some $\bar{\tau} \in \left[\underline{\tau}, \tau_0\right]$. Therefore, there exists a regular linear operator $\overline{W} \in \mathbb{L}(\mathcal{W}, \mathcal{W})$ such that $W_k \to \overline{W}$ as $k \to \infty$. We now define

$$\bar{x} := \overline{W}\begin{pmatrix} -K'(\bar{u})^* \bar{v} \\ K(\bar{u}) \end{pmatrix}. \tag{4.4.81}$$

If we recall the definition of H in (4.1.15), we can easily see that $\bar{w} \in H^{-1}(0)$ is equivalent to

$$\bar{x} \in \overline{W} S(\bar{w}). \tag{4.4.82}$$

In order to show the latter inclusion, we note that G and CVaR_β^* are proper, convex, and lower semi-continuous. Thus, according to [BC17, Proposition 16.36], the graph of S is sequentially weakly-strongly closed, i.e. if $w_{n_k+1} \rightharpoonup \bar{w}$ P-a.s. and $S\left(w_{n_k+1}\right) \ni W_{n_k}^{-1} x_{n_k+1} \to \overline{W}^{-1}\bar{x}$ P-a.s. as $k \to \infty$ imply that $\bar{x} \in \overline{W} S(\bar{w})$ P-a.s. Therefore, all we need to show is that $x_{n_k+1} \to \bar{x}$ P-a.s. as $k \to \infty$ by examining the terms of the sum of (4.4.79) separately.

Due to condition (iii) and the convergence of the primal step size sequence $(\tau_k)_{k\in\mathbb{N}_0}$, we know that the sequences $(\varphi_k)_{k\in\mathbb{N}_0}$ and hence $(Z_k)_{k\in\mathbb{N}_0}$ converge (strongly) to some $\overline{Z} \in \mathbb{L}(\mathcal{W},\mathcal{W})$ with $\overline{Z} \neq 0$ as well. Therefore, the convergence assumption $Z_k M_k (w_{k+1} - w_k) \to 0$ P-a.s. as $k \to \infty$ implies that the third term of the sum in (4.4.79) converges, i.e.

$$M_{n_k}\left(w_{n_k+1} - w_{n_k}\right) \to 0 \quad \text{P-a.s.} \tag{4.4.83}$$

as $k \to \infty$. Furthermore, since $\delta < \kappa$, it follows from Lemma 4.4.8 that there exists $\varepsilon > 0$ such that

$$Z_k M_k = \begin{pmatrix} \varphi_k \mathrm{Id}_{\mathcal{U}} & \eta K'(u_k)^* \\ \eta K'(u_k) & \psi \mathrm{Id}_{\mathcal{V}} \end{pmatrix} \geq \varepsilon \mathrm{Id} \quad \text{P-a.s.} \tag{4.4.84}$$

for all $k \in \mathbb{N}_0$. Again, due to the assumption that $Z_k M_k (w_{k+1} - w_k) \to 0$ P-a.s. as $k \to \infty$, it follows that

$$\varepsilon \left\| w_{n_k+1} - w_{n_k} \right\|_{\mathcal{W}}^2 \leq \left\| w_{n_k+1} - w_{n_k} \right\|_{Z_{n_k} M_{n_k}}^2 \to 0 \quad \text{P-a.s.}, \tag{4.4.85}$$

which we need in the following.

Now we consider the term $D_{n_k}(w_{n_k+1}) = D_{n_k}^{LIN}(w_{n_k+1}) + D_{n_k}^{CGF}(w_{n_k+1})$. From (4.1.16) and Assumption 4.4.7 we know that

$$D_{n_k}^{LIN}(w_{n_k+1}) = \begin{pmatrix} \tau_{n_k}\left[K'(u_{n_k})^* - K'(u_{n_k+1})^*\right] v_{n_k+1} \\ \sigma\left[K(u_{n_k+1}) - K(u_{n_k}) - K'(u_{n_k})(u_{n_k+1} - u_{n_k})\right] \end{pmatrix}. \tag{4.4.86}$$

Due to Assumption 4.4.2 (Locally Lipschitz K'), condition (viii), and (4.4.85), we have

$$\begin{aligned} &\left\| \tau_{n_k}\left[K'(u_{n_k})^* - K'(u_{n_k+1})^*\right] v_{n_k+1} \right\|_{\mathcal{U}} \\ &\qquad \leq \tau_{n_k} L' \left\| u_{n_k+1} - u_{n_k} \right\|_{\mathcal{U}} \left\| v_{n_k+1} \right\|_{\mathcal{V}} \to 0 \quad \text{P-a.s.} \end{aligned} \tag{4.4.87}$$

as $k \to \infty$, since $v_{n_k+1} \rightharpoonup \bar{v}$ implies the boundedness of $(v_{n_k+1})_{k\in\mathbb{N}_0}$. Similarly, due to Assumption 4.4.2 (Locally Lipschitz K') and (4.4.6), we get

$$\begin{aligned} &\left\| \sigma\left[K(u_{n_k+1}) - K(u_{n_k}) - K'(u_{n_k})(u_{n_k+1} - u_{n_k})\right] \right\|_{\mathcal{V}} \\ &\qquad \leq \frac{\sigma L'}{2} \left\| u_{n_k+1} - u_{n_k} \right\|_{\mathcal{U}}^2 \to 0 \quad \text{P-a.s.} \end{aligned} \tag{4.4.88}$$

and thus $D_{n_k}^{LIN}(w_{n_k+1}) \to 0$ P-a.s. as $k \to \infty$. For the convergence of the discrepancy from CGF, we recall from (4.1.17) that

$$D_{n_k}^{CGF}(w_{n_k+1}) = \begin{pmatrix} \tau_{n_k}\left[\hat{K}'_{n_k}(u_{n_k})^* - K'(u_{n_k})^*\right](2v_{n_k+1} - v_{n_k}) \\ \sigma\left[K(u_{n_k}) - \hat{K}_{n_k}(u_{n_k})\right] \end{pmatrix}. \tag{4.4.89}$$

Lemma 4.4.11 (ii) applied to every $(\bar{u},\bar{v}) \in \mathcal{Z}$ implies that

$$\mathsf{E}\left(\left\| \left[\hat{K}'_{n_k}\left(u_{n_k}\right)^* - K'\left(u_{n_k}\right)^*\right]\left(2v_{n_k+1} - v_{n_k}\right)\right\|_{\mathcal{U}} \middle| \mathcal{F}_{n_k}\right)$$
$$\leq 2L'R_{\mathcal{U}}\left(3R_{\mathcal{V}} + \Lambda\right) M {n_k}^{-2} \quad (4.4.90)$$

holds P-a.s. for all $k \in \mathbb{N}$, where $\Lambda := \sup_{(\bar{u},\bar{v})\in\mathcal{Z}} \|\bar{v}\|_{\mathcal{V}} < \infty$ since $\mathcal{Z}$ is bounded. Now, Definition 2.5.4 (Conditional Expectation) implies that $\mathsf{E}\left(\mathsf{E}\left(X \middle| \mathcal{F}_{n_k}\right)\mathbf{1}_A\right) = \mathsf{E}\left(X\mathbf{1}_A\right)$ for every $X \in \mathcal{R}\left(\mathcal{F}, \mathcal{U}\right)$ and $A \in \mathcal{F}_{n_k}$. Since $\mathcal{F}_{n_k}$ is a σ-algebra of Θ, we can choose $A := \Theta$, yielding $\mathsf{E}\left(\mathsf{E}\left(X \middle| \mathcal{F}_{n_k}\right)\right) = \mathsf{E}\left(X\right)$. Therefore, taking the expectation of (4.4.90) and summing over k yields

$$\begin{aligned}\sum_{k=1}^{\infty} \mathsf{E}\left(\left\| \left[\hat{K}'_{n_k}\left(u_{n_k}\right)^* - K'\left(u_{n_k}\right)^*\right]\left(2v_{n_k+1} - v_{n_k}\right)\right\|_{\mathcal{U}}\right)& \\ \leq 2L'R_{\mathcal{U}}\left(3R_{\mathcal{V}} + \Lambda\right) M \sum_{k=1}^{\infty} {n_k}^{-2}& \\ \leq 2L'R_{\mathcal{U}}\left(3R_{\mathcal{V}} + \Lambda\right) M \sum_{k=1}^{\infty} k^{-2}.&\end{aligned} \quad (4.4.91)$$

We know from [Dan12] that $\sum_{k=1}^{\infty} k^{-2} = \frac{\pi^2}{6}$, thus the right-hand side of (4.4.91) is finite. This is why we can apply [Kle13, Theorem 6.12 (i)], which gives us the convergence

$$\left[\hat{K}'_{n_k}\left(u_{n_k}\right)^* - K'\left(u_{n_k}\right)^*\right]\left(2v_{n_k+1} - v_{n_k}\right) \to 0 \quad \text{P-a.s.} \quad (4.4.92)$$

as $k \to \infty$. Since $\tau_{n_k} \to \bar{\tau}$ as $k \to \infty$, the first component of $D^{CGF}_{n_k}(w_{n_k+1})$ in (4.4.89) hence converges P-a.s. to 0. For the second component, we analogously apply Lemma 4.4.11 (i) to every $(\bar{u}, \bar{v}) \in \mathcal{Z}$, yielding

$$\mathsf{E}\left(\left\|\hat{K}_{n_k}\left(u_{n_k}\right) - K\left(u_{n_k}\right)\right\|_{\mathcal{V}} \middle| \mathcal{F}_{n_k}\right) \leq 2LR_{\mathcal{U}}M{n_k}^{-2} \quad \text{P-a.s.} \quad (4.4.93)$$

for all $k \in \mathbb{N}$. As above, this implies

$$\sum_{k=1}^{\infty} \mathsf{E}\left(\left\|\hat{K}_{n_k}\left(u_{n_k}\right) - K\left(u_{n_k}\right)\right\|_{\mathcal{V}}\right) \leq 2LR_{\mathcal{U}}M\frac{\pi^2}{6} < \infty \quad (4.4.94)$$

and thus

$$\hat{K}_{n_k}\left(u_{n_k}\right) - K\left(u_{n_k}\right) \to 0 \quad \text{P-a.s.} \quad (4.4.95)$$

as $k \to \infty$. Altogether, we have proven that

$$D_{n_k}\left(w_{n_k+1}\right) = D^{LIN}_{n_k}\left(w_{n_k+1}\right) + D^{CGF}_{n_k}\left(w_{n_k+1}\right) \to 0 \quad \text{P-a.s.} \quad (4.4.96)$$

as $k \to \infty$.

For the first term of the sum in (4.4.79), we note that condition (ix) implies

$$W_{n_k}\begin{pmatrix} -K'\left(u_{n_k+1}\right)^* v_{n_k+1} \\ K\left(u_{n_k+1}\right)\end{pmatrix} \to \overline{W}\begin{pmatrix} -K'\left(\bar{u}\right)^* \bar{v} \\ K\left(\bar{u}\right)\end{pmatrix} = \bar{x} \quad \text{P-a.s.} \quad (4.4.97)$$

since $\left(u_{n_k+1}, v_{n_k+1}\right) \rightharpoonup (\bar{u}, \bar{v})$ as $k \to \infty$.

Altogether, we have shown that $x_{n_k+1} \to \bar{x}$ P-a.s., which implies $\bar{x} \in \overline{W}S(\overline{w})$ P-a.s. and thus $\overline{w} \in H^{-1}(0)$ P-a.s. We know from Assumption 4.4.13 that $(w_k) \in \mathcal{R}(\mathcal{F}, \mathcal{D})$ for all $k \in \mathbb{N}_0$ and since $\mathcal{D}$ is compact and closed, it is also weakly closed according to [CV20, Lemma 1.10]. Therefore, we have $\overline{w} \in \mathcal{D}$ P-a.s. and since $\overline{w} \in H^{-1}(0)$ P-a.s. is also true, it follows from Assumption 4.4.13 (iii) that $\overline{w} \in \mathcal{Z}$ P-a.s.

(v) In order to prove Theorem 4.3.6 (v), let $\left(w_{n_k}\right)_{k\in\mathbb{N}}$ be a weakly convergent subsequence. If we recall the definition of the preconditioning operator M_k in (4.1.18), we see that conditions (viii) and (x) imply

$$M_{n_k} = \begin{pmatrix} \mathrm{Id}_{\mathcal{U}} & \tau_{n_k} K'\left(u_{n_k}\right)^* \\ \sigma K'\left(u_{n_k}\right) & \mathrm{Id}_{\mathcal{V}} \end{pmatrix} \rightharpoonup \begin{pmatrix} \mathrm{Id}_{\mathcal{U}} & \bar{\tau}\left(\overline{K'}\right)^* \\ \sigma\overline{K'} & \mathrm{Id}_{\mathcal{V}} \end{pmatrix} \quad \text{P-a.s.} \tag{4.4.98}$$

as $k \to \infty$. Furthermore, we have already seen that $(\varphi_k)_{k\in\mathbb{N}_0}$ is monotonically increasing and bounded from above. Therefore, there exists $\overline{\varphi} \in (0, \infty)$ such that

$$R_\infty := \begin{pmatrix} \overline{\varphi}\mathrm{Id}_{\mathcal{U}} & \eta\left(\overline{K'}\right)^* \\ \eta\overline{K'} & \psi\mathrm{Id}_{\mathcal{V}} \end{pmatrix} \in \mathcal{R}\left(\mathcal{F}, \mathbb{L}\left(\mathcal{W}, \mathcal{W}\right)\right) \tag{4.4.99}$$

satisfies $Z_{n_k} M_{n_k} w \to R_\infty w$ P-a.s. as $k \to \infty$.

We can now apply Theorem 4.3.6, since we have shown that all of its conditions are satisfied. This yields the existence of $\hat{w} \in \mathcal{R}(\mathcal{F}, \mathcal{Z})$ such that $w_k \rightharpoonup \hat{w}$ P-a.s. as $k \to \infty$, which concludes the proof. □

Note that condition (viii) of Theorem 4.4.14, which requires that there exists a $\underline{\tau} \in (0, \infty)$ such that $\tau_k \geq \underline{\tau}$ for all $k \in \mathbb{N}_0$, implies that the sequence $(\varphi_k)_{k\in\mathbb{N}_0}$ is bounded from above (as we have already seen in the above proof). This conflicts with Assumption 4.4.7 (ii) if $\tilde{\gamma}_G > 0$, since $\varphi_{k+1} = \varphi_k + 2\tilde{\gamma}_G\eta$ implies that $\varphi_k \to \infty$ as $k \to \infty$. However, we can remedy the situation by setting $\tilde{\gamma}_G = 0$ after a finite number of iterations [CV17, Remark 2.1]. This stops the acceleration due to the strong convexity of G but allows us to prove the convergence of the algorithm.

Furthermore, it is not trivial to show under which circumstances the assumption that $u_k \in \mathcal{B}$ P-a.s. for all $k \in \mathbb{N}_0$ in condition (ii) holds. However, as a consequence of the bounded domains of G and CVaR_β^*, we know from Lemma 4.1.1 and Assumption 4.4.13 that $(u_k, v_k) \in \mathcal{D}$ P-a.s. for all $k \in \mathbb{N}_0$, i.e. $u_k \in \mathbb{B}_{R_{\mathcal{U}}}(\bar{u})$ P-a.s. for all $k \in \mathbb{N}_0$, a radius $R_{\mathcal{U}} > 0$, and some $(\bar{u}, \bar{v}) \in H^{-1}(0)$. Hence, if $\mathbb{B}_{R_{\mathcal{U}}}(\bar{u}) \subset \mathcal{B}$ holds, condition (ii) of Theorem 4.4.14 is satisfied automatically. The meaning of this inclusion is that the assumptions from Section 4.4.1 hold at least within the domain of G, which is of course the case if they hold globally. In this case, Theorem 4.4.14 shows global convergence for any starting vector $(u_0, v_0) \in \mathcal{U} \times \mathcal{V}$.

4.4.4 Local Step Size Bound

In condition (ii) of Theorem 4.4.14, we require that $u_k \in \mathcal{B}$ P-a.s. for all $k \in \mathbb{N}_0$, i.e. the primal iterates stay inside a region where the assumptions from Section 4.4.1 are satisfied. The proof that this condition holds under certain assumptions regarding the initial step size was, for a similar algorithm, conducted in [CV20, Lemma 4.6]. However, due to the randomization we introduced

in Section 4.2, this result cannot be easily adapted to our framework. Nevertheless, the following lemma provides a primal step size bound which sets a limit on how far the next iterate can escape from a given neighborhood around a critical point. It is adapted from [CMV19, Lemma 3.8].

Lemma 4.4.15 (Local Step Size Bound)

Let $k \in \mathbb{N}_0$ be arbitrary and $(\bar{u}, \bar{v}) = \bar{w} \in H^{-1}(0)$. Let further $\mathcal{U}_K^1$ be the neighborhood from Definition 4.4.1 (Locally Lipschitz K) and assume that Assumption 4.4.2 (Locally Lipschitz K') holds for a neighborhood $\mathcal{U}_K^2$. Furthermore, we assume that

(i) the inclusions

$$u_j \in \mathbb{B}_{r_{\mathcal{U},k}}(\bar{u}) \quad \text{for all } j \in \{0, \dots, k\},$$
$$\text{and} \quad v_j \in \mathbb{B}_{R_{\mathcal{V}}}(\bar{v}) \quad \text{for all } j \in \{k, k+1\},$$

are satisfied for some radii $r_{\mathcal{U},k}, R_{\mathcal{V}} > 0$;

(ii) $(u_{k+1}, v_{k+1}) = w_{k+1}$ satisfies (RI);

(iii) the step size bound

$$\tau_k \leq \frac{\delta_{\mathcal{U},k}}{\left(3R_{\mathcal{V}} + \|\bar{v}\|_{\mathcal{V}}\right)\left(\|K'(u_k)\|_{\mathbb{L}(\mathcal{U},\mathcal{V})} + 2L' r_{\mathcal{U},k}\Gamma_k\right) + \left\|K'(\bar{u})^* \bar{v}\right\|_{\mathcal{U}}}$$

holds P-a.s. for $\Gamma_k := \sum_{j=1}^{k} \prod_{i=j}^{k} \operatorname{ess\,sup}_{\xi \in \Xi} \chi_{A_i^c}(\xi)$ and a constant $\delta_{\mathcal{U},k} \geq 0$;

(iv) the inclusion $\mathbb{B}_{r_{\mathcal{U},k}+\delta_{\mathcal{U},k}}(\bar{u}) \subset \mathcal{U}_K^1 \cap \mathcal{U}_K^2$ is true.

Then, the following inclusion holds as well:

$$u_{k+1} \in \mathbb{B}_{r_{\mathcal{U},k}+\delta_{\mathcal{U},k}}(\bar{u}).$$

Proof: Unless stated otherwise, all equations in this proof are meant to hold P-a.s. We need to show that

$$\|u_{k+1} - \bar{u}\|_{\mathcal{U}} \leq r_{\mathcal{U},k} + \delta_{\mathcal{U},k} \tag{4.4.100}$$

holds. We do this by dividing the proof into two parts. In the first part, we show that

$$\|u_{k+1} - \bar{u}\|_{\mathcal{U}} \leq r_{\mathcal{U},k} + C_{\mathcal{U},k+1} \tag{4.4.101}$$

for some $C_{\mathcal{U},k+1} \geq 0$, and in the second part, we prove that

$$C_{\mathcal{U},k+1} \leq \delta_{\mathcal{U},k} \tag{4.4.102}$$

follows from the assumed step size bound.

(i) Since w_{k+1} satisfies (RI), we can multiply the inclusion by $(u_{k+1} - \bar{u}, 0)^{\mathsf{T}} \in \mathcal{W}$, yielding

$$\begin{aligned} 0 \in \tau_k \left\langle \partial G(u_{k+1}) + K'(u_{k+1})^* v_{k+1}, u_{k+1} - \bar{u} \right\rangle_{\mathcal{U}} + \tau_k \left\langle D_{k,1}(w_{k+1}), u_{k+1} - \bar{u} \right\rangle_{\mathcal{U}} \\ + \left\langle u_{k+1} - u_k + \tau_k K'(u_k)^* (v_{k+1} - v_k), u_{k+1} - \bar{u} \right\rangle_{\mathcal{U}}, \end{aligned} \tag{4.4.103}$$

where

$$D_{k,1}(w_{k+1}) := \left[K'(u_k)^* - K'(u_{k+1})^*\right] v_{k+1} + \left[\hat{K}_k'(u_k)^* - K'(u_k)^*\right](2v_{k+1} - v_k). \quad (4.4.104)$$

The three-point version of Pythagoras' identity

$$\langle u_{k+1} - u_k, u_{k+1} - \bar{u}\rangle_{\mathcal{U}} = \frac{1}{2}\|u_{k+1} - u_k\|_{\mathcal{U}}^2 - \frac{1}{2}\|u_k - \bar{u}\|_{\mathcal{U}}^2 + \frac{1}{2}\|u_{k+1} - \bar{u}\|_{\mathcal{U}}^2 \quad (4.4.105)$$

yields

$$\begin{aligned}\|u_k - \bar{u}\|_{\mathcal{U}}^2 \in\ & 2\tau_k\left\langle \partial G(u_{k+1}) + K'(u_{k+1})^* v_{k+1} + D_{k,1}(w_{k+1}), u_{k+1} - \bar{u}\right\rangle_{\mathcal{U}} \\ & + 2\tau_k\left\langle K'(u_k)^*(v_{k+1} - v_k), u_{k+1} - \bar{u}\right\rangle_{\mathcal{U}} \\ & + \|u_{k+1} - u_k\|_{\mathcal{U}}^2 + \|u_{k+1} - \bar{u}\|_{\mathcal{U}}^2.\end{aligned} \quad (4.4.106)$$

Since $0 \in H(\overline{w})$, we can subtract $\left\langle \partial G(\bar{u}) + K'(\bar{u})^* \bar{v}, u_{k+1} - \bar{u}\right\rangle_{\mathcal{U}}$, which implies

$$\begin{aligned}\|u_k - \bar{u}\|_{\mathcal{U}}^2 \in\ & 2\tau_k\langle \partial G(u_{k+1}) - \partial G(\bar{u}), u_{k+1} - \bar{u}\rangle_{\mathcal{U}} \\ & + 2\tau_k\left\langle K'(u_{k+1})^* v_{k+1} - K'(\bar{u})^*\bar{v} + D_{k,1}(w_{k+1}), u_{k+1} - \bar{u}\right\rangle_{\mathcal{U}} \\ & + 2\tau_k\left\langle K'(u_k)^*(v_{k+1} - v_k), u_{k+1} - \bar{u}\right\rangle_{\mathcal{U}} \\ & + \|u_{k+1} - u_k\|_{\mathcal{U}}^2 + \|u_{k+1} - \bar{u}\|_{\mathcal{U}}^2.\end{aligned} \quad (4.4.107)$$

The first term on the right-hand side can be estimated using the monotonicity of ∂G such that the Cauchy-Schwarz inequality yields

$$\|u_k - \bar{u}\|_{\mathcal{U}}^2 \geq -2C_{\mathcal{U},k+1}\|u_{k+1} - \bar{u}\|_{\mathcal{U}} + \|u_{k+1} - u_k\|_{\mathcal{U}}^2 + \|u_{k+1} - \bar{u}\|_{\mathcal{U}}^2, \quad (4.4.108)$$

with

$$C_{\mathcal{U},k+1} := \tau_k\left\|K'(u_{k+1})^* v_{k+1} - K'(\bar{u})^*\bar{v} + K'(u_k)^*(v_{k+1} - v_k) + D_{k,1}(w_{k+1})\right\|_{\mathcal{U}}. \quad (4.4.109)$$

If we rearrange (4.4.108) and use that $\|u_{k+1} - \bar{u}\|_{\mathcal{U}} \leq \|u_{k+1} - u_k\|_{\mathcal{U}} + \|u_k - \bar{u}\|_{\mathcal{U}}$, we get

$$\|u_{k+1} - u_k\|_{\mathcal{U}}^2 + \|u_{k+1} - \bar{u}\|_{\mathcal{U}}^2 \leq \|u_k - \bar{u}\|_{\mathcal{U}}^2 + 2C_{\mathcal{U},k+1}\left(\|u_{k+1} - u_k\|_{\mathcal{U}} + \|u_k - \bar{u}\|_{\mathcal{U}}\right), \quad (4.4.110)$$

which is equivalent to

$$\left(\|u_{k+1} - u_k\|_{\mathcal{U}} - C_{\mathcal{U},k+1}\right)^2 + \|u_{k+1} - \bar{u}\|_{\mathcal{U}}^2 \leq \left(\|u_k - \bar{u}\|_{\mathcal{U}} + C_{\mathcal{U},k+1}\right)^2. \quad (4.4.111)$$

Therefore, we have

$$\| u_{k+1} - \bar{u} \|_{\mathcal{U}} \leq \| u_k - \bar{u} \|_{\mathcal{U}} + C_{\mathcal{U},k+1} \leq r_{\mathcal{U},k} + C_{\mathcal{U},k+1}. \tag{4.4.112}$$

(ii) Now, we prove that $C_{\mathcal{U},k} \leq \delta_{\mathcal{U},k}$. Expanding the definition of $D_{k,1}(w_{k+1})$ in (4.4.104) yields

$$\begin{aligned} C_{\mathcal{U},k+1} = \tau_k \Big\| -K'(\bar{u})^* \bar{v} + K'(u_k)^* (2v_{k+1} - v_k) \\ + \left[\hat{K}_k'(u_k)^* - K'(u_k)^* \right] (2v_{k+1} - v_k) \Big\|, \end{aligned} \tag{4.4.113}$$

and together with equation (4.4.47) it follows that

$$\begin{aligned} \tau_k^{-1} C_{\mathcal{U},k+1} \leq \left\| K'(\bar{u})^* \bar{v} \right\|_{\mathcal{U}} + \| K'(u_k) \|_{\mathbb{L}(\mathcal{U},\mathcal{V})} \| 2v_{k+1} - v_k \|_{\mathcal{V}} \\ + \sum_{j=1}^{k} \left\| K'(u_{j-1}) - K'(u_j) \right\|_{\mathbb{L}(\mathcal{U},\mathcal{V})} \left\| \Pi_{\cap_{i=j}^{k} A_i^c} (2v_{k+1} - v_k) \right\|_{\mathcal{V}}. \end{aligned} \tag{4.4.114}$$

Since we assumed that $v_j \in \mathbb{B}_{R_{\mathcal{V}}}(\bar{v})$ for all $j \in \{k, k+1\}$, we can use the inequality $\| 2v_{k+1} - v_k \|_{\mathcal{V}} \leq 2 \| v_{k+1} - \bar{v} \|_{\mathcal{V}} + \| v_k - \bar{v} \|_{\mathcal{V}} + \| \bar{v} \|_{\mathcal{V}} \leq 3R_{\mathcal{V}} + \| \bar{v} \|_{\mathcal{V}}$ together with the local Lipschitz continuity of K' and the reasoning of the proof of Lemma 4.4.11 to see that

$$\begin{aligned} \tau_k^{-1} C_{\mathcal{U},k+1} \leq \left\| K'(\bar{u})^* \bar{v} \right\|_{\mathcal{U}} + \| K'(u_k) \|_{\mathbb{L}(\mathcal{U},\mathcal{V})} \left(3R_{\mathcal{V}} + \| \bar{v} \|_{\mathcal{V}} \right) \\ + 2L' r_{\mathcal{U},k} \| 2v_{k+1} - v_k \|_{\mathcal{V}} \sum_{j=1}^{k} \operatorname*{ess\,sup}_{\xi \in \Xi} \prod_{i=j}^{k} \chi_{A_i^c}(\xi). \end{aligned} \tag{4.4.115}$$

Here, we can interchange the essential supremum and the product operator, yielding

$$\begin{aligned} \tau_k^{-1} C_{\mathcal{U},k+1} \leq \left(3R_{\mathcal{V}} + \| \bar{v} \|_{\mathcal{V}} \right) \left(\| K'(u_k) \|_{\mathbb{L}(\mathcal{U},\mathcal{V})} + 2L' r_{\mathcal{U},k} \sum_{j=1}^{k} \prod_{i=j}^{k} \operatorname*{ess\,sup}_{\xi \in \Xi} \chi_{A_i^c}(\xi) \right) \\ + \left\| K'(\bar{u})^* \bar{v} \right\|_{\mathcal{U}}. \end{aligned} \tag{4.4.116}$$

The desired inequality $C_{\mathcal{U},k+1} \leq \delta_{\mathcal{U},k}$ now follows directly from the assumed primal step size bound. □

Chapter 5

Simplex Projection

In this chapter, we describe the metric projection onto the so-called *bounded probability simplex*, which turns out to be equal to the proximal operator of CVaR_β^*, once we have discretized the probability space. After presenting this discretization in the first section, we state the optimization problem that needs to be solved in order to compute the projection, followed by the corresponding optimality condition. In the fourth section, we present an efficient algorithm for computing this projection and prove its convergence afterwards. The content of Sections 5.2 to 5.5 is mostly taken from the master's thesis [Ang18], where a similar problem was considered.

5.1 Discretization of the Probability Space

In Section 3.1 we defined a separable probability space $(\Xi, \mathcal{A}, \mathbb{P})$ and the space of random variables $\mathcal{V} := L^2_{\mathbb{P}}(\Xi) := L^2(\Xi, \mathcal{A}, \mathbb{P})$. The approach we use to discretize the probability space is the sample average approximation (SAA) introduced in Section 2.5.2. Therefore, we need to bring the objective function into a form like (2.5.8). Since G does not involve random variables, we only need to focus on $\mathrm{CVaR}_\beta \circ K$ here. If we assume that the PDE constraint introduced in Section 3.2 has $d \in \mathbb{N}$ uncertain coefficients in $\mathbb{R}$, we can use a d-dimensional random vector

$$\hat{x} := (\hat{x}_1, \dots, \hat{x}_d) : \Xi \to \mathbb{R}^d \tag{5.1.1}$$

with $\hat{x}_i \in \mathcal{V}$ for all $i \in \{1, \dots, d\}$ to model this uncertainty. Let then $f \colon \mathcal{U} \times \mathbb{R}^d \to \mathbb{R}$ be given such that

$$f\big(u, \hat{x}(\cdot)\big) = K(u)(\cdot), \tag{5.1.2}$$

$$\text{i.e.} \quad \Xi \ni \xi \mapsto f\big(u, \hat{x}(\xi)\big) = K(u)(\xi) \tag{5.1.3}$$

for all $u \in \mathcal{U}$. Therefore, the first term of the sum of the objective function of Problem 3.1.1 can be written as

$$\begin{aligned} \mathrm{CVaR}_\beta\big(K(u)\big) &= \mathrm{CVaR}_\beta\big(f(u, \hat{x}(\cdot))\big) \\ &= \min_{\alpha \in \mathbb{R}} \left(\alpha + \frac{1}{1-\beta} \mathbb{E}\left(\big(f\big(u, \hat{x}(\cdot)\big) - \alpha\big)^+ \right) \right) \end{aligned} \tag{5.1.4}$$

for all $u \in \mathcal{U}$, where we have used Proposition 2.5.11 (ii) for the last equality. Now, we assume that we have a sample $\left\{ \xi^j \in \mathbb{R}^d \;\middle|\; j \in \{1, \dots, S\} \right\}$ of $S \in \mathbb{N}$ realizations of the random vector $\hat{x}$, e.g. obtained by Monte Carlo sampling. The discretization of the integral within the expected

value yields the approximation

$$\mathrm{CVaR}_\beta(\hat{z}) \approx \min_{\alpha\in\mathbb{R}} \left(\alpha + \frac{1}{(1-\beta)S} \sum_{j=1}^{S} (z_j - \alpha)^+ \right), \tag{5.1.5}$$

where $\hat{z} := K(u)$ for some $u \in \mathcal{U}$ and $z_j := \hat{z}\left(\xi^j\right) = K(u)\left(\xi^j\right) = f\left(u, \xi^j\right) \in \mathbb{R}$ for all $j \in \{1, \dots, S\}$. In the following, we will denote the right-hand side of (5.1.5) by $\mathrm{CVaR}_\beta(\mathbf{z})$ with $\mathbf{z} \in \mathbb{R}^S$, so the argument of CVaR_β provides the information whether we are dealing with the exact CVaR_β or its approximation.

Since the dual step of Algorithm 4.1 requires us to compute the proximal operator of CVaR_β^*, we need to investigate how this operator can be approximated. The following lemma is a modification of [Ang18, Lemma 3.2.1] and shows that CVaR_β^* is equal to the metric projection (with respect to the weighted norm) onto a specific set.

Lemma 5.1.1 (Proximal Operator of CVaR_β^*)

Let $S \in \mathbb{N}$ be the sample size, $\Sigma \in \mathbb{L}\left(\mathbb{R}^S, \mathbb{R}^S\right)$ a self-adjoint and strongly monotone operator, $\beta \in (0, 1)$ a probability level, and CVaR_β the approximation as defined in (5.1.5). Then, the proximal operator of the Fenchel conjugate of CVaR_β is, for all $\mathbf{z} \in \mathbb{R}^S$, given by

$$\mathrm{prox}^{\Sigma}_{\mathrm{CVaR}_\beta^*}(\mathbf{z}) = \mathrm{proj}^{\Sigma}_{\Delta}(\mathbf{z}) := \arg\min_{\mathbf{y}\in\Delta} \frac{1}{2} \| \mathbf{z} - \mathbf{y} \|^2_{\Sigma^{-1}},$$

where

$$\Delta := \left\{ \mathbf{y} \in \mathbb{R}^S \;\middle|\; \mathbf{y}^\top \mathbf{1} = 1 \text{ and } 0 \le y_j \le \frac{1}{(1-\beta)S} \text{ for all } j \in \{1, \dots, S\} \right\}$$

is the so-called *bounded probability simplex* and $\mathbf{1} \in \mathbb{R}^S$ denotes the vector of all ones.

Proof: Let $\mathbf{z} \in \mathbb{R}^S$. First, we recall from (5.1.5) that

$$\mathrm{CVaR}_\beta(\mathbf{z}) = \min_{\alpha\in\mathbb{R}} \left(\alpha + \frac{1}{(1-\beta)S} \sum_{j=1}^{S} (z_j - \alpha)^+ \right). \tag{5.1.6}$$

The use of the function $(\cdot)^+$ can be avoided by introducing new variables $w_j \in \mathbb{R}$ for every $j \in \{1, \dots, S\}$. The minimization on the right-hand side is then equivalent to

$$\begin{aligned} \min_{\alpha\in\mathbb{R},\, \mathbf{w}\in\mathbb{R}^S} \quad & \alpha + \frac{1}{(1-\beta)S} \sum_{j=1}^{S} w_j \\ \text{s.t.} \quad & \alpha + w_j \ge z_j \qquad \forall j \in \{1, \dots, S\}, \\ & w_j \ge 0 \qquad \forall j \in \{1, \dots, S\}. \end{aligned} \tag{5.1.7}$$

This linear optimization problem has a non-empty feasible set (e.g. $w_j := \max\{0, z_j\}$ for all $j \in \{1, \dots, S\}$ and $\alpha := 0$ satisfy its constraints), and since

$$\begin{aligned} \alpha + \frac{1}{(1-\beta)S} \sum_{j=1}^{S} w_j &\geq \frac{1}{S} \sum_{j=1}^{S} (\alpha + w_j) \\ &\geq \frac{1}{S} \sum_{j=1}^{S} z_j \\ &> -\infty, \end{aligned} \tag{5.1.8}$$

its objective function is bounded from below. Thus, problem (5.1.7) has an optimal solution and is, due to strong duality [BT97b, Theorem 4.4], equivalent to its dual problem. Following [BT97b, Chapter 4.2], the dual can be written as

$$\begin{aligned} \max_{\mathbf{y} \in \mathbb{R}^S} \quad & \sum_{j=1}^{S} y_j z_j \\ \text{s.t.} \quad & 0 \leq y_j \leq \frac{1}{(1-\beta)S} \quad \forall j \in \{1, \dots, S\}, \\ & \sum_{j=1}^{S} y_j = 1. \end{aligned} \tag{5.1.9}$$

Now, the definitions of Δ and the indicator function δ_Δ immediately imply

$$\mathrm{CVaR}_\beta(\mathbf{z}) = \sup_{\mathbf{y} \in \mathbb{R}^S} \left(\mathbf{y}^\top \mathbf{z} - \delta_\Delta(\mathbf{y}) \right). \tag{5.1.10}$$

This is, according to Definition 2.2.6, the Fenchel conjugate of δ_Δ, i.e.

$$\mathrm{CVaR}_\beta(\mathbf{z}) = \delta_\Delta^*(\mathbf{z}). \tag{5.1.11}$$

Since Δ is convex and closed, δ_Δ is weakly lower semi-continuous [CV20, Lemma 2.5] and convex, and thus also lower semi-continuous [CV20, Corollary 3.2]. Furthermore, Δ is non-empty, hence δ_Δ is proper. We can therefore apply Theorem 2.2.7 (Fenchel-Moreau-Rockafellar) with equality, yielding

$$\mathrm{CVaR}_\beta^*(\mathbf{z}) = \delta_\Delta^{**}(\mathbf{z}) = \delta_\Delta(\mathbf{z}). \tag{5.1.12}$$

Now, Definition 2.2.14 (Weighted Proximal Operator) implies that

$$\begin{aligned} \mathrm{prox}^{\Sigma}_{\mathrm{CVaR}_\beta^*}(\mathbf{z}) &= \arg\min_{\mathbf{y} \in \mathbb{R}^S} \left(\frac{1}{2} \| \mathbf{z} - \mathbf{y} \|^2_{\Sigma^{-1}} + \delta_\Delta(\mathbf{y}) \right) \\ &= \arg\min_{\mathbf{y} \in \Delta} \frac{1}{2} \| \mathbf{z} - \mathbf{y} \|^2_{\Sigma^{-1}} \\ &= \mathrm{proj}^{\Sigma}_{\Delta}(\mathbf{z}). \end{aligned} \tag{5.1.13}$$

□

Apparently, the metric projection (with respect to the weighted norm) onto the bounded probability simplex plays a major role in Algorithm 4.1, since it needs to be computed in every iteration to update the dual variable. Fortunately, under an additional assumption, this projection can be computed very efficiently by using a specially tailored algorithm, which we develop in the following sections.

5.2 Problem Formulation

As in Section 5.1, we assume that $S \in \mathbb{N}$ is the sample size and $\beta \in (0,1)$ the probability level of CVaR_β. Additionally, we assume that there exist $\sigma_1, \dots, \sigma_S > 0$ such that $\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_S)$. Recall that the bounded probability simplex is defined as the set

$$\Delta := \left\{ \mathbf{y} \in \mathbb{R}^S \;\middle|\; \mathbf{y}^\top \mathbf{1} = 1 \text{ and } 0 \le y_j \le \frac{1}{(1-\beta)S} \text{ for all } j \in \{1, \dots, S\} \right\}. \tag{5.2.1}$$

If we want to compute $\operatorname{proj}_\Delta^\Sigma(\mathbf{z})$ for any $\mathbf{z} = (z_1, \dots, z_S) \in \mathbb{R}^S$, we need to find the solution to the following optimization problem:

$$\begin{aligned} \min_{\mathbf{y} \in \mathbb{R}^S} \quad & \frac{1}{2} \| \mathbf{z} - \mathbf{y} \|_{\Sigma^{-1}}^2 \\ \text{s.t.} \quad & 0 \le y_j \le \frac{1}{(1-\beta)S} \quad \forall j \in \{1, \dots, S\}, \\ & \sum_{j=1}^S y_j = 1. \end{aligned} \tag{5.2.2}$$

This is a quadratic program with a strictly convex objective function, and several solutions have been described for similar problems. For example, John Duchi et al. developed an algorithm to compute the projection onto the *positive simplex* [Duc+08], i.e. they considered a set like Δ but with an arbitrary value on the right-hand side of the equation in (5.2.1) and no upper bound on $\mathbf{y}$. The key idea of their proof is sorting the elements of the point being projected in a descending order and finding the last index that satisfies a certain condition. Weiran Wang and Miguel Á. Carreira-Perpiñán slightly changed this problem and used the so called *probability simplex* instead [WC13], which means that the condition $\mathbf{y}^\top \mathbf{1} = 1$ has to be satisfied for all $\mathbf{y}$ in the set (i.e. the right-hand side of the equation is not arbitrary anymore). They proposed a proof that uses the KKT conditions and also the sorting as in the previously mentioned paper. A similar approach was used by Nelson Maculan and Geraldo Galdino de Paula Jr. [MdP89].

All aforementioned authors did not consider the possibility of an upper bound on the elements of the respective set like $y_j \le \frac{1}{(1-\beta)S}$ in (5.2.1). Yunmei Chen and Xiaojing Ye used tools from the field of convex analysis like the proximal operator and the Fenchel conjugate to construct an algorithm where this upper bound is equal to 1 [CY11].

The algorithm presented in the following sections is based on the ideas in [WC13] and modified such that there is an upper bound on $\mathbf{y}$. This upper bound can be any real, positive number. While the optimality condition in the next section still allows for a diagonal weight matrix $\Sigma \in \mathbb{R}^{S \times S}$ as defined above, this assumption is tightened for the algorithm in Section 5.4 by requiring that $\Sigma = \sigma \mathrm{Id}$ with a $\sigma > 0$.

5.3 Optimality Condition

In order to derive the optimality condition for the above problem (5.2.2), we apply the KKT conditions [NW06, Theorem 12.1]. The Lagrangian of the problem is

$$\mathcal{L}(\mathbf{y}, \nu, \mu, \lambda) := \frac{1}{2} \| \mathbf{z} - \mathbf{y} \|^2_{\Sigma^{-1}} - \nu^\top \mathbf{y} - \mu^\top (p\mathbf{1} - \mathbf{y}) - \lambda \left(\mathbf{y}^\top \mathbf{1} - 1 \right) \tag{5.3.1}$$

where $\lambda \in \mathbb{R}$ and $\nu, \mu \in \mathbb{R}^S$ are the Lagrange multipliers and $p := \frac{1}{(1-\beta)S}$. The following KKT conditions hold for the optimal solution $\mathbf{y}^*$:

$$\nabla_{\mathbf{y}} \mathcal{L}(\mathbf{y}^*, \nu^*, \mu^*, \lambda^*) = \Sigma (\mathbf{y}^* - \mathbf{z}) - \nu^* + \mu^* - \lambda^* \mathbf{1} = 0, \tag{5.3.2}$$

$$\mathbf{y}^{*\top} \mathbf{1} - 1 = 0, \tag{5.3.3}$$

$$y_j^* \geq 0 \qquad \forall j \in \{1, \ldots, S\}, \tag{5.3.4}$$

$$p - y_j^* \geq 0 \qquad \forall j \in \{1, \ldots, S\}, \tag{5.3.5}$$

$$\nu_j^* \geq 0 \qquad \forall j \in \{1, \ldots, S\}, \tag{5.3.6}$$

$$\mu_j^* \geq 0 \qquad \forall j \in \{1, \ldots, S\}, \tag{5.3.7}$$

$$y_j^* \nu_j^* = 0 \qquad \forall j \in \{1, \ldots, S\}, \tag{5.3.8}$$

$$\mu_j^* \left(p - y_j^* \right) = 0 \qquad \forall j \in \{1, \ldots, S\}. \tag{5.3.9}$$

Now we use the complementarity conditions (5.3.8) and (5.3.9) to derive an explicit description of the optimal solution $\mathbf{y}^*$. Let $j \in \{1, \ldots, S\}$. We distinguish the following cases:

- if $y_j^* > 0$ then $\nu_j^* = 0$;
  - if $y_j^* < p$ then $\mu_j^* = 0$; it follows from (5.3.2) that

$$p > y_j^* = z_j + \sigma_j^{-1} \lambda^* > 0; \tag{5.3.10}$$

  - if $y_j^* = p$ then $\mu_j^* \geq 0$; it follows from (5.3.2) that

$$z_j + \sigma_j^{-1} \lambda^* \geq z_j + \sigma_j^{-1} \left(\lambda^* - \mu_j^* \right) = y_j^* = p > 0; \tag{5.3.11}$$

- if $y_j^* = 0$ then $\nu_j^* \geq 0$; since $p > 0$ we have $y_j^* < p$ and therefore $\mu_j^* = 0$; it follows from (5.3.2) that

$$z_j + \sigma_j^{-1} \lambda^* = -\sigma_j^{-1} \nu_j^* \leq 0. \tag{5.3.12}$$

It can easily be shown that the following $\mathbf{y}^*$ satisfies the KKT conditions in all of the above cases:

$$y_j^* = \max \left\{ \min \left\{ z_j + \sigma_j^{-1} \lambda^*, p \right\}, 0 \right\}. \tag{5.3.13}$$

Since we do not know the Lagrangian multiplier λ^* in advance, we need to develop an algorithm that is able to compute it.

5.4 Algorithm

As mentioned before, we additionally assume from now on that there exists $\sigma > 0$ such that $\Sigma = \sigma \mathrm{Id} \in \mathbb{L}\left(\mathbb{R}^S, \mathbb{R}^S\right)$. In this case, the metric projection in Lemma 5.1.1 simplifies to the Euclidean projection

$$\mathrm{proj}_{\Delta}^{\Sigma}(\mathbf{z}) = \operatorname*{arg\,min}_{\mathbf{y}\in\Delta} \frac{\sigma}{2} \| \mathbf{z} - \mathbf{y} \|_2^2 = \operatorname*{arg\,min}_{\mathbf{y}\in\Delta} \frac{1}{2} \| \mathbf{z} - \mathbf{y} \|_2^2 =: \mathrm{proj}_{\Delta}(\mathbf{z}), \tag{5.4.1}$$

which does not depend on the operator Σ anymore. Therefore, problem (5.2.2) simplifies to

$$\begin{aligned} \min_{\mathbf{y}\in\mathbb{R}^S} \quad & \frac{1}{2} \| \mathbf{z} - \mathbf{y} \|_2^2 \\ \text{s.t.} \quad & 0 \le y_j \le \frac{1}{(1-\beta)S} \quad \forall j \in \{1, \dots, S\}, \\ & \sum_{j=1}^{S} y_j = 1, \end{aligned} \tag{5.4.2}$$

where the weighted norm $\| \cdot \|_{\Sigma^{-1}}$ has been replaced by the Euclidean norm $\| \cdot \|_2$. The corresponding optimality condition is obtained by choosing $\sigma := 1$ in (5.3.13), yielding

$$y_j^* = \max\left\{ \min\left\{ z_j + \lambda^*, p \right\}, 0 \right\}. \tag{5.4.3}$$

Algorithm 5.1 uses this optimality condition to find the optimal solution $\mathbf{y}^*$ to problem (5.4.2).

Algorithm 5.1 (Euclidean Projection onto the Bounded Probability Simplex)

Input: $\mathbf{z} \in \mathbb{R}^S$, $p = \frac{1}{(1-\beta)S}$

Sort $\mathbf{z}$ into $\mathbf{u}$: $u_1 \ge u_2 \ge \cdots \ge u_S$, define $u_0 := u_1 + p$.

Compute $\zeta := \max\left\{ j \in \{0, \dots, S\} \;\middle|\; jp + \sum_{i=j+1}^{S} \max\{0, u_i + p - u_j\} \le 1 \right\}$.

if $\zeta p + \sum_{i=\zeta+1}^{S} \max\{0, u_i + p - u_\zeta\} = 1$ **then**

  $\lambda^* := p - u_\zeta$

else

  $\rho := \max\left\{ j \in \{\zeta + 1, \dots, S\} \;\middle|\; u_j + \frac{1}{j-\zeta}\left(1 - \zeta p - \sum_{i=\zeta+1}^{j} u_i\right) > 0 \right\}$

  $\lambda^* := \frac{1}{\rho-\zeta}\left(1 - \zeta p - \sum_{i=\zeta+1}^{\rho} u_i\right)$

end

Output: $\mathbf{y}^* \in \mathbb{R}^S$ with $y_j^* := \max\left\{ \min\left\{ z_j + \lambda^*, p \right\}, 0 \right\}$ for all $j \in \{1, \dots, S\}$

The algorithm has the following geometric interpretation: Place the values $z_1, \dots, z_S$ as points on the abscissa. Then the optimal solution is given by a shift of these points such that the ones

to the right of the ordinate sum up to 1, and whenever a point would be greater than p, it is set to p. An implementation in Julia can be found in [Ang22, SimplexProj.jl].

Note that finding ζ in Line 2 of Algorithm 5.1 could take a great share of the computation time, especially if it is done by simply checking the condition

$$jp + \sum_{i=j+1}^{S} \max\{0, u_i + p - u_j\} \leq 1 \tag{5.4.4}$$

for every $j \in \{0, \dots, S\}$, starting at $j = 0$ and terminating if the condition is not satisfied any more. Certainly, the actual computational cost depends on the given input $\mathbf{z}$ and p, but yet it is reasonable to use a smarter way of finding ζ.

Let us assume that we want to find the greatest integer $j \in \{m, \dots, n\}$ $(m, n \in \mathbb{N})$ such that $\varphi(j) \leq 1$, where $\varphi\colon \{m, \dots, n\} \to \mathbb{R}$ is a non-decreasing function. In our framework, φ would be defined as

$$j \mapsto \varphi(j) := jp + \sum_{i=j+1}^{S} \max\{0, u_i + p - u_j\}. \tag{5.4.5}$$

In inequality (5.5.11) we show that φ is indeed non-increasing. The recursive Algorithm 5.2 is a very simple solution that can in fact perform better in some cases than the successive checking of $\varphi(j) \leq 1$, as we will see in Section 6.3.6.

Algorithm 5.2 (Find ζ)

Input: $m, n \in \mathbb{N}$, $\varphi\colon \{m, \dots, n\} \to \mathbb{R}$ non-decreasing

1 **if** $\varphi(m) > 1$ **then** $\zeta := m - 1$
2 **else if** $\varphi(n) \leq 1$ **then** $\zeta := n$
3 **else if** $m + 1 = n$ **then** $\zeta := m$
4 **else**
5 $\quad j := m + \lfloor \frac{n-m}{2} \rfloor$
6 $\quad$ Run the algorithm again with $\{m, \dots, j\}$ as the domain of φ; the solution is denoted by k.
7 $\quad$ **if** $k = j$ **then**
8 $\quad\quad$ Run the algorithm again with $\{j + 1, \dots, n\}$ as the domain of φ; the solution is denoted by l.
9 $\quad\quad$ **if** $l = j$ **then** $\zeta := k$
10 $\quad\quad$ **else** $\zeta := l$
11 $\quad$ **else** $\zeta := k$
12 **end**

Output: ζ

Lines 1 to 3 handle some trivial cases. Then the domain of φ is cut in two halves (see Line 5) and the algorithm is run again with the lower half as the domain of φ (see Line 6). If the solution k is not equal to the new upper bound j, it must be the final solution (see Line 11). Otherwise, the final solution is either k itself or it lies in the upper half of the domain, i.e. in $\{j+1,\dots,S\}$. This is checked by running the algorithm again with the upper half as the domain of φ (see Line 8). Note that $\lfloor\cdot\rfloor$ denotes the function of rounding down to the next integer that is less or equal to the argument. An implementation in Julia can be found in [Ang22, SimplexProj.jl].

Algorithm 5.3 (Find ζ With an Initial Guess)

Input: $\mathbf{u} \in \mathbb{R}^S$ s.t. $u_1 \geq \cdots \geq u_S$, $p = \frac{1}{(1-\beta)S}$, *optional:* initial guess $\zeta_0 \in \{1,\dots,n\}$

1 Define $\varphi(j) := jp + \sum_{i=j+1}^{n} \max\{0, u_i + p - u_j\}$ for every $j \in \{1,\dots,S\}$.
2 **if** ζ_0 *is empty* **then**
3 Call Algorithm 5.2 with $\varphi\colon \{1,\dots,S\} \to \mathbb{R}$ and denote the solution by ζ.
4 **else**
5 $\zeta_1 := \zeta_0$
6 **while** *true* **do**
7 **if** $\varphi(\zeta_1) \leq 1$ **then**
8 $\zeta := \zeta_1$
9 Update $\zeta_1 \leftarrow \zeta_1 + 1$
10 **else**
11 **if** *the previous if-condition was true during the last iteration* **then** break
12 **else**
13 Update $\zeta_1 \leftarrow \zeta_1 - 1$
14 **if** $\zeta_1 > 0$ **then** $\zeta := \zeta_1$
15 **else** break
16 **end**
17 **end**
18 **end**
19 **end**

Output: ζ

Since Algorithm 5.1 is called in every iteration of Algorithm 4.1, we can make another improvement that might speed up the search for ζ. In this case, it is reasonable to assume that the input of Algorithm 5.1 is not significantly different from the input of the previous function call. Therefore, the index ζ is likely to change only slightly from one call to the next one. Algorithm 5.3 enables us to pass an initial guess of ζ and iterate up- or downward until the real ζ is found. However, if no initial guess is provided, it just calls Algorithm 5.2. An implementation in Julia can also be found in [Ang22, SimplexProj.jl].

5.5 Convergence

The proof of convergence of Algorithm 5.1 is based on [WC13, Chapter 3] and extended such that it is applicable to our modified assumptions.

Theorem 5.5.1 (Convergence of Projection Algorithm)

Let $S \in \mathbb{N}$ be the sample size, $\beta \in (0,1)$ a probability level, $p := \frac{1}{(1-\beta)S}$, and $\mathbf{z} \in \mathbb{R}^S$. Then, Algorithm 5.1 finds the optimal solution to the problem (5.4.2), i.e. the output $\mathbf{y}^* = (y_1^*, \dots, y_S^*)$ satisfies

$$y_j^* = \max\left\{\min\left\{z_j + \lambda^*, p\right\}, 0\right\}$$

for all $j \in \{1, \dots, S\}$ and some $\lambda^* \in \mathbb{R}$.

Proof: Without loss of generality, we assume that the components of $\mathbf{z}$ are sorted and $\mathbf{y}^*$ uses the same ordering, i.e.

$$z_1 \geq \cdots \geq z_S \quad \text{and} \quad y_1^* \geq \cdots \geq y_S^*. \tag{5.5.1}$$

Therefore, we do not need to consider the sorting in Line 1 of Algorithm 5.1. Let furthermore $\rho \in \{1, \dots, S\}$ be the greatest index with $y_\rho^* > 0$ and $\zeta \in \{1, \dots, S\}$ the greatest index with $z_\zeta + \lambda^* \geq p$, where λ^* is the Lagrange multiplier of the optimal solution described in Section 5.3, i.e.

$$y_\rho^* > 0 \quad \text{and} \quad y_j^* = 0 \qquad \text{for all } j \in \{\rho + 1, \dots, S\}, \tag{5.5.2}$$

$$z_\zeta + \lambda^* \geq p \quad \text{and} \quad z_j + \lambda^* < p \qquad \text{for all } j \in \{\zeta + 1, \dots, S\}. \tag{5.5.3}$$

It is possible that there is no $\zeta \in \{1, \dots, S\}$ with $z_\zeta + \lambda^* \geq p$ at all. In this case we define $\zeta := 0$. If we disregard the case where $\mathbf{z} = (0, \dots, 0)$, which has the solution $\mathbf{y}^* = (0, \dots, 0)$, there must be a $\rho \in \{1, \dots, S\}$ as defined above. The KKT condition (5.3.3) implies

$$1 = \sum_{j=1}^{S} y_j^* = \sum_{j=1}^{\rho} y_j^*, \tag{5.5.4}$$

and with (5.3.10) and (5.3.11) we get

$$1 = \sum_{\substack{j=1 \\ y_j^* = p}}^{\rho} p + \sum_{\substack{j=1 \\ y_j^* < p}}^{\rho} (z_j + \lambda^*) \tag{5.5.5}$$

$$= \sum_{\substack{j=1 \\ z_j + \lambda^* \geq p}}^{\rho} p + \sum_{\substack{j=1 \\ z_j + \lambda^* < p}}^{\rho} (z_j + \lambda^*) \tag{5.5.6}$$

$$= \sum_{j=1}^{\min\{\rho,\zeta\}} p + \sum_{j=\min\{\rho,\zeta\}+1}^{\rho} (z_j + \lambda^*) \tag{5.5.7}$$

$$= \zeta p + \sum_{j=\zeta+1}^{\rho} (z_j + \lambda^*). \tag{5.5.8}$$

The last equation holds since $\zeta \leq \rho$. Otherwise, there would be a $j \in \{1, \dots, S\}$ such that $\rho < j \leq \zeta$, which implies $y_j^* = 0$. The choice of ζ then implies $p \leq z_j + \lambda^*$ and together with (5.3.12) we have $p \leq z_j + \lambda^* \leq 0$. This leads to a contradiction to the choice of p.

Since we do not know the value of λ^*, we need to calculate ζ. If we eliminate ρ from the sum in equation (5.5.8), we get

$$\zeta p + \sum_{j=\zeta+1}^{S} \max\{0, z_j + \lambda^*\} = 1. \tag{5.5.9}$$

This is true since for all $j \in \{\rho + 1, \dots, S\}$ we have $y_j^* = 0$ (see (5.5.2)), which implies $z_j + \lambda^* \leq 0$ (see (5.3.12)).

To motivate the next idea, we recall the geometric interpretation of Algorithm 5.1. We assume that all points are non-positive. We start shifting them to the right and whenever a point reaches p, it will not be shifted any further. Obviously, there must be a last point that reaches p. All of the following points (i.e. smaller values) can still be shifted to the right but will never reach p. We will see that the following idea even makes sense if the assumption of all points being non-positive does not hold.

The previously mentioned shift is determined by the optimal Lagrange multiplier λ^*. We know from (5.5.3) that $p - z_\zeta \leq \lambda^*$ and $\lambda^* < p - z_j$ for all $j \in \{\zeta + 1, \dots, S\}$. If we take the left-hand side of (5.5.9), replace ζ by any $j \in \{0, \dots, S\}$ and λ^* by $p - z_j$, we get the sequence

$$\left(q_j\right)_{j \in \{0,\dots,S\}} := \left(jp + \sum_{i=j+1}^{S} \max\{0, z_i + p - z_j\} \right)_{j \in \{0,\dots,S\}}, \tag{5.5.10}$$

where $z_0 := z_1 + p$ guarantees $z_i + p - z_0 \leq 0$ for all $i \in \{1, \dots, S\}$. In the framework of the geometric interpretation, the latter inequality is equivalent to the assumption of all points being non-positive. The sequence $\left(q_j\right)_{j \in \{0,\dots,S\}}$ is non-decreasing because for every $j \in \{0, \dots, S-1\}$ we have

$$\begin{aligned} q_{j+1} &= jp + \sum_{i=j+1}^{S} \max\{0, z_i + p - z_{j+1}\} + \underbrace{p - \max\{0, z_{j+1} + p - z_{j+1}\}}_{=0} \\ &\overset{(5.5.1)}{\geq} jp + \sum_{i=j+1}^{S} \max\{0, z_i + p - z_j\} \\ &= q_j. \end{aligned} \tag{5.5.11}$$

Note that this property is the justification for using Algorithm 5.2 to find ζ. The assumption $\beta \in (0, 1)$ implies that $q_S \geq 1$, and since $q_0 = 0$, there is a maximal $j \in \{0, \dots, S\}$ satisfying $q_j \leq 1$. This maximal index turns out to be the index of the mentioned *last point that reaches p*, i.e.

$$\zeta = \max\left\{ j \in [0, S] \cap \mathbb{N} \;\middle|\; q_j \leq 1 \right\} \tag{5.5.12}$$

as in Line 2 of Algorithm 5.1, where $\zeta = 0$ means that none of the points *reaches p*.

With this, we are able to calculate λ^*. If $q_\zeta = 1$, equation (5.5.9) is already satisfied and the algorithm terminates with

$$\lambda^* = p - z_\zeta. \tag{5.5.13}$$

Otherwise, we have $p - z_\zeta < \lambda^* < p - z_{\zeta+1}$. Taking ρ into account again, we get $\zeta \neq \rho$ (otherwise (5.5.8) yields $\zeta p = 1$ in contradiction to $q_\zeta < 1$). Together with $\zeta \leq \rho$ we have $\zeta < \rho$. This allows us to rearrange (5.5.8) such that

$$\lambda^* = \frac{1}{\rho - \zeta} \left(1 - \zeta p - \sum_{i=\zeta+1}^{\rho} z_i \right). \tag{5.5.14}$$

The last step is to prove that

$$\rho = \max \left\{ j \in (\zeta, S] \cap \mathbb{N} \;\middle|\; z_j + \frac{1}{j - \zeta} \left(1 - \zeta p - \sum_{i=\zeta+1}^{j} z_i \right) > 0 \right\}. \tag{5.5.15}$$

Let $j \in \{\zeta + 1, \ldots, S\}$. We consider three different cases:

- if $j = \rho$:

$$z_\rho + \frac{1}{\rho - \zeta} \left(1 - \zeta p - \sum_{i=\zeta+1}^{\rho} z_i \right) \overset{(5.5.14)}{=} z_\rho + \lambda^* \overset{(5.3.11)}{>} 0; \tag{5.5.16}$$

- if $\zeta < j < \rho$:

$$\begin{aligned} & z_j + \frac{1}{j - \zeta} \left(1 - \zeta p - \sum_{i=\zeta+1}^{j} z_i \right) \\ & \qquad = \frac{1}{j - \zeta} \left((j - \zeta) z_j + \underbrace{1 - \zeta p - \sum_{i=\zeta+1}^{\rho} z_i}_{\overset{(5.5.14)}{=} (p - \zeta)\lambda^*} + \sum_{i=j+1}^{\rho} z_i \right) \\ & \qquad = \frac{1}{j - \zeta} \left((j - \zeta)(z_j + \lambda^*) + \sum_{i=j+1}^{\rho} (z_i + \lambda^*) \right) \overset{(5.3.11)}{>} 0; \end{aligned} \tag{5.5.17}$$

- if $j > \rho$:

$$\begin{aligned} & z_j + \frac{1}{j - \zeta} \left(1 - \zeta p - \sum_{i=\zeta+1}^{j} z_i \right) \\ & \qquad = \frac{1}{j - \zeta} \left((j - \zeta) z_j + \underbrace{1 - \zeta p - \sum_{i=\zeta+1}^{\rho} z_i}_{\overset{(5.5.14)}{=} (\rho - \zeta)\lambda^*} - \sum_{i=\rho+1}^{j} z_i \right) \\ & \qquad = \frac{1}{j - \zeta} \left((\rho - \zeta) \underbrace{(z_j + \lambda^*)}_{\overset{(5.3.12)}{\leq} 0} + \sum_{i=\rho+1}^{j} \underbrace{(z_j - z_i)}_{\overset{(5.5.1)}{\leq} 0} \right) \leq 0. \end{aligned} \tag{5.5.18}$$

This justifies the choice of ρ and guarantees that we find the optimal λ^* in any possible case. Therefore, we can construct $\mathbf{y}^* = \left(y_1^*, \ldots, y_S^*\right)$ as in (5.4.3), i.e.

$$y_j^* = \max\left\{\min\left\{z_j + \lambda^*, p\right\}, 0\right\} \tag{5.5.19}$$

for all $j \in \{1, \ldots, S\}$. The reasoning of Section 5.3 then shows that $\mathbf{y}^*$ satisfies the KKT conditions and is hence the optimal solution to problem (5.4.2). □

Chapter 6

Applications

In this chapter we consider two exemplary problems that stem from [KS16] and show how our algorithm performs in practice. In the first two sections we develop step size choices and index selection rules that satisfy the assumptions made in Section 4.4. After that, we consider a problem constrained by an elliptic PDE with a discontinuous and uncertain coefficient in the third section, where we prove that the aforementioned assumptions are indeed satisfied for this particular example. Furthermore, we derive a discretized version of Algorithm 4.1 and present numerical results for different parameter settings using the Julia code provided in [Ang22]. The last section deals with a constraint in form of a parabolic equation and also shows how the algorithm behaves for different choices of parameters.

6.1 Choice of Step Sizes

In this section we present a possible way to choose the step sizes in order to satisfy Assumption 4.4.7 (Step-Size-Testing-Relation). We have already seen in the proof of Theorem 4.4.14 that, under the given assumptions, there exists a $C \in (0, \infty)$ such that $\| K'(u_k) \|_{\mathbb{L}(\mathcal{U},\mathcal{V})} \leq C$ P-a.s. for all $k \in \mathbb{N}$. Furthermore, we have seen in (4.4.13) that a sufficient condition for Assumption 4.4.7 (iii) is

$$\sigma\tau_k \| K'(u_k) \|^2_{\mathbb{L}(\mathcal{U},\mathcal{V})} < 1 \quad \text{P-a.s.} \tag{6.1.1}$$

for all $k \in \mathbb{N}_0$. Hence, if we are able to compute C in advance, we could set $\tilde{\gamma}_G = 0$ (i.e. no acceleration) and choose constant step sizes $\tau = \tau_k > 0$ for all $k \in \mathbb{N}_0$ and $\sigma > 0$ such that

$$\sigma\tau < C^{-2}. \tag{6.1.2}$$

In the case of non-constant (primal) step sizes, we could choose some acceleration parameter $\tilde{\gamma}_G \in (0, \gamma_G)$ (where γ_G is the monotonicity factor of ∂G, see Assumption 4.4.3) and initial step sizes $\tau_0 > 0$ and $\sigma > 0$ such that

$$\sigma\tau_0 < C^{-2}. \tag{6.1.3}$$

This inequality is automatically satisfied for all τ_k as well, since Assumption 4.4.7 (i) and (ii) imply that $(\tau_k)_{k\in\mathbb{N}_0}$ is monotonically decreasing. In practice, if we do not know the constant C but have already obtained a discretization[1] of $K'(u_0)^*$ in form of a matrix $K'(\mathbf{u}_0)^* \in \mathbb{R}^{N\times S}$ for the starting vector $\mathbf{u}_0 \in \mathbb{R}^N$, we can obtain an estimate of the norm by using the equation

$$\| K'(u_0)^* \|^2_{\mathbb{L}(\mathcal{V},\mathcal{U})} \approx \| K'(\mathbf{u}_0)^* \|^2_{\mathbb{L}(\mathbb{R}^S,\mathbb{R}^N)} = \lambda_{max}, \tag{6.1.4}$$

[1] We use the adjoint here, since this is the operator we discretize in Section 6.3.4, see (6.3.97).

where $\lambda_{max} \in \mathbb{R}$ is the largest eigenvalue of the symmetric matrix $M := K'(\mathbf{u}_0)K'(\mathbf{u}_0)^* \in \mathbb{R}^{S\times S}$; see [Bjö15, Section 1.1.7]. This eigenvalue can be estimated by the following algorithm, which is called *power method* and stems from [Bjö15, Section 3.3.1].

Algorithm 6.1 (Power Method)

Input: $k_{max} \in \mathbb{N}$, $M \in \mathbb{R}^{S\times S}$, $\mathbf{z}_0 \in \mathbb{R}^S$ with $\| \mathbf{z}_0 \|_2 = 1$

Initialize $k := 0$.

repeat

  $\mathbf{z}'_k := M\mathbf{z}_k$.

  $\mathbf{z}_{k+1} := \frac{\mathbf{z}'_k}{\| \mathbf{z}'_k \|_2}$.

  Update $k \leftarrow k + 1$.

until $k = k_{max}$

Output: $\rho := \mathbf{z}_{k_{max}}^\top M\mathbf{z}_{k_{max}}$

The sequence $(\mathbf{z}_k)_{k\in\mathbb{N}_0}$ generated this way converges to the eigenvector of M corresponding to the largest eigenvalue λ_{max}. The so-called *Rayleigh quotient* ρ approximates this eigenvalue well enough after only a few iterations.

6.2 Index Selection Rules

As we have seen in Section 4.4.2, the main assumption regarding the index selection for CGF is Assumption 4.4.10 (Reduction of CGF), i.e. there exists a constant $M > 0$ such that the sequence $(A_k)_{k\in\mathbb{N}} \subset \mathcal{R}(\mathcal{F}, \mathcal{A})$ satisfies

$$\mathsf{E}\left(\operatorname*{ess\,sup}_{\xi\in\Xi} \chi_{A_k^c}(\xi) \,\middle|\, \mathcal{F}_k \right) \leq Mk^{-3} \quad \mathsf{P}\text{-a.s.} \tag{6.2.1}$$

for all $k \in \mathbb{N}$. In this section, we present two possible index selection rules satisfying this assumption.

First, recall from Section 5.1 that we assumed that the PDE constraint introduced in Section 3.2 has $d \in \mathbb{N}$ uncertain coefficients in $\mathbb{R}$. We used a d-dimensional random vector $\hat{x}\colon \Xi \to \mathbb{R}^d$ to model this uncertainty and obtained a sample $\left\{\xi^j \in \mathbb{R}^d \,\middle|\, j \in \{1,\dots,S\}\right\}$ of $S \in \mathbb{N}$ realizations of this random vector. If we consider the set $\left\{\xi \in \Xi \,\middle|\, \hat{x}(\xi) = \xi_j \text{ for some } j \in \{1,\dots,S\}\right\}$ and assume that, for each $j \in \{1,\dots,S\}$, there exists exactly one $\xi \in \Xi$ satisfying $\hat{x}(\xi) = \xi_j$, then it is clear that Ξ can be identified by the set $\{1,\dots,S\}$. Furthermore, $\mathcal{A}$ can be identified by the power set of $\{1,\dots,S\}$, i.e. $\mathcal{A} = 2^{\{1,\dots,S\}}$. This gives us the justification to approximate every random variable $v \in \mathcal{V} = L^2(\Xi, \mathcal{A}, \mathbb{P})$ by a vector $\mathbf{v} \in \mathbb{R}^S$. The characteristic function $\chi_A\colon \Xi \to \{0,1\}$ for a set $A \in \mathcal{A}$ can then be represented by a vector $\mathbf{x}_A = (x_{A,1},\dots,x_{A,S}) \in \{0,1\}^S$ with

$$x_{A,j} = \chi_A(j) = \begin{cases} 1, & \text{if } j \in A, \\ 0, & \text{if } j \notin A, \end{cases} \tag{6.2.2}$$

for all $j \in \{1, \dots, S\}$. Consequently, the function Π_A, which was defined in (4.1.3) for every element $A \in \mathcal{A}$, then simply maps a vector $\mathbf{v} = (v_1, \dots, v_S) \in \mathbb{R}^S$ to itself, except that all components with index not in A are set to 0, i.e.

$$\Pi_A(\mathbf{v})_j = x_{A,j} v_j = \begin{cases} v_j, & \text{if } j \in A, \\ 0, & \text{if } j \notin A, \end{cases} \tag{6.2.3}$$

for all $j \in \{1, \dots, S\}$. This is the reason why the term *component-wise gradient freezing*, as introduced in Section 4.1, makes sense, since the definitions of $\hat{K}_k$ and $\hat{K}_k(\cdot)^*$ in (4.1.5) and (4.1.6), respectively, imply that only those *components* with indices in A_k are updated. The others are *frozen* in the sense that they just remain at their respective values of the previous iteration.

Now, if we take a look at (6.2.1) again, we see that it can be rewritten as

$$\mathsf{E}\left(\max_{j \in \{1,\dots,S\}} \left(1 - x_{A_k,j}\right) \middle| \mathcal{F}_k \right) \leq Mk^{-3} \quad \text{P-a.s.} \tag{6.2.4}$$

for all $k \in \mathbb{N}$, where $A_k \colon \Theta \to 2^{\{1,\dots,S\}}$ is the randomized index set for all $k \in \mathbb{N}$. Following the reasoning after Assumption 4.4.10, if we assume that the sequence $(A_k)_{k \in \mathbb{N}}$ is independent, we can simplify the left-hand side to get

$$\mathsf{E}\left(\max_{j \in \{1,\dots,S\}} \left(1 - x_{A_k,j}\right) \right) = 1 - \mathsf{E}\left(\min_{j \in \{1,\dots,S\}} x_{A_k,j} \right) \leq Mk^{-3} \tag{6.2.5}$$

for all $k \in \mathbb{N}$. Since $\min_{j \in \{1,\dots,S\}} x_{A_k,j}$ is a random variable taking values in $\{0, 1\}$, we can equivalently write

$$1 - \mathsf{P}\Big(\forall j \in \{1, \dots, S\} : x_{A_k,j} = 1 \Big) = \mathsf{P}\Big(\exists j \in \{1, \dots, S\} : x_{A_k,j} = 0 \Big) \leq Mk^{-3} \tag{6.2.6}$$

for all $k \in \mathbb{N}$. Therefore, the left-hand side is the probability that at least one component is frozen in iteration k.

The following lemma presents a possible strategy one could pursue in order to satisfy (6.2.6).

Lemma 6.2.1 (Index Selection Rule №1)

Let $M > 0$ and $(q_k)_{k \in \mathbb{N}} \subset [0, 1]$ be an arbitrary sequence of probabilities. If, in each iteration $k \in \mathbb{N}$, we independently select every index in $\{1, \dots, S\}$ with probability

$$p_k := \begin{cases} q_k, & \text{if } k < M^{1/3}, \\ \max\left\{ q_k, \left(1 - Mk^{-3}\right)^{1/S} \right\}, & \text{if } k \geq M^{1/3}, \end{cases}$$

then Assumption 4.4.10 is satisfied.

Proof: Let $k \in \mathbb{N}$. Due to the the definition of p_k we have, in any case, that

$$p_k^S \geq 1 - Mk^{-3}. \tag{6.2.7}$$

Since $p_k = \mathsf{P}\left(x_{A_k,j} = 1\right)$ for all $j \in \{1, \dots, S\}$ and the indices are selected independently,

this implies that

$$1 - \mathrm{P}\Big(\forall j \in \{1, \dots, S\} : x_{A_k,j} = 1\Big) = 1 - \prod_{j=1}^{S} \mathrm{P}\left(x_{A_k,j} = 1\right) = 1 - p_k^S \leq Mk^{-3}. \tag{6.2.8}$$

Therefore, inequality (6.2.6) is satisfied and hence Assumption 4.4.10 as well. □

Note that the sequence $(q_k)_{k\in\mathbb{N}} \subset [0,1]$ is not necessary to satisfy Assumption 4.4.10. However, if we would just set $p_k := \max\left\{0, \left(1 - Mk^{-3}\right)^{1/S}\right\}$, then no indices would be selected at all before reaching an iteration with $k > M^{1/3}$.

In the case that we do not want to select every index equally likely, we can use the following lemma, which presents another index selection rule.

Lemma 6.2.2 (Index Selection Rule №2)

Let $(q_k)_{k\in\mathbb{N}} \subset [0,1]$ be a sequence with $q_k \to 1$ as $k \to \infty$. If, in each iteration $k \in \mathbb{N}$, we select $\lceil q_k S \rceil$ indices in $\{1, \dots, S\}$ such that the sequence $(A_k)_{k\in\mathbb{N}}$ is independent, then Assumption 4.4.10 is satisfied.

Proof: Since $q_k \to 1$ as $k \to \infty$, we know that there exists $K \in \mathbb{N}$ such that $\lceil q_k S \rceil = S$ for all $k \geq K$. This means that, if $k \geq K$, we select all indices in iteration k, i.e.

$$\mathrm{P}\Big(\exists j \in \{1, \dots, S\} : x_{A_k,j} = 0\Big) = 0 \tag{6.2.9}$$

for all $k \geq K$. We can therefore choose any $M \geq K^3$ to see that inequality (6.2.6) and hence Assumption 4.4.10 is satisfied. □

Note that the question *how* the desired number of indices is selected in each iteration is not answered in this lemma. Therefore, any approach that maintains the independence of the sequence $(A_k)_{k\in\mathbb{N}}$ is permissible here.

6.3 Elliptic Equation with a Discontinuous Coefficient

In this chapter, we use Algorithm 4.1 to solve a problem constrained by an elliptic partial differential equation that is described in [KS16, Section 6.1]. After the problem's formulation, we show that the assumptions of Section 4.4 are satisfied and how the PDE and the involved functions are discretized. We finish with numerical results for different parameter settings.

6.3.1 Problem Formulation

We consider the optimal control of a linear elliptic PDE with a discontinuous coefficient where the location of the discontinuity is uncertain. Let $\mathcal{U} := L^2(\Omega)$ be the space of control variables with $\Omega := (-1, 1)$, and $\mathcal{V} := L^2_{\mathbb{P}}(\Xi)$ the space of random variables with a separable probability

space $(\Xi, \mathcal{A}, \mathbb{P})$. Let further $\beta \in (0,1)$ be the probability level of CVaR and $\alpha := 10^{-4}$ the weight of the penalty term. The problem described in [KS16, Section 6.1] can be formulated as

$$\min_{u \in \mathcal{U}} \quad \frac{1}{2}\mathrm{CVaR}_\beta \left[\int_\Omega \left(y\left(\hat{\xi}(\cdot), x; u\right) - 1 \right)^2 \, \mathrm{d}x \right] + \frac{\alpha}{2} \int_\Omega u(x)^2 \, \mathrm{d}x, \tag{6.3.1}$$

where $\hat{\xi} \colon \Xi \to \Lambda := [-0.1, 0.1] \times [-0.5, 0.5]$ is a two-dimensional random vector with uniform density $\rho \equiv 5$ and $y := y(u) \in L^2_\rho\left(\Lambda; H^1_0(\Omega)\right)$ solves the weak form of

$$-\partial_x\Big(\epsilon(\lambda, x)\partial_x y(\lambda, x)\Big) = f(\lambda, x) + u(x), \qquad \forall (\lambda, x) \in \Lambda \times \Omega, \tag{6.3.2a}$$

$$y(\lambda, -1) = y(\lambda, 1) = 0, \qquad \forall \lambda \in \Lambda. \tag{6.3.2b}$$

For the sake of simplicity, we write $y(\lambda, x; u)$ or $y(\lambda, x)$ (if the dependency on u is obvious) instead of $y(u)(\lambda)(x)$ for all $(\lambda, x) \in \Lambda \times \Omega$ and $u \in L^2(\Omega)$. The discontinuous coefficient is modeled as

$$\epsilon(\lambda, x) := 0.1\chi_{(-1,\lambda_1]}(x) + 10\chi_{(\lambda_1,1)}(x) \quad \text{for all } (\lambda, x) \in \Lambda \times \Omega. \tag{6.3.3}$$

The function $f \colon \Lambda \times \Omega \to \mathbb{R}$ is defined by $(\lambda, x) \mapsto f(\lambda, x) := \exp\left(-(x - \lambda_2)^2\right)$. In order to bring problem (6.3.1) in the form of Problem 3.1.1, i.e.

$$\min_{u \in \mathcal{U}} \quad \mathrm{CVaR}_\beta(K(u)) + G(u), \tag{6.3.4}$$

we additionally introduce a bounded subset

$$\mathcal{U}_{ad} := \{u \in \mathcal{U} \mid u_a(x) \le u(x) \le u_b(x) \text{ for } \mu\text{-a.e. } x \in \Omega\} \tag{6.3.5}$$

for some $u_a, u_b \in \mathcal{U}$, and define the functions

$$G \colon \mathcal{U} \to \overline{\mathbb{R}}, \qquad u \mapsto G(u) := \frac{\alpha}{2} \int_\Omega u(x)^2 \, \mathrm{d}x + \delta_{\mathcal{U}_{ad}}(u), \tag{6.3.6}$$

$$K \colon \mathcal{U} \to \mathcal{V}, \qquad u \mapsto K(u) := \frac{1}{2} \int_\Omega \left(y\left(\hat{\xi}(\cdot), x; u\right) - 1 \right)^2 \, \mathrm{d}x. \tag{6.3.7}$$

The resulting problem is slightly different from [KS16, Section 6.1], since we restrict the controls to the set $\mathcal{U}_{ad}$. However, this is necessary in order to satisfy Assumption 3.1.2, stating that the domain of G is bounded.

6.3.2 Satisfaction of the Assumptions

The first assumption we need to prove is Assumption 3.1.2 (Properties of G). Due to the specific definition of G, we can even show *strong* convexity here, which directly implies the strong monotonicity of ∂G as required in Assumption 4.4.3.

Lemma 6.3.1 (Properties of G)

The mapping $G \colon \mathcal{U} \to \overline{\mathbb{R}}$ as defined in (6.3.6) has a bounded domain. Moreover, it is proper, lower semi-continuous, and strongly convex with factor α. Consequently, the convex subdifferential $\partial G \colon \mathcal{U} \rightrightarrows \mathcal{U}^*$ is strongly monotone with factor α.

Proof: Clearly, G has a bounded domain $\mathrm{dom}\,(G) = \mathcal{U}_{ad}$. Since $\mathcal{U}_{ad}$ is non-empty and $\|u\|_{\mathcal{U}} < \infty$ for all $u \in \mathcal{U}$, we can easily see that G is proper. Furthermore, due to the continuity of $\|\cdot\|_{\mathcal{U}}$, G is continuous on $\mathcal{U}_{ad}$. Since $\mathcal{U}_{ad}$ is closed, we also have

$$\delta_{\mathcal{U}_{ad}}(u) \leq \liminf_{k\to\infty} \delta_{\mathcal{U}_{ad}}(u_k) \tag{6.3.8}$$

for every sequence $(u_k)_{k\in\mathbb{N}} \subset \mathcal{U}$ with $u_k \rightharpoonup u \in \mathcal{U}$. Hence, $\delta_{\mathcal{U}_{ad}}$ is lower semi-continuous, and due to Lemma 2.2.5 (ii), G is lower semi-continuous as well.

In order to show the strong convexity in the sense of Lemma 2.2.11 (i), let $u, u' \in \mathcal{U}$ and $\theta \in [0,1]$. Without loss of generality, we assume that $u, u' \in \mathcal{U}_{ad}$. Then,

$$\begin{aligned} &G\big(\theta u + (1-\theta)u'\big) - \theta G(u) - (1-\theta)G(u') \\ &\qquad = \frac{\alpha}{2}\left(\|\,\theta u + (1-\theta)u'\,\|_{\mathcal{U}}^2 - \theta\,\|\,u\,\|_{\mathcal{U}}^2 - (1-\theta)\,\|\,u'\,\|_{\mathcal{U}}^2\right) \\ &\qquad = -\frac{\alpha}{2}\theta(1-\theta)\left(\|\,u\,\|_{\mathcal{U}}^2 - 2\,\langle u, u'\rangle_{\mathcal{U}} + \|\,u'\,\|_{\mathcal{U}}^2\right) \\ &\qquad = -\theta(1-\theta)\frac{\alpha}{2}\,\|\,u-u'\,\|_{\mathcal{U}}^2, \end{aligned} \tag{6.3.9}$$

which implies

$$G\big(\theta u + (1-\theta)u'\big) + \theta(1-\theta)\frac{\alpha}{2}\,\|\,u-u'\,\|_{\mathcal{U}}^2 \leq \theta G(u) + (1-\theta)G(u'). \tag{6.3.10}$$

Hence, G is strongly convex with factor α, and Lemma 2.2.11 implies that ∂G is strongly monotone. □

In order to show the remaining assumptions, we first derive the weak form of (6.3.2), since we will need it later. Let $\varphi \in L^2_\rho\left(\Lambda; H^1_0(\Omega)\right)$ be a test function. Multiplying the left-hand side of (6.3.2) with φ, integrating over $x \in \Omega$ and $\lambda \in \Lambda$, and integrating by parts yields

$$\begin{aligned} &\int_\Lambda \rho(\lambda) \int_\Omega -\partial_x\big(\epsilon(\lambda,x)\partial_x y(\lambda,x)\big)\varphi(\lambda,x)\,\mathrm{d}x\,\mathrm{d}\lambda \\ &\qquad = \int_\Lambda \rho(\lambda) \int_\Omega \epsilon(\lambda,x)\partial_x y(\lambda,x)\partial_x\varphi(\lambda,x)\,\mathrm{d}x\,\mathrm{d}\lambda =: a(y,\varphi). \end{aligned} \tag{6.3.11}$$

Multiplying the right-hand side of (6.3.2) with φ and integrating yields

$$\int_\Lambda \rho(\lambda) \int_\Omega \big(f(\lambda,x) + u(x)\big)\varphi(\lambda,x)\,\mathrm{d}x\,\mathrm{d}\lambda =: F(\varphi). \tag{6.3.12}$$

Therefore, in order to solve equation (6.3.2), we need to find $y \in L^2_\rho\left(\Lambda; H^1_0(\Omega)\right)$ such that

$$a(y,\varphi) = F(\varphi) \quad \text{for all } \varphi \in L^2_\rho\left(\Lambda; H^1_0(\Omega)\right). \tag{6.3.13}$$

The same problem, with the expectation instead of CVaR, was studied in [Kou+13, Section 5.1]. The authors showed that, for any $u \in \mathcal{U}$, (6.3.13) has a unique solution $y = y(u) \in L^2_\rho\left(\Lambda; H^1_0(\Omega)\right)$,

and that the mapping $\lambda \mapsto y(u)(\lambda)$ is continuous, hence $y \in L^\infty_\rho\left(\Lambda; H^1_0(\Omega)\right)$. Due to the equivalence in (2.5.6) and the image measure theorem [Kle13, Theorem 4.10], this implies that

$$\hat{y} := y \circ \hat{\xi} \in L^\infty\left(\Xi; H^1_0(\Omega)\right) \subset L^4\left(\Xi; H^1_0(\Omega)\right) =: \mathcal{Y} \tag{6.3.14}$$

is the unique solution to

$$\hat{a}(\hat{y}, \varphi) = \hat{F}(\varphi) \quad \text{for all } \varphi \in L^2\left(\Xi; H^1_0(\Omega)\right), \tag{6.3.15}$$

where

$$\hat{a}(y, \varphi) := \int_\Xi \int_\Omega \epsilon\left(\hat{\xi}(\xi), x\right) \partial_x y(\xi, x) \partial_x \varphi(\xi, x) \, \mathrm{d}x \, \mathrm{d}\mathbb{P}(\xi) \tag{6.3.16}$$

and

$$\hat{F}(\varphi) := \int_\Xi \int_\Omega \left(f\left(\hat{\xi}(\xi), x\right) + u(x)\right) \varphi(\xi, x) \, \mathrm{d}x \, \mathrm{d}\mathbb{P}(\xi) \tag{6.3.17}$$

for all $y, \varphi \in L^2\left(\Xi; H^1_0(\Omega)\right)$. It follows from the definition of the Bochner space (see Definition 2.4.4) that $\int_\Xi \| \hat{y}_u(\xi) \|^4_{H^1(\Omega)} \, \mathrm{d}\mathbb{P}(\xi) < \infty$, and together with Theorem 2.6.3 (Sobolev Embedding) we get

$$\int_\Xi \| \hat{y}_u(\xi) - 1 \|^4_{L^2(\Omega)} \, \mathrm{d}\mathbb{P}(\xi) < \infty, \tag{6.3.18}$$

which implies that $K(u) \in \mathcal{V} = L^2(\Xi)$. Hence, the mapping $K \colon \mathcal{U} \to \mathcal{V}$ is well defined. Furthermore, we can write K as $K = J \circ S$ where $S \colon \mathcal{U} \to \mathcal{Y}$ is the PDE solution operator and $J \colon \mathcal{Y} \to \mathcal{V}$ the Nemytskii operator

$$J(y) = \tilde{J}\big(y(\cdot), \cdot\big) \tag{6.3.19}$$

for all $y \in \mathcal{Y}$ with the tracking-type function from (3.2.4) (with $\overline{y} \equiv 1$), i.e.

$$\tilde{J} \colon H^1_0(\Omega) \times \Xi \to \mathbb{R}, \quad (y, \xi) \mapsto \tilde{J}(y, \xi) := \frac{1}{2} \| y - 1 \|^2_{L^2(\Omega)}. \tag{6.3.20}$$

In the following, we show that the assumptions of Section 4.4.1 are satisfied globally so that we do not need to worry about condition (ii) of Theorem 4.4.14, which requires the iterates to stay within the desired ball $\mathcal{B}$ (where the aforementioned assumptions are satisfied).

Lemma 6.3.2 (Lipschitz Continuity of K)

The mapping $K \colon \mathcal{U} \to \mathcal{V}$ as defined in (6.3.7) is Lipschitz continuous.

Proof: Let $y(u), y(u') \in L^\infty_\rho\left(\Lambda; H^1_0(\Omega)\right)$ solve the weak form of (6.3.2) for some controls $u, u' \in \mathcal{U}$.

First, we show that the PDE solution operator $S \colon \mathcal{U} \to \mathcal{Y}$ is Lipschitz continuous. If we subtract the respective PDEs and define $\tilde{y}(\lambda, x) := y(u)(\lambda)(x) - y(u')(\lambda)(x)$ and $\tilde{u}(x) := u(x) - u'(x)$ for all $\lambda \in \Lambda$ and $x \in \Omega$, we get

$$-\partial_x\Big[\epsilon(\lambda, x) \partial_x \tilde{y}(\lambda, x)\Big] = \tilde{u}(x) \quad \text{for all } (\lambda, x) \in \Lambda \times \Omega. \tag{6.3.21}$$

The Lax-Milgram lemma [Trö10, Lemma 2.2] applied to the weak form of this PDE yields

$$\| \tilde{y}(\lambda, \cdot) \|_{H^1(\Omega)} \leq c \| \tilde{u} \|_{\mathcal{U}} \tag{6.3.22}$$

for all $\lambda \in \Lambda$ and some $c > 0$ independent of $\tilde{u}$ and λ. If we expand $\tilde{y}$ and $\tilde{u}$, we get

$$\| y(u)(\lambda) - y(u')(\lambda) \|_{H^1(\Omega)} \leq c \| u - u' \|_{\mathcal{U}}, \tag{6.3.23}$$

hence the mapping $y(\cdot)(\lambda) \colon \mathcal{U} \to H_0^1(\Omega)$ is Lipschitz continuous with constant c for all $\lambda \in \Lambda$. Together with the notation $\hat{y}_u := y(u) \circ \hat{\xi}$ and $\hat{y}_{u'} := y(u') \circ \hat{\xi}$, this implies that

$$\begin{aligned} \| \hat{y}_u - \hat{y}_{u'} \|_{\mathcal{Y}} &= \left(\int_\Xi \| \hat{y}_u(\xi) - \hat{y}_{u'}(\xi) \|_{H^1(\Omega)}^4 \, \mathrm{d}\mathbb{P}(\xi) \right)^{1/4} \\ &= \left(\int_\Lambda \| y(u)(\lambda) - y(u')(\lambda) \|_{H^1(\Omega)}^4 \, \mathrm{d}\mathbb{P}_{\hat{\xi}}(\lambda) \right)^{1/4} \\ &\leq c \| u - u' \|_{\mathcal{U}} \end{aligned} \tag{6.3.24}$$

where we have used the image measure theorem [Kle13, Theorem 4.10] for the second equality as well as $\mathbb{P}_{\hat{\xi}}(\Lambda) = \mathbb{P}(\Xi) = 1$ for the inequality. Therefore, the PDE solution operator $S \colon \mathcal{U} \to \mathcal{Y}$ with $u \mapsto S(u) := \hat{y}_u = y(u) \circ \hat{\xi}$ is Lipschitz continuous.

Now, let $\xi \in \Xi$ and $M > 0$ such that $\| \hat{y}_u(\xi) \|_{H^1(\Omega)}, \| \hat{y}_{u'}(\xi) \|_{H^1(\Omega)} \leq M$, which is possible since $\hat{y}_u, \hat{y}_{u'} \in L^\infty\left(\Xi; H_0^1(\Omega)\right)$. Together with the definition of J in the proof of Lemma 6.3.3, we can estimate

$$\begin{aligned} | J(\hat{y}_u)(\xi) - J(\hat{y}_{u'})(\xi) | &= \frac{1}{2} \left| \| \hat{y}_u(\xi) - 1 \|_{L^2(\Omega)}^2 - \| \hat{y}_{u'}(\xi) - 1 \|_{L^2(\Omega)}^2 \right| \\ &\leq a(\xi) \| \hat{y}_u(\xi) - \hat{y}_{u'}(\xi) \|_{L^2(\Omega)}, \end{aligned} \tag{6.3.25}$$

where we have used

$$a(\xi) := \frac{1}{2} \left(\| \hat{y}_u(\xi) - 1 \|_{L^2(\Omega)} + \| \hat{y}_{u'}(\xi) - 1 \|_{L^2(\Omega)} \right) \tag{6.3.26}$$

and the reverse triangle inequality. Since $a(\xi) \leq M + \mu(\Omega) =: c'$, we have

$$| J(\hat{y}_u)(\xi) - J(\hat{y}_{u'})(\xi) | \leq c' \| \hat{y}_u(\xi) - \hat{y}_{u'}(\xi) \|_{L^2(\Omega)}. \tag{6.3.27}$$

Inequality (6.3.27) holds for $\mathbb{P}$-almost every $\xi \in \Xi$, hence we can further estimate

$$\begin{aligned} \| J(\hat{y}_u) - J(\hat{y}_{u'}) \|_{\mathcal{V}} &= \left(\int_\Xi | J(\hat{y}_u)(\xi) - J(\hat{y}_{u'})(\xi) |^2 \, \mathrm{d}\mathbb{P}(\xi) \right)^{1/2} \\ &\leq c' \left(\int_\Xi \| \hat{y}_u(\xi) - \hat{y}_{u'}(\xi) \|_{H^1(\Omega)}^2 \, \mathrm{d}\mathbb{P}(\xi) \right)^{1/2} \\ &= c' \| \hat{y}_u - \hat{y}_{u'} \|_{L^2\left(\Xi; H_0^1(\Omega)\right)}, \end{aligned} \tag{6.3.28}$$

where the last inequality is due to (6.3.27) and Theorem 2.6.3 (Sobolev Embedding). If, for every $y \in L^4\left(\Xi; H_0^1(\Omega)\right) = \mathcal{Y}$, we define $f_y(\xi) := \| y(\xi) \|_{H^1(\Omega)}^2$, then Hölder's inequality

[Hin+09, Lemma 1.3] yields

$$\begin{aligned} \| y \|_{L^2(\Xi;H_0^1(\Omega))} &= \left(\left\| f_y \right\|_{L^1(\Xi)} \right)^{1/2} \\ &\leq \left(\left\| f_y \right\|_{L^2(\Xi)} \| 1 \|_{L^2(\Xi)} \right)^{1/2} = \| y \|_{\mathcal{Y}}, \end{aligned} \tag{6.3.29}$$

where we have used that $\| 1 \|_{L^2(\Xi)} = \mathbb{P}(\Xi)^{1/2} = 1$. Applying this together with (6.3.24) to (6.3.28) yields

$$\| J(\hat{y}_u) - J(\hat{y}_{u'}) \|_{\mathcal{V}} \leq c' \| \hat{y}_u - \hat{y}_{u'} \|_{\mathcal{Y}} \leq L \| u - u' \|_{\mathcal{U}} \tag{6.3.30}$$

with $L := cc'$. Since the constants c and c' are independent of u and u', this inequality holds for all $u, u' \in \mathcal{U}$. Due to the definition $S(u) = \hat{y}_u$, we conclude the desired Lipschitz continuity of $K = J \circ S$, i.e.

$$\| K(u) - K(u') \|_{\mathcal{V}} \leq L \| u - u' \|_{\mathcal{U}} \tag{6.3.31}$$

for all $u, u' \in \mathcal{U}$. □

The following lemma shows that K also satisfies Assumption 3.2.4.

Lemma 6.3.3 (Continuous Differentiability of K)

The mapping $K \colon \mathcal{U} \to \mathcal{V}$ as defined in (6.3.7) is continuously Fréchet differentiable.

Proof: We know from (3.2.4) (with $\bar{y} \equiv 1$) that Assumption 3.2.2 (Properties of $\tilde{J}$) is satisfied. Moreover, $y \mapsto \tilde{J}(y, \xi)$ is Fréchet differentiable for all $\xi \in \Xi$ and the derivative

$$\tilde{J}_y \colon H_0^1(\Omega) \times \Xi \to \mathbb{L}\left(H_0^1(\Omega), \mathbb{R}\right), \quad (y, \xi) \mapsto \tilde{J}_y(y, \xi) := \langle y - 1, \cdot \rangle_{L^2(\Omega)} \tag{6.3.32}$$

satisfies Assumption 3.2.2 (i) (Carathéodory). Due to the Cauchy-Schwarz inequality and Theorem 2.6.3 (Sobolev Embedding), the growth condition

$$\left\| \tilde{J}_y(y, \xi) \right\|_{\mathbb{L}(H_0^1(\Omega), \mathbb{R})} = \sup_{\substack{h \in H_0^1(\Omega) \\ \| h \|_{H^1(\Omega)} = 1}} \left| \langle y - 1, h \rangle_{L^2(\Omega)} \right| \leq a(\xi) + \| y \|_{H^1(\Omega)} \tag{6.3.33}$$

holds for all $y \in H_0^1(\Omega)$ and $\xi \in \Xi$ with $a \equiv \mu(\Omega)^{1/2}$. Hence, we can apply Proposition 3.2.3 with $q = r = 4$ to see that $J \colon \mathcal{Y} \to \mathcal{V}$ is continuously Fréchet differentiable.

To show the continuous Fréchet differentiability of $S \colon \mathcal{U} \to \mathcal{Y}$, we define the space

$$\mathcal{X} := L^2\left(\Xi; H_0^1(\Omega)\right) \tag{6.3.34}$$

and note that the mapping $\hat{a} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as defined in (6.3.16) is bounded and $\mathcal{X}$-elliptic in the sense of the Lax-Milgram lemma [Trö10, Lemma 2.2]. Therefore, we know from [Trö10, Lemma 2.35] that the operator $A \colon \mathcal{X} \to \mathcal{X}^*$ defined by

$$\langle Ay, \varphi \rangle_{\mathcal{Y}^*, \mathcal{Y}} = \hat{a}(y, \varphi) \quad \text{for all } y, \varphi \in \mathcal{X} \tag{6.3.35}$$

is linear, continuous, and bijective, and the inverse operator $A^{-1} \colon \mathcal{X}^* \to \mathcal{X}$ is continuous

as well. For the same reason, the operator $B\colon \mathcal{U} \to \mathcal{X}^*$ defined by

$$\langle Bu, \varphi \rangle_{\mathcal{Y}^*,\mathcal{Y}} = \int_\Xi \int_\Omega u(x)\varphi(\xi, x) \, \mathrm{d}x \, \mathrm{d}\mathbb{P}(\xi) \quad \text{for all } u \in \mathcal{U}, \varphi \in \mathcal{X} \tag{6.3.36}$$

is linear and continuous. The PDE, which induces the weak form (6.3.15), can then be expressed as

$$e(y, u) := Ay - C\hat{f} - Bu = 0 \in \mathcal{X}^* \tag{6.3.37}$$

for all $u \in \mathcal{U}$ and $y \in \mathcal{Y}$, where $\hat{f} \in \mathcal{X}$ is defined by $\hat{f}(\xi)(x) := f\left(\hat{\xi}(\xi), x\right)$ and C is some linear operator mapping $\hat{f}$ into $\mathcal{X}^*$. Consequently, the solution operator can be written as

$$u \mapsto S(u) = A^{-1}Bu + A^{-1}C\hat{f} \tag{6.3.38}$$

with the linear and continuous operator $A^{-1}B\colon \mathcal{U} \to \mathcal{X}$, hence $S\colon \mathcal{U} \to \mathcal{X}$ is continuous as well. Furthermore, we already know from the proof of Lemma 6.3.2 that $S\colon \mathcal{U} \to \mathcal{Y} \subset \mathcal{X}$ is Lipschitz continuous. Therefore, the representation in (6.3.38) yields that $S\colon \mathcal{U} \to \mathcal{Y}$ is continuously Fréchet differentiable with derivative

$$u \mapsto S'(u) = A^{-1}B \in \mathbb{L}\left(\mathcal{U}, \mathcal{Y}\right), \tag{6.3.39}$$

which is independent of $u \in \mathcal{U}$.

Together with the continuous Fréchet differentiability of J, we can apply the chain rule [CV20, Theorem 2.7] again, which yields that $K = J \circ S$ is Fréchet differentiable with

$$K'(u) = J'\big(S(u)\big) \circ S'(u) \tag{6.3.40}$$

for all $u \in \mathcal{U}$. Since J and S are continuously Fréchet differentiable, K' is the composition of continuous functions and as such continuous as well. □

In the following lemma, we show that Assumption 4.4.2 is also satisfied globally.

Lemma 6.3.4 (Lipschitz Continuity of K')

The mapping $K'\colon \mathcal{U} \to \mathbb{L}\left(\mathcal{U}, \mathcal{V}\right)$ with $K\colon \mathcal{U} \to \mathcal{V}$ as defined in (6.3.7) is Lipschitz continuous.

Proof: We have already seen in the proof of Lemma 6.3.3 that the PDE solution operator $S\colon \mathcal{U} \to \mathcal{Y}$ is continuously Fréchet differentiable with derivative $u \mapsto S'(u) = A^{-1}B \in \mathbb{L}\left(\mathcal{U}, \mathcal{Y}\right)$ independent of $u \in \mathcal{U}$. The proof of Lemma 6.3.3 further shows that

$$K'(u) = J'\big(S(u)\big) \circ S'(u) = J'\big(S(u)\big) \circ A^{-1}B \tag{6.3.41}$$

for all $u \in \mathcal{U}$.

Therefore, we focus on the Lipschitz continuity of $u \mapsto J'\big(S(u)\big) \in \mathbb{L}\left(\mathcal{Y}, \mathcal{V}\right)$. Due to the definition in (6.3.20), we have

$$J'\big(S(u)\big)y = \langle S(u)(\cdot) - 1, y(\cdot) \rangle_{L^2(\Omega)} \tag{6.3.42}$$

for all $y \in \mathcal{Y}$. Let now $u, u' \in \mathcal{U}$. The definition of the operator norm yields

$$\begin{aligned}
&\left\| J'\big(S(u)\big) - J'\big(S(u')\big) \right\|_{\mathbb{L}(\mathcal{Y},\mathcal{V})} \\
&\qquad = \sup_{\substack{y\in\mathcal{Y} \\ \| y \|_{\mathcal{Y}}=1}} \left\| \langle S(u)(\cdot) - S(u')(\cdot), y(\cdot)\rangle_{L^2(\Omega)} \right\|_{\mathcal{V}} \\
&\qquad \leq \sup_{\substack{y\in\mathcal{Y} \\ \| y \|_{\mathcal{Y}}=1}} \left(\int_\Xi \| S(u)(\xi) - S(u')(\xi) \|^2_{L^2(\Omega)} \, \| y(\xi) \|^2_{L^2(\Omega)} \, \mathrm{d}\mathbb{P}(\xi) \right)^{1/2},
\end{aligned} \tag{6.3.43}$$

where we have used the Cauchy-Schwarz inequality. Using Hölder's inequality, we can further estimate

$$\begin{aligned}
&\sup_{\substack{y\in\mathcal{Y} \\ \| y \|_{\mathcal{Y}}=1}} \left(\int_\Xi \| S(u)(\xi) - S(u')(\xi) \|^2_{L^2(\Omega)} \, \| y(\xi) \|^2_{L^2(\Omega)} \, \mathrm{d}\mathbb{P}(\xi) \right)^{1/2} \\
&\quad \leq \sup_{\substack{y\in\mathcal{Y} \\ \| y \|_{\mathcal{Y}}=1}} \left(\int_\Xi \| S(u)(\xi) - S(u')(\xi) \|^4_{L^2(\Omega)} \, \mathrm{d}\mathbb{P}(\xi) \right)^{1/4} \left(\int_\Xi \| y(\xi) \|^4_{L^2(\Omega)} \, \mathrm{d}\mathbb{P}(\xi) \right)^{1/4} \\
&\quad \leq \sup_{\substack{y\in\mathcal{Y} \\ \| y \|_{\mathcal{Y}}=1}} \| S(u) - S(u') \|_{\mathcal{Y}} \, \| y \|_{\mathcal{Y}} \\
&\quad = \| S(u) - S(u') \|_{\mathcal{Y}},
\end{aligned} \tag{6.3.44}$$

where we have used Theorem 2.6.3 (Sobolev Embedding) for the last inequality. We know from the proof of Lemma 6.3.2 that $S\colon \mathcal{U} \to \mathcal{Y}$ is Lipschitz continuous, hence we can combine (6.3.43) and (6.3.44) to see that there exists some constant $L' > 0$ such that

$$\left\| J'\big(S(u)\big) - J'\big(S(u')\big) \right\|_{\mathbb{L}(\mathcal{Y},\mathcal{V})} \leq L' \, \| u - u' \|_{\mathcal{U}} \tag{6.3.45}$$

for all $u, u' \in \mathcal{U}$. Together with the representation of $S'(u)$ in (6.3.41), we can estimate

$$\begin{aligned}
&\| K'(u) - K'(u') \|_{\mathbb{L}(\mathcal{U},\mathcal{V})} \\
&\qquad = \sup_{\substack{\tilde{u}\in\mathcal{U} \\ \| \tilde{u} \|_{\mathcal{U}}=1}} \left\| \big(K'(u) - K'(u')\big)\tilde{u} \right\|_{\mathcal{V}} \\
&\qquad = \sup_{\substack{\tilde{u}\in\mathcal{U} \\ \| \tilde{u} \|_{\mathcal{U}}=1}} \left\| \Big(J'\big(S(u)\big) - J'\big(S(u')\big)\Big) A^{-1} B \tilde{u} \right\|_{\mathcal{V}} \\
&\qquad \leq \left\| J'\big(S(u)\big) - J'\big(S(u')\big) \right\|_{\mathbb{L}(\mathcal{Y},\mathcal{V})} \left\| A^{-1} B \right\|_{\mathbb{L}(\mathcal{U},\mathcal{Y})} \sup_{\substack{\tilde{u}\in\mathcal{U} \\ \| \tilde{u} \|_{\mathcal{U}}=1}} \| \tilde{u} \|_{\mathcal{U}} \\
&\qquad \leq L \, \| u - u' \|_{\mathcal{U}}
\end{aligned} \tag{6.3.46}$$

for all $u, u' \in \mathcal{U}$, where $L := L' \left\| A^{-1} B \right\|_{\mathbb{L}(\mathcal{U},\mathcal{Y})} > 0$ is independent of u and u'. Hence, $u \mapsto K'(u)$ is Lipschitz continuous. □

The validity of Assumption 4.4.4, which leads to the three-point condition on K in Lemma 4.4.5, is proven in the following lemma.

Lemma 6.3.5

Let $K\colon \mathcal{U} \to \mathcal{V}$ as defined in (6.3.7). For given $\tilde{\gamma}_G \in [0, \alpha)$ there exist a constant $\gamma_{\mathcal{U}} > 0$ such that

$$(\alpha - \tilde{\gamma}_G) \| u - u' \|_{\mathcal{U}}^2 + \left\langle \left(K'(u) - K'(u')\right)(u - u'), v \right\rangle_{\mathcal{V}} \geq \gamma_{\mathcal{U}} \| u - u' \|_{\mathcal{U}}^2$$

for all $u, u' \in \mathcal{U}$ and $v \in \mathcal{V}$.

Proof: Since $\mathcal{Y} \ni y \mapsto \tilde{J}(y, \xi)$ is convex for all $\xi \in \Xi$, this is also true for $\mathcal{Y} \ni y \mapsto J(y)(\xi)$. Furthermore, we know from (6.3.37) that the PDE solution operator $\mathcal{U} \ni u \mapsto S(u) = A^{-1}Bu + A^{-1}C\hat{f}$ is affine. Therefore, K is pointwise convex, i.e. the mapping $\mathcal{U} \ni u \mapsto J\big(S(u)\big)(\xi) = K(u)(\xi)$ is convex for all $\xi \in \Xi$. Now let $F_v\colon \mathcal{U} \to \mathbb{R}$ for all $v \in \mathcal{V}$ be defined by

$$F_v(u) := \langle K(u), v \rangle_{\mathcal{V}} \tag{6.3.47}$$

for all $u \in \mathcal{U}$. Due to the pointwise convexity of K as well as the monotonicity and linearity of the integral, we can estimate

$$\begin{aligned} F_v\big(\theta u + (1-\theta)u'\big) &= \int_\Xi K\big(\theta u + (1-\theta)u'\big)(\xi) v(\xi) \, \mathrm{d}\mathbb{P}(\xi) \\ &\leq \int_\Xi \Big(\theta\big(K(u)(\xi)v(\xi)\big) + (1-\theta)\big(K(u')(\xi)v(\xi)\big)\Big) \, \mathrm{d}\mathbb{P}(\xi) \\ &= \theta F_v(u) + (1-\theta) F_v(u') \end{aligned} \tag{6.3.48}$$

for all $u, u' \in \mathcal{U}$ and $\theta \in [0, 1]$, hence F_v is convex for all $v \in \mathcal{V}$. We know from Section 2.2.4 that, in this case, $u \mapsto \partial F_v(u)$ is monotone. Since F_v is differentiable with $F_v'(u) = \langle K'(u)(\cdot), v \rangle_{\mathcal{V}} \in \mathbb{L}(\mathcal{U}, \mathbb{R})$ for all $u \in \mathcal{U}$, this means that that $u \mapsto \{F_v'(u)\}$ is monotone for all $v \in \mathcal{V}$, i.e.

$$\left\langle F_v'(u) - F_v'(u'), u - u' \right\rangle_{\mathcal{U}^*, \mathcal{U}} = \left\langle \left(K'(u) - K'(u')\right)(u - u'), v \right\rangle_{\mathcal{V}} \geq 0 \tag{6.3.49}$$

for all $u, u' \in \mathcal{U}$ and $v \in \mathcal{V}$. Since $\alpha > \tilde{\gamma}_G$, there exists $\gamma_{\mathcal{U}} > 0$ such that $\alpha - \tilde{\gamma}_G \geq \gamma_{\mathcal{U}}$, which proves the asserted inequality. □

The lemma shows that Assumption 4.4.4 holds globally with $\gamma_G := \alpha$, if $\tilde{\gamma}_G < \gamma_G$. However, this is only a minor restriction since the exact choice of $\tilde{\gamma}_G$ only has an impact on the acceleration due to the strong convexity of G and does not affect the convergence of the algorithm.

6.3.3 PDE Discretization

As in Section 5.1, let $S \in \mathbb{N}$ and $\left\{\xi^j \in \Lambda \,\middle|\, j \in \{1, \ldots, S\}\right\}$ be the set of samples from the random variable $\hat{\xi}$. For all $j \in \{1, \ldots, S\}$ we consider the weak form of (6.3.2) with fixed $\lambda := \xi^j$, which reads

$$a_{\xi^j}(y, \varphi) = F_{\xi^j}(\varphi) \quad \text{for all } \varphi \in H_0^1(\Omega) \tag{6.3.50}$$

for the state $y = y\left(\xi^j\right) \in H_0^1(\Omega)$, where

$$a_{\xi^j}(y, \varphi) := \int_\Omega \epsilon\left(\xi^j, x\right) \partial_x y(x) \partial_x \varphi(x) \, \mathrm{d}x \tag{6.3.51}$$

and

$$F_{\xi^j}(\varphi) := \int_\Omega \left(f\left(\xi^j, x\right) + u(x)\right) \varphi(x) \, \mathrm{d}x. \tag{6.3.52}$$

The PDE (6.3.2) is discretized using the finite element method, following [Cla21], with inner grid points $x_1, \ldots, x_N \in \Omega$ ($N \in \mathbb{N}$) such that

$$-1 =: x_0 < x_1 < \cdots < x_{N+1} := 1. \tag{6.3.53}$$

For the approximation by piecewise linear polynomials, we define

$$\mathcal{S} := \left\{ v \in C^0(\Omega) \;\middle|\; v|_{[x_{i-1}, x_i]} \in P_1 \text{ for all } i \in \{1, \ldots, N+1\} \right\} \tag{6.3.54}$$

as well as

$$\mathcal{S}_{BC} := \left\{ v \in C^0(\Omega) \;\middle|\; v|_{[x_{i-1}, x_i]} \in P_1 \text{ for all } i \in \{1, \ldots, N+1\},\ v(-1) = v(1) = 0 \right\}, \tag{6.3.55}$$

where P_1 is the space of all linear polynomials. The subscript BC stands for *boundary condition* and means that functions in $\mathcal{S}_{BC}$ automatically satisfy the boundary condition (6.3.2b). Obviously, we have $\mathcal{S}_{BC} \subset \mathcal{S}$. A basis $(\varphi_0, \ldots, \varphi_{N+1})$ of $\mathcal{S}$ is given by the linear B-splines

$$\varphi_i \colon \Omega \to \mathbb{R},\ x \mapsto \varphi_i(x) := \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}}, & x \in [x_{i-1}, x_i), \\ \frac{x_{i+1} - x}{x_{i+1} - x_i}, & x \in [x_i, x_{i+1}], \\ 0, & \text{else}, \end{cases} \tag{6.3.56}$$

for all $i \in \{1, \ldots, N\}$, as well as

$$\varphi_0 \colon \Omega \to \mathbb{R},\ x \mapsto \varphi_0(x) := \begin{cases} \frac{x_1 - x}{x_1 - x_0}, & x \in [x_0, x_1], \\ 0, & \text{else}, \end{cases} \tag{6.3.57}$$

and

$$\varphi_{N+1} \colon \Omega \to \mathbb{R},\ x \mapsto \varphi_{N+1}(x) := \begin{cases} \frac{x - x_N}{x_{N+1} - x_N}, & x \in [x_N, x_{N+1}), \\ 0, & \text{else}. \end{cases} \tag{6.3.58}$$

A basis of $\mathcal{S}_{BC}$ is then given by $(\varphi_1, \ldots, \varphi_N)$. Therefore, it is sufficient to require equation (6.3.50) only for these basis functions, i.e. for each scenario $j \in \{1, \ldots, S\}$ we need to find a $y_j \in \mathcal{S}_{BC}$ such that

$$a_{\xi^j}\left(y_j, \varphi_i\right) = F_{\xi^j}(\varphi_i) \quad \text{for all } i \in \{1, \ldots, N\}. \tag{6.3.59}$$

For a given $y_j \in \mathcal{S}_{BC}$, let $\mathbf{y}_j := (y_{1,j}, \ldots, y_{N,j})^\top \in \mathbb{R}^N$ be the vector of the coefficients of y_j with respect to the basis functions, i.e.

$$y_j = \sum_{i=1}^{N} y_{i,j} \varphi_i \quad \text{for all } j \in \{1, \ldots, S\}. \tag{6.3.60}$$

Since $\varphi_i(x_i) = 1$ and $\varphi_k(x_i) = 0$ for all $i, k \in \{1, \dots, N\}$ with $k \neq i$, the coefficients $y_{1,j}, \dots, y_{N,j}$ of the approximate solution $y_j \in \mathcal{S}_{BC}$ to (6.3.59) are, at the grid points, equal to the function values of the true solution $y\left(\xi^j\right) \in H_0^1(\Omega)$ to (6.3.50), i.e.

$$y_j(x_i) = y_{i,j} = y\left(\xi^j\right)(x_i) \quad \text{for all } i \in \{1, \dots, N\}. \tag{6.3.61}$$

The functions $f\colon \Lambda \times \Omega \to \mathbb{R}$ and $u\colon \Omega \to \mathbb{R}$ are discretized the same way, i.e. the vectors

$$(f_{0,j}, \cdots, f_{N+1,j})^\top \in \mathbb{R}^{N+2}, \qquad \text{and} \qquad (u_0, \dots, u_{N+1})^\top \in \mathbb{R}^{N+2} \tag{6.3.62}$$

contain the coefficients with respect to the basis of $\mathcal{S}$ such that

$$f\left(\xi^j, x\right) \approx \sum_{i=0}^{N+1} f_{i,j}\varphi_i(x) \qquad \text{and} \qquad u(x) \approx \sum_{i=0}^{N+1} u_i\varphi_i(x) \tag{6.3.63}$$

for all $j \in \{1, \dots, S\}$ and $x \in \Omega$. Since f and u do not have to satisfy the boundary condition, we need the basis functions φ_0 and φ_{N+1} here, too. However, we will primarily use the coefficients with respect to the basis of $\mathcal{S}_{BC}$ in the following, which is why we define

$$\mathbf{f}_j := (f_{1,j}, \cdots, f_{N,j})^\top \in \mathbb{R}^N \qquad \text{and} \qquad \mathbf{u} := (u_1, \dots, u_N)^\top \in \mathbb{R}^N. \tag{6.3.64}$$

Furthermore, we define the so called *stiffness matrix* $D_j := \left(\left(D_j\right)_{i,k}\right)_{i,k\in\{1,\dots,N\}} \in \mathbb{R}^{N\times N}$ by

$$\left(D_j\right)_{i,k} := a_{\xi^j}(\varphi_i, \varphi_k) \tag{6.3.65}$$

for all $i, k \in \{1, \dots, N\}$ and $j \in \{1, \dots, S\}$, which is self-adjoint due to the symmetry of a_{ξ^j}. Then, equation (6.3.59) is approximated by

$$D_j\mathbf{y}_j = M\left(\mathbf{f}_j + \mathbf{u}\right). \tag{6.3.66}$$

Here, $M := (M_{i,k})_{i,k\in\{1,\dots,N\}} \in \mathbb{R}^{N\times N}$ refers to the *mass matrix*, which is defined by

$$M_{i,k} := \langle\varphi_i, \varphi_k\rangle_{\mathcal{U}} \tag{6.3.67}$$

for all $i, k \in \{1, \dots, N\}$, and which is obviously self-adjoint. Note that, on the right hand side of (6.3.66), we have used that f and u are approximated by piecewise linear functions, i.e.

$$f\left(\xi^j, x\right) + u(x) = \sum_{i=1}^{N} \left(f_{i,j} + u_i\right)\varphi_i(x) \tag{6.3.68}$$

for all $x \in \Omega$ and scenarios $j \in \{1, \dots, S\}$.

We assume that $\epsilon\colon \Lambda \times \Omega \to \mathbb{R}$ is constant between each pair of adjacent grid points, i.e.

$$\epsilon\left(\xi^j, x\right) = \epsilon\left(\xi^j, x_i\right) =: \epsilon_{i,j} \tag{6.3.69}$$

for all $x \in (x_{i-1}, x_i]$ and all $i \in \{1, \dots, N+1\}$. With this assumption, we can easily compute the entries of D_j for the particular basis functions given in (6.3.56). Let $j \in \{1, \dots, S\}$. For every $i \in \{1, \dots, N\}$ we have

$$\begin{aligned}\left(D_j\right)_{i,i} &= \int_\Omega \epsilon\left(\xi^j, x\right)\left(\partial_x \varphi_i(x)\right)^2 \,\mathrm{d}x \\ &= \int_{x_{i-1}}^{x_i} \epsilon\left(\xi^j, x_i\right)\left(\frac{1}{x_i - x_{i-1}}\right)^2 \mathrm{d}x + \int_{x_i}^{x_{i+1}} \epsilon\left(\xi^j, x_{i+1}\right)\left(\frac{1}{x_{i+1} - x_i}\right)^2 \mathrm{d}x \\ &= \frac{\epsilon_{i,j}}{\Delta x_i} + \frac{\epsilon_{i+1,j}}{\Delta x_{i+1}} \end{aligned} \tag{6.3.70}$$

with $\Delta x_i := x_i - x_{i-1}$ for all $i \in \{2, \dots, N\}$. To compute the minor diagonal of D_j, let $i \in \{2, \dots, N\}$. Due to symmetry we have

$$\begin{aligned}\left(D_j\right)_{i,i-1} = \left(D_j\right)_{i-1,i} &= \int_\Omega \epsilon\left(\xi^j, x\right)\partial_x \varphi_i(x)\partial_x \varphi_{i-1}(x)\,\mathrm{d}x \\ &= -\int_{x_{i-1}}^{x_i} \epsilon\left(\xi^j, x_i\right)\left(\frac{1}{x_i - x_{i-1}}\right)^2 \mathrm{d}x \\ &= -\frac{\epsilon_{i,j}}{\Delta x_i}. \end{aligned} \tag{6.3.71}$$

All remaining entries of D_j are zero since the corresponding basis functions have no overlapping region where they are both non-zero. Similarly, we obtain

$$M_{i,i} = \langle \varphi_i, \varphi_i \rangle_{\mathcal{U}} = \frac{\Delta x_i + \Delta x_{i+1}}{3} \qquad \text{for all } i \in \{1, \dots, N\}, \tag{6.3.72}$$

$$M_{i,i-1} = M_{i-1,i} = \langle \varphi_i, \varphi_{i-1} \rangle_{\mathcal{U}} = \frac{\Delta x_i}{6} \qquad \text{for all } i \in \{2, \dots, N\}, \tag{6.3.73}$$

using integration by substitution. Since D_j is symmetric and positive definite for all $j \in \{1, \dots, S\}$, equation (6.3.66) can be written as

$$\mathbf{y}_j = D_j^{-1} M\left(\mathbf{f}_j + \mathbf{u}\right). \tag{6.3.74}$$

6.3.4 Function Discretization

We recall from the description of Algorithm 4.1 that the functions involved in iteration $k \in \mathbb{N}$, which we discretize in the following, are

$$\operatorname{prox}_{\mathrm{CVaR}_\beta^*}^{\Sigma_k} : \mathcal{V} \to \mathcal{V}, \qquad \operatorname{prox}_G^{T_k} : \mathcal{U} \to \mathcal{U}, \tag{6.3.75}$$

$$\hat{K}_k : \mathcal{U} \to \mathcal{V}, \qquad \text{and} \quad \hat{K}_k'(\cdot)^* : \mathcal{U} \to \mathbb{L}\left(\mathcal{V}, \mathcal{U}\right). \tag{6.3.76}$$

As we have seen in (5.4.1), the proximal operator of CVaR_β^* is independent of the scalar step size operator Σ_k and reads

$$\operatorname{prox}_{\mathrm{CVaR}_\beta^*}(\mathbf{z}) = \operatorname{proj}_\Delta(\mathbf{z}), \tag{6.3.77}$$

where

$$\Delta := \left\{ \mathbf{y} \in \mathbb{R}^S \;\middle|\; \mathbf{y}^\top \mathbf{1} = 1 \text{ and } 0 \le y_j \le \frac{1}{(1-\beta)S} \text{ for all } j \in \{1, \dots, S\} \right\} \tag{6.3.78}$$

is the bounded probability simplex.

In order to derive the explicit form of $\operatorname{prox}_G^{T_k}$ for iteration $k \in \mathbb{N}$, we define the function $\phi_k^u \colon \mathcal{U} \to \mathbb{R}$ by

$$z \mapsto \phi_k^u(z) := \frac{1}{2} \| z - u \|_{T_k^{-1}}^2 + \frac{\alpha}{2} \| z \|_{\mathcal{U}}^2 \tag{6.3.79}$$

for a given $u \in \mathcal{U}$ with α as above. Definition 2.2.14 (Weighted Proximal Operator) yields

$$\operatorname{prox}_G^{T_k}(u) = \operatorname*{arg\,min}_{z \in \mathcal{U}} \left(\phi_k^u(z) + \delta_{\mathcal{U}_{ad}}(z) \right) \tag{6.3.80}$$

for all $u \in \mathcal{U}$. Let $\bar{u} \in \mathcal{U}$ be the minimum on the right-hand side, which satisfies the Fermat principle in Theorem 2.2.3, i.e.

$$0 \in \partial \left(\phi_k^u + \delta_{\mathcal{U}_{ad}} \right)(\bar{u}) = \partial \phi_k^u(\bar{u}) + \partial \delta_{\mathcal{U}_{ad}}(\bar{u}), \tag{6.3.81}$$

where we have used the sum rule with equality [CV20, Theorem 4.14]. Since ϕ_k^u is Gâteaux differentiable with derivative

$$\begin{aligned} \mathrm{D}\phi_k^u(z) &= \langle z - u, \cdot \rangle_{T_k^{-1}} + \alpha \langle z, \cdot \rangle_{\mathcal{U}} \\ &= \langle (\mathrm{Id} + \alpha T_k) z - u, \cdot \rangle_{T_k^{-1}} \in \mathcal{U}^* \end{aligned} \tag{6.3.82}$$

for all $z \in \mathcal{U}$, and $T_k = \tau_k \mathrm{Id}$ is positive definite, condition (6.3.81) is satisfied if

$$-u^* \in \partial \delta_{\mathcal{U}_{ad}}(\bar{u}), \tag{6.3.83}$$

where $u^* \in \mathcal{U}$ is the Riesz representation of $\langle (\mathrm{Id} + \alpha T_k)\bar{u} - u, \cdot \rangle_{\mathcal{U}} \in \mathcal{U}^*$. Due to Lemma 2.2.15, this is equivalent to

$$\bar{u} = \operatorname{prox}_{\delta_{\mathcal{U}_{ad}}}^{\Gamma} \left(\bar{u} - \Gamma u^* \right) \tag{6.3.84}$$

for some self-adjoint and strongly monotone operator $\Gamma \in \mathbb{L}(\mathcal{U}, \mathcal{U})$. Especially, if we set $\Gamma := (\mathrm{Id} + \alpha T_k)^{-1} = (1 + \alpha\tau_k)^{-1}\mathrm{Id}$ and use the fact that $\operatorname{prox}_{\delta_{\mathcal{U}_{ad}}}^{\Gamma}$ is the projection onto $\mathcal{U}_{ad}$, we get

$$\bar{u} = \operatorname{proj}_{\mathcal{U}_{ad}} \left((1 + \alpha\tau_k)^{-1} u \right). \tag{6.3.85}$$

Plugging this into (6.3.80) and using the explicit form of the projection in [Ceg12, Section 4.1.6] yields

$$\operatorname{prox}_G^{T_k}(u) = \max \left\{ \min \left\{ (1 + \alpha\tau_k)^{-1} u, u_b \right\}, u_a \right\}, \tag{6.3.86}$$

for all $u \in \mathcal{U}$, i.e.

$$\left(\operatorname{prox}_G^{T_k}(u) \right)(x) = \begin{cases} u_a(x), & \text{if } \left((1 + \alpha\tau_k)^{-1} u \right)(x) < u_a(x), \\ (1 + \alpha\tau_k)^{-1} u(x), & \text{if } u_a(x) \le \left((1 + \alpha\tau_k)^{-1} u \right)(x) \le u_b(x), \\ u_b(x), & \text{if } \left((1 + \alpha\tau_k)^{-1} u \right)(x) > u_b(x), \end{cases} \tag{6.3.87}$$

for all $u \in \mathcal{U}$ and $x \in \Omega$. The i-th component of the discretization $\operatorname{prox}_G^{T_k} \colon \mathbb{R}^N \to \mathbb{R}^N$ with $T_k = \tau_k \mathrm{Id}$ is therefore given by

$$\left(\operatorname{prox}_G^{T_k}(\mathbf{u}) \right)_i = \max \left\{ \min \left\{ (1 + \alpha\tau_k)^{-1} u_i, u_i^b \right\}, u_i^a \right\} \tag{6.3.88}$$

for all $i \in \{1, \dots, N\}$ and $\mathbf{u} = (u_1, \dots, u_N) \in \mathbb{R}^N$, where $u_i^a := u_a(x_i)$ and $u_i^b := u_b(x_i)$.

In order to derive the discretization of $\hat{K}_k'$ and $\hat{K}_k'(\cdot)^*$ as defined in (4.1.5) and (4.1.6), respectively, we recall from (6.2.3) that

$$\Pi_A(\mathbf{v})_j = \begin{cases} v_j, & \text{if } j \in A, \\ 0, & \text{if } j \notin A, \end{cases} \tag{6.3.89}$$

for all $\mathbf{v} = (v_1, \dots, v_S) \in \mathbb{R}^S$, $A \subset \{1, \dots, S\}$, and $j \in \{1, \dots, S\}$. Therefore, we only need to focus on the discretization of K and $K'(\cdot)^*$, since they are independent of the projection onto A.

Let $j \in \{1, \dots, S\}$. We can write the j-th component of K as

$$\begin{aligned} K(u)_j &= \frac{1}{2} \left\| y\left(\xi^j, \cdot; u\right) - 1 \right\|_{\mathcal{U}}^2 \\ &= \frac{1}{2} \int_\Omega \left(y\left(\xi^j, x; u\right) - 1 \right)^2 \, \mathrm{d}x \\ &\approx \frac{1}{2} \sum_{i=1}^{N+1} \int_{x_{i-1}}^{x_i} \left(y_{i-1,j} \varphi_{i-1}(x) + y_{i,j} \varphi_i(x) - 1 \right)^2 \mathrm{d}x \\ &= \sum_{i=1}^{N+1} \frac{1}{2\Delta x_i} \int_{x_{i-1}}^{x_i} \left(\left(y_{i-1,j} - 1\right) x_i - \left(y_{i,j} - 1\right) x_{i-1} + \left(y_{i,j} - y_{i-1,j}\right) x \right)^2 \mathrm{d}x \\ &=: K(\mathbf{u})_j \end{aligned} \tag{6.3.90}$$

for all $u \in \mathcal{U}$ with $y_{0,j} = y_{N+1,j} = 0$ due to the boundary condition (6.3.2b). Using integration by substitution, we get

$$\begin{aligned} K(\mathbf{u})_j &= \frac{1}{6} \sum_{i=1}^{N+1} \frac{\Delta x_i}{y_{i,j} - y_{i-1,j}} \left(\left(y_{i,j} - 1\right)^3 - \left(y_{i-1,j} - 1\right)^3 \right) \\ &= \frac{1}{6} \sum_{i=1}^{N+1} \Delta x_i \Big(\left(y_{i,j} - 1\right)^2 + \left(y_{i,j} - 1\right)\left(y_{i-1,j} - 1\right) + \left(y_{i-1,j} - 1\right)^2 \Big), \end{aligned} \tag{6.3.91}$$

where we have used that $a^3 - b^3 = \left(a^2 + ab + b^2\right)(a - b)$ for all $a, b \in \mathbb{R}$. In order to discretize $K'(\cdot)^*$, let $y'\left(\xi^j; u\right) : \mathcal{U} \to H_0^1$ be the Fréchet derivative of $\mathcal{U} \ni u' \mapsto y\left(u'\right)\left(\xi^j\right)$ in $u \in \mathcal{U}$. We can derive from equation (6.3.74) that

$$\left(y'(\xi^j; u)h\right)(x_i) \approx \left(D_j^{-1} M \mathbf{h}\right)_i =: y'_{i,j} \in \mathbb{R} \tag{6.3.92}$$

for every $i \in \{1, \dots, N\}$ and $h \in \mathcal{U}$ with $\mathbf{h} := \left(h(x_1), \dots, h(x_N)\right)^\top \in \mathbb{R}^N$. Together with (6.3.91), it follows that the j-th component of the Fréchet derivative $K'(u)_j : \mathcal{U} \to \mathbb{R}$ is approximated for all $h \in \mathcal{U}$ by

$$\begin{aligned} K'(u)_j h &\approx \frac{1}{6} \sum_{i=1}^{N+1} \Delta x_i \Big(2\left(y_{i,j} - 1\right) y'_{i,j} + y'_{i,j}\left(y_{i-1,j} - 1\right) + \left(y_{i,j} - 1\right) y'_{i-1,j} + 2\left(y_{i-1,j} - 1\right) y'_{i-1,j} \Big) \\ &= \frac{1}{6} \sum_{i=1}^{N+1} \Delta x_i \Big(\left(2y_{i,j} + y_{i-1,j} - 3\right) y'_{i,j} + \left(2y_{i-1,j} + y_{i,j} - 3\right) y'_{i-1,j} \Big) \\ &=: K'(\mathbf{u})_j \mathbf{h}. \end{aligned} \tag{6.3.93}$$

Using $y'_{0,j} = y'_{N+1,j} = 0$ and rearranging yields

$$K'(\mathbf{u})_j \mathbf{h} = X_j^\top \mathbf{y}'_j, \tag{6.3.94}$$

where $\mathbf{y}'_j := \left(y'_{1,j}, \dots, y'_{N,j}\right)^\top \in \mathbb{R}^N$ and

$$X_j := \frac{1}{6} \begin{pmatrix} \Delta x_1 \left(2y_{1,j} - 3\right) + \Delta x_2 \left(2y_{1,j} + y_{2,j} - 3\right) \\ \Delta x_2 \left(2y_{2,j} + y_{1,j} - 3\right) + \Delta x_3 \left(2y_{2,j} + y_{3,j} - 3\right) \\ \vdots \\ \Delta x_{N-1} \left(2y_{N-1,j} + y_{N-2,j} - 3\right) + \Delta x_N \left(2y_{N-1,j} + y_{N,j} - 3\right) \\ \Delta x_N \left(2y_{N,j} + y_{N-1,j} - 3\right) + \Delta x_{N+1} \left(2y_{N,j} - 3\right) \end{pmatrix} \in \mathbb{R}^N. \tag{6.3.95}$$

It follows from (6.3.92) that

$$K'(\mathbf{u})_j = X_j^\top D_j^{-1} M \tag{6.3.96}$$

for all $j \in \{1, \dots, S\}$, and transposing the matrix with rows $\left\{K'(\mathbf{u})_j \,\middle|\, j \in \{1, \dots, S\}\right\}$ yields the adjoint operator of $K'(\mathbf{u})$, i.e.

$$K'(\mathbf{u})^* = \begin{pmatrix} MD_1^{-1}X_1 & \cdots & MD_S^{-1}X_S \end{pmatrix} \in \mathbb{R}^{N \times S}. \tag{6.3.97}$$

Although we do not need a discretization of the objective function for the algorithm itself, it is still reasonable to consider the progression of the objective function value when investigating the convergence in Section 6.3.6. Without loss of generality, let $\mathbf{z} = (z_1, \dots, z_S) \in \mathbb{R}^S$ be an ordered vector, i.e. $z_1 \leq \cdots \leq z_S$. In order to find an expression for $\mathrm{CVaR}_\beta(\mathbf{z})$, we consider a discrete random variable $\hat{z} \colon \Xi \to \{z_1, \dots, z_S\}$ with uniform distribution. This means that

$$\mathbb{P}\left(\left\{\xi \in \Xi \,\middle|\, \hat{z}(\xi) = z_j\right\}\right) = S^{-1} \tag{6.3.98}$$

for all $j \in \{1, \dots, S\}$, hence the distribution function is

$$F_{\hat{z}}(\alpha) = S^{-1} \sum_{j=1}^{S} \chi_{(-\infty,\alpha]}(z_j) \tag{6.3.99}$$

for all $\alpha \in \mathbb{R}$. Definition 2.5.6 then yields for a probability level $\beta \in (0, 1)$ that

$$\begin{aligned} \mathrm{VaR}_\beta(\hat{z}) &= \min\left\{\alpha \in \mathbb{R} \mid F_{\hat{z}}(\alpha) \geq \beta\right\} \\ &= \min\left\{z \in \{z_1, \dots, z_S\} \,\middle|\, \sum_{j=1}^{S} \chi_{(-\infty,z]}(z_j) \geq \beta S\right\} \\ &= \min\left\{z \in \{z_1, \dots, z_S\} \,\middle|\, \sum_{j=1}^{S} \chi_{(-\infty,z]}(z_j) = \lceil \beta S \rceil\right\} \\ &= z_{\lceil \beta S \rceil} \end{aligned} \tag{6.3.100}$$

due to the ascending order of $\mathbf{z}$. With Proposition 2.5.11 (iv) and $m := \lceil \beta S \rceil$, we get

$$\begin{aligned} \mathrm{CVaR}_\beta(\hat{z}) &= z_m + \frac{1}{1-\beta}\mathbb{E}\left((\hat{z} - z_m)^+\right) \\ &= z_m + \frac{1}{(1-\beta)S}\sum_{j=1}^{S}(z_j - z_m)^+ \\ &= z_m + \frac{1}{(1-\beta)S}\sum_{j=m+1}^{S}(z_j - z_m) \\ &= \frac{m - \beta S}{(1-\beta)S} z_m + \frac{1}{(1-\beta)S}\sum_{j=m+1}^{S} z_j. \end{aligned} \tag{6.3.101}$$

Therefore, we can use

$$\mathrm{CVaR}_\beta(\mathbf{z}) = \frac{\lceil \beta S \rceil - \beta S}{(1-\beta)S} z_{\lceil \beta S \rceil} + \frac{1}{(1-\beta)S}\sum_{j=\lceil \beta S \rceil + 1}^{S} z_j \tag{6.3.102}$$

to compute the Conditional Value-at-Risk of a vector $\mathbf{z} \in \mathbb{R}^S$ that is sorted in ascending order.

The function G can be discretized by calculating the integral in (6.3.6) analogously to (6.3.90) and (6.3.91), yielding

$$\begin{aligned} G(\mathbf{u}) &= \frac{\alpha}{2}\sum_{i=1}^{N+1}\int_{x_{i-1}}^{x_i}\left(u_{i-1}\varphi_{i-1}(x) + u_i\varphi_i(x) - 1\right)^2 \mathrm{d}x \\ &= \frac{\alpha}{6}\sum_{i=1}^{N+1}\Delta x_i\left(u_i^2 + u_i u_{i-1} + u_{i-1}^2\right) \end{aligned} \tag{6.3.103}$$

for every $\mathbf{u} \in \mathcal{U}_{ad} = \left\{z \in \mathbb{R}^N \;\middle|\; u_a(x_i) \le z_i \le u_b(x_i) \text{ for all } i \in \{1, \dots, N\}\right\}$ and $u_0 = u_{N+1} = 0$. Combining (6.3.91), (6.3.102), and (6.3.103) yields the approximation of the objective function in (6.3.4), i.e.

$$F(\mathbf{u}) := \mathrm{CVaR}_\beta(K(\mathbf{u})) + G(\mathbf{u}) \tag{6.3.104}$$

for all $\mathbf{u} \in \tilde{\mathcal{U}}_{ad}$.

As a comparison, we will also consider a risk-neutral modification of problem (6.3.4) in Section 6.3.6, where the Conditional Value-at-Risk is replaced by the expected value. Therefore, we follow the order in Section 5.1 and discretize $\mathbb{E}$ first and determine the proximal operator of the Fenchel conjugate afterwards. Recall from Section 2.5.2 that the sample average approximation yields

$$\mathbb{E}(\mathbf{z}) = \frac{1}{S}\sum_{j=1}^{S} z_j = \frac{1}{S}\langle \mathbf{1}, \mathbf{z}\rangle_{\mathbb{R}^S} \tag{6.3.105}$$

for every $\mathbf{z} \in \mathbb{R}^S$. According to Definition 2.2.6, the Fenchel conjugate of $\mathbb{E}$ is then given by

$$\mathbb{E}^*(\mathbf{z}) = \sup_{\mathbf{v} \in \mathbb{R}^S}\left\langle \mathbf{z} - S^{-1}\mathbf{1}, \mathbf{v}\right\rangle_{\mathbb{R}^S} = \delta_{\{S^{-1}\mathbf{1}\}}(\mathbf{z}) \tag{6.3.106}$$

for every $\mathbf{z} \in \mathbb{R}^S$. Therefore, Definition 2.2.14 and the independence of the step size σ (see (6.3.77)) yield

$$\mathrm{prox}_{\mathbb{E}^*}(\mathbf{z}) = \mathrm{proj}_{\{S^{-1}\mathbf{1}\}}(\mathbf{z}) \tag{6.3.107}$$

for all $\mathbf{z} \in \mathbb{R}^S$. As we have mentioned in Section 2.5.4, one could also view the expected value as the Conditional Value-at-Risk for the probability level $\beta = 0$. This is compliant with the form of the proximal operator of CVaR_β^* presented in (6.3.77), since, in this case, the probability simplex reads

$$\Delta = \left\{ \mathbf{y} \in \mathbb{R}^S \;\middle|\; \mathbf{y}^\top \mathbf{1} = 1 \text{ and } 0 \leq y_j \leq S^{-1} \; \forall j \in \{1, \ldots, S\} \right\} = \left\{ S^{-1}\mathbf{1} \right\}. \tag{6.3.108}$$

Therefore, we have

$$\mathrm{prox}_{\mathrm{CVaR}_0^*}(\mathbf{z}) = \mathrm{proj}_{\{S^{-1}\mathbf{1}\}}(\mathbf{z}) = \mathrm{prox}_{\mathbb{E}^*}(\mathbf{z}) \tag{6.3.109}$$

for all $\mathbf{z} \in \mathbb{R}^S$.

6.3.5 Discrete Algorithm

If we combine the results of Section 6.3.3 and Section 6.3.4, we can formulate a discrete version of Algorithm 4.1. The primal and dual iterates are denoted by $\mathbf{u}_k = \left(u_1^k, \ldots, u_N^k\right) \in \mathbb{R}^N$ and $\mathbf{v}_k = \left(v_1^k, \ldots, v_S^k\right) \in \mathbb{R}^N$, respectively. The required inputs are the stopping criterion tolerance $\varepsilon > 0$, the starting vectors $\mathbf{u}_0 \in \mathbb{R}^N$, $\mathbf{v}_0 \in \mathbb{R}^S$ and the initial step size operators $T_0 \in \mathbb{R}^{N \times N}$, $\Sigma_0 \in \mathbb{R}^{S \times S}$. Furthermore, following Section 6.2, we denote by $A_k \in \mathcal{A} = 2^{\{1,\ldots,S\}}$ the set of the indices in iteration $k \in \mathbb{N}_0$, for which the gradients are *not* frozen. However, there is no freezing in the first iteration due to the choice of $A_0 := \{1, \ldots, S\}$.

Algorithm 6.2 (Discrete Formulation)

Initialize $k := 0$.
repeat
  if $k \geq 1$ **then**
    Randomly select $A_k \in \mathcal{A}$.
  end
  Compute $\mathbf{y}_{j,k} := D_j^{-1} M \left(\mathbf{f}_j + \mathbf{u}_k\right)$ for every $j \in A_k$.
  Compute $K(\mathbf{u}_k)_j$ by (6.3.91) for every $j \in A_k$.
  $\mathbf{v}_{k+1} := \mathrm{prox}_{\mathrm{CVaR}_\beta^*}\left(\mathbf{v}_k + \Sigma_k K(\mathbf{u}_k)\right)$
  $\bar{\mathbf{v}}_{k+1} := 2\mathbf{v}_{k+1} - \mathbf{v}_k$
  Compute column j of $K'(\mathbf{u}_k)^*$ by (6.3.97) for every $j \in A_k$.
  $\mathbf{u}_{k+1} := \mathrm{prox}_G^{T_k}\left(\mathbf{u}_k - T_k K'(\mathbf{u}_k)^* \bar{\mathbf{v}}_{k+1}\right)$
  Determine step size operators $T_{k+1} \in \mathbb{R}^{N \times N}$ and $\Sigma_{k+1} \in \mathbb{R}^{S \times S}$.
  Update $k \leftarrow k + 1$.
until $\| u_k - u_{k-1} \|_2 < \varepsilon$ **and** $\| v_k - v_{k-1} \|_2 < \varepsilon$

An implementation of Algorithm 6.2 in Julia can be found in [Ang22, Algorithm.jl].

6.3.6 Numerical Results

In the following, we present some numerical results using the Julia code provided in [Ang22, EEDC.jl]. Throughout this section, we define the bounds within $\mathcal{U}_{ad}$ by -10 and 10, respectively, i.e.

$$\mathcal{U}_{ad} = \left\{ \mathbf{z} \in \mathbb{R}^N \;\middle|\; -10 \leq z_i \leq 10 \text{ for all } i \in \{1, \dots, N\} \right\}, \tag{6.3.110}$$

and use the starting vectors

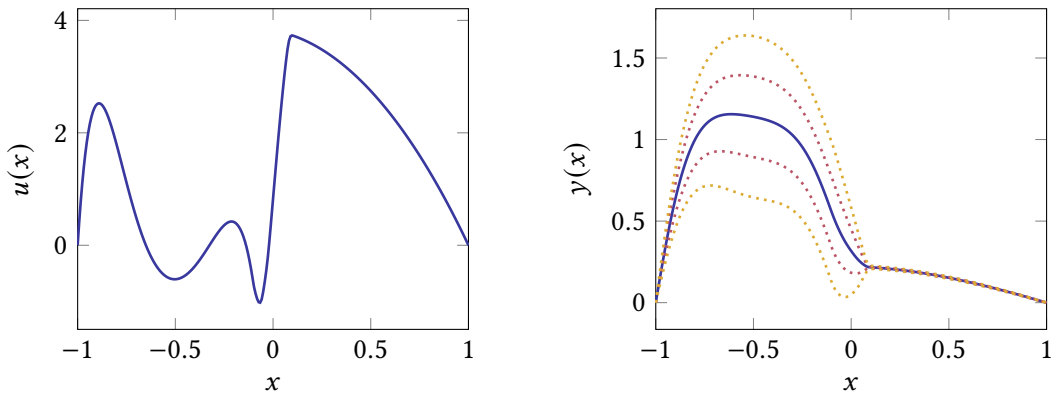
$$\mathbf{u}_0 = (0, \dots, 0) \in \mathbb{R}^N \quad \text{and} \quad \mathbf{v}_0 = (0, \dots, 0) \in \mathbb{R}^S. \tag{6.3.111}$$

For a given dual step size $\sigma > 0$, the initial primal step size is defined by

$$\tau_0 := 0.99 \, (\sigma \rho)^{-1}, \tag{6.3.112}$$

where ρ is the Rayleigh quotient determined by 5 iterations of Algorithm 6.1 with the matrix $K'(\mathbf{u}_0)K'(\mathbf{u}_0)^*$ and the starting vector $\mathbf{z}_0 := S^{-2}\mathbf{1}$. Note that, although G is strongly monotone, we can not accelerate the algorithm by choosing $\tilde{\gamma}_G > 0$ for this particular problem in practice. However, this allows us to set $\tilde{\gamma}_G := 0$ in Lemma 6.3.5, which resolves the conflict with condition (viii) of Theorem 4.4.14 as discussed at the end of Section 4.4.3.

To give a first impression of the problem, we first present in Figure 6.1 the optimal control and state of problem (6.3.1) for a probability level of $\beta = 0.9$, $S = 1{,}000$ scenarios, and $N = 256$ equidistant grid points. The solution was computed using constant scalar step sizes with $\sigma = 0.01$ and a stopping criterion tolerance of $\varepsilon = 10^{-10}$. The big jump in the control around $x = 0$ can be explained by the change of the PDE coefficient $\epsilon(x)$ from 0.1 to 10, which appears at some x within $[-0.1, 0.1]$.

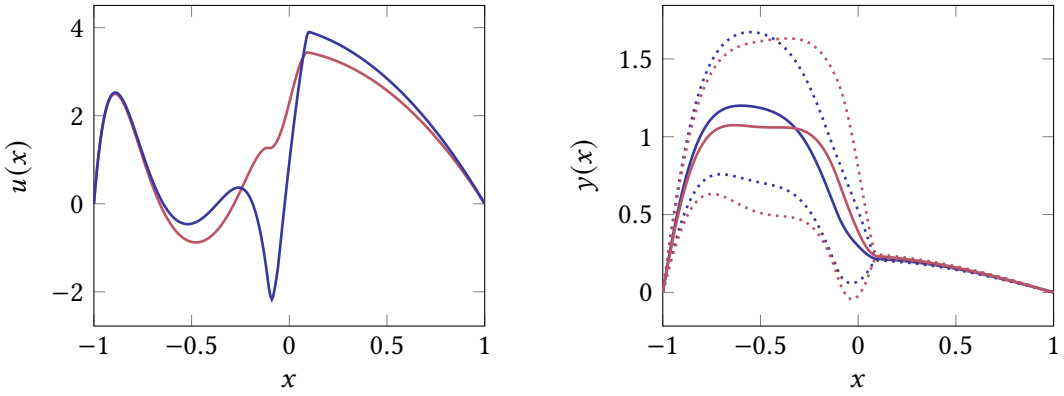


(a) Optimal control

(b) Optimal state (mean ± one and two standard deviations)

Figure 6.1: Example of optimal control and state for $\beta = 0.9$

Figure 6.2 illustrates how the use of the Conditional Value-at-Risk as a risk measure affects the solution. The blue graph shows the risk-averse control and state for a probability level of $\beta = 0.99$ and all other parameters as in Figure 6.1. The red graph shows the risk-neutral case where the Conditional Value-at-Risk is replaced by the expected value as described in (6.3.107). One can see that, especially in the region where the discontinuity of the coefficient appears, the risk-averse optimal state has a smaller deviation from its mean than the risk-neutral state. Apparently, this is achieved by the sharp dip of the risk-averse optimal control around $x = 0$.

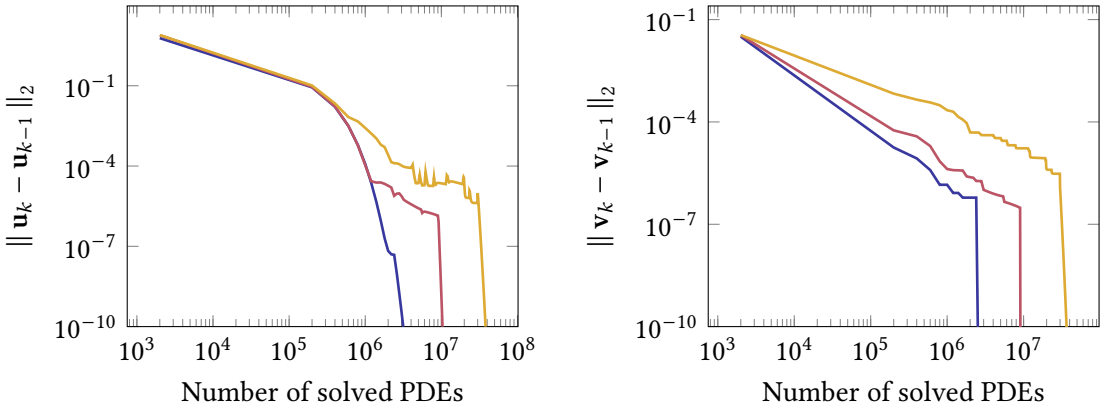


(a) Optimal control

(b) Optimal state (mean (solid) ± two (dotted) standard deviations)

Figure 6.2: Comparison of risk-averse ($\beta = 0.99$) and risk-neutral case

In the following we examine the convergence behavior for different choices of the probability level β. Figure 6.3 shows how the norm of the difference of successive iterates changes in relation to the number of solved PDEs. This norm is particularly interesting because it is the norm of the residual of the fixed-point iteration (4.1.2), which is the basis of our algorithm. The results were computed using $\sigma = 0.01$ with constant scalar step sizes, $S = 1{,}000$ scenarios, $N = 256$ grid points, and a stopping criterion tolerance of $\varepsilon = 10^{-10}$. Apparently, in order to reach the required tolerance, a higher probability level leads to more PDE solves. Interestingly, this behavior is not visible in Figure 6.4, where the function values approach their asymptotes around the same number of PDE solves.



(a) Convergence of primal variable

(b) Convergence of dual variable

Figure 6.3: Convergence behavior for $\beta = 0.01$, $\beta = 0.5$, and $\beta = 0.99$

Before investigating the convergence behavior for different choices of parameters, we briefly show how the use of Algorithm 5.2 and Algorithm 5.3 can decrease the computation time in practice. We already mentioned in Section 5.4 that simply checking the condition

$$jp + \sum_{i=j+1}^{S} \max\{0, u_i + p - u_j\} \leq 1 \tag{6.3.113}$$

for all $j \in \{0, \dots, S\}$ (starting with $j = 0$ until it is not satisfied any more) in order to compute

the index ζ in Algorithm 5.1 can be computationally demanding in some cases. Here, the vector $(u_1, \dots, u_S) \in \mathbb{R}^S$ does not refer to the control, but to the sorted argument of $\operatorname{prox}_{\mathrm{CVaR}^*_\beta}$, i.e. $u_1 \geq \cdots \geq u_S$. While in the most examples we consider in this section, there is no significant difference in terms of computation time, we can deliberately choose a combination of parameters to make the performance difference visible. Solving the problem with parameters $N = 256$, $S = 1{,}000$, $\beta = 0.01$, and $\varepsilon = 10^{-10}$ takes about 89 seconds, if we use the straightforward approach explained above (using the Julia function `findfirst`) and constant scalar step sizes with $\sigma = 0.01$. In contrast, using Algorithm 5.2 and Algorithm 5.3 results in a computation time of about 76 seconds, which is about 15% less. An explanation for the difference in this specific case could be that the combination of S and β leads to an upper bound of $p := ((1-\beta)S)^{-1}$ in the definition of the probability simplex (see (6.3.78)) that is close to 0. Therefore, condition (6.3.113) is violated only for large indices, which means that the repeated checking for the condition takes long when starting with $j = 0$.

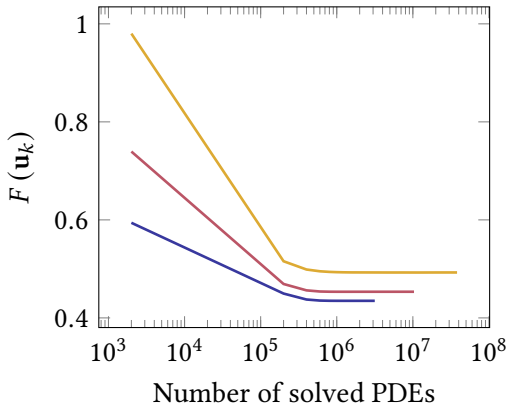


Figure 6.4: Convergence behavior for $\beta = 0.01$, $\beta = 0.5$, and $\beta = 0.99$

Since the core of this work lies in the component-wise gradient freezing (CGF), we show in the following how this method can reduce the number of PDE solves in Line 6 and Line 10 of Algorithm 6.2. For the subsequent examples, we also use $N = 256$ grid points, $S = 1{,}000$ scenarios, $\varepsilon = 10^{-10}$, and $\sigma = 0.01$ with constant scalar step sizes as above. First, we note that one way to save PDE solves could be to modify Line 10 of Algorithm 6.2 such that only the columns of $K'(\mathbf{u}_k)^*$ with index $j \in A_k \setminus B_k$, where

$$B_k := \left\{ j \in \{1, \dots, S\} \;\middle|\; (\bar{\mathbf{v}}_{k+1})_j = 0 \right\}, \tag{6.3.114}$$

are computed. The idea is that we do not need these columns in Line 11 because they are multiplied with 0 anyway. However, the algorithm changed this way no longer fits into our theoretical framework, since it uses two different index sets for CGF, which are also not stochastically independent. However, we present in Table 6.1 how many iterations and PDE solves are required to reach the stopping criterion tolerance in dependence of the probability level β. The number of iterations is not affected by the use of B_k. One can see that there is almost no saving in the number of solved PDEs if β is close to 0. However, the larger β becomes, the more PDE solves we can save. In contrast, CGF is able to save more PDE solves if β is close to 0, as we will see later. The reason for this effect becomes clear if we recall the definition of the bounded probability simplex given in Lemma 5.1.1, which is

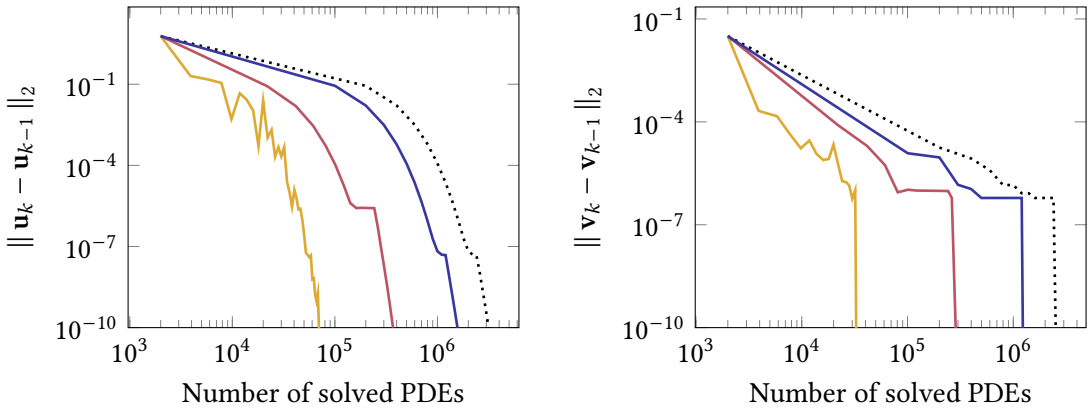
$$\Delta := \left\{ \mathbf{y} \in \mathbb{R}^S \;\middle|\; \mathbf{y}^\top \mathbf{1} = 1 \text{ and } 0 \leq y_j \leq \frac{1}{(1-\beta)S} \text{ for all } j \in \{1, \dots, S\} \right\}. \tag{6.3.115}$$

If β is close to 0, then the upper bound $\frac{1}{(1-\beta)S}$ is close to $\frac{1}{S}$, which means that almost all coordinates of the projection's result in Line 8 of Algorithm 6.2 need to be at this bound in order to satisfy the equality in the definition of Δ. As a consequence, there cannot be many coordinates $j \in \{1, \dots, S\}$ such that $(\bar{\mathbf{v}}_{k+1})_j = 2\,(v_{k+1})_j - (v_k)_j = 0$, since $v_{k+1}, v_k \in \Delta$ for all $k \in \mathbb{N}$. Therefore, the cardinality of B_k is often close to 0, hence the effect of disregarding indices in B_k is very small. On the other hand, one can observe that the cardinality of B_k is close to S if β is close to 1. This is the reason why very few PDE solves are needed in Line 10 in this case and the savings by using B_k are almost 50% for $\beta = 0.99$.

	# solved PDEs		
β	without B_k	with B_k	savings
0.01	$3.16 \cdot 10^6$	$3.15 \cdot 10^6$	0.47%
0.1	$1.23 \cdot 10^7$	$1.17 \cdot 10^7$	5.13%
0.5	$1.04 \cdot 10^7$	$7.78 \cdot 10^6$	25.16%
0.9	$1.95 \cdot 10^7$	$1.07 \cdot 10^7$	44.93%
0.99	$3.81 \cdot 10^7$	$1.93 \cdot 10^7$	49.26%

Table 6.1: Number of solved PDEs with/without B_k for different values of β

In the following, we investigate how this compares to the savings we achieve by the use of CGF. We start with index selection rule №1 as described in Lemma 6.2.1 with a constant sequence of probabilities $q_k := q$ for some $q \in [0, 1]$ and all $k \in \mathbb{N}$, and $M := 10^{20}$. Note that M is deliberately chosen so large that the stopping criterion is likely to be reached in an iteration $k < M^{1/3}$ and we can observe the maximal possible effect of using CGF[2]. In Figure 6.5 we show the convergence behavior for a probability level of $\beta = 0.01$ and different values of q.



(a) Convergence of primal variable

(b) Convergence of dual variable

Figure 6.5: Convergence behavior for $\beta = 0.01$ with CGF and index selection rule №1 for $q = 0.5$, $q = 0.1$, $q = 0.01$, and without CGF (dotted)

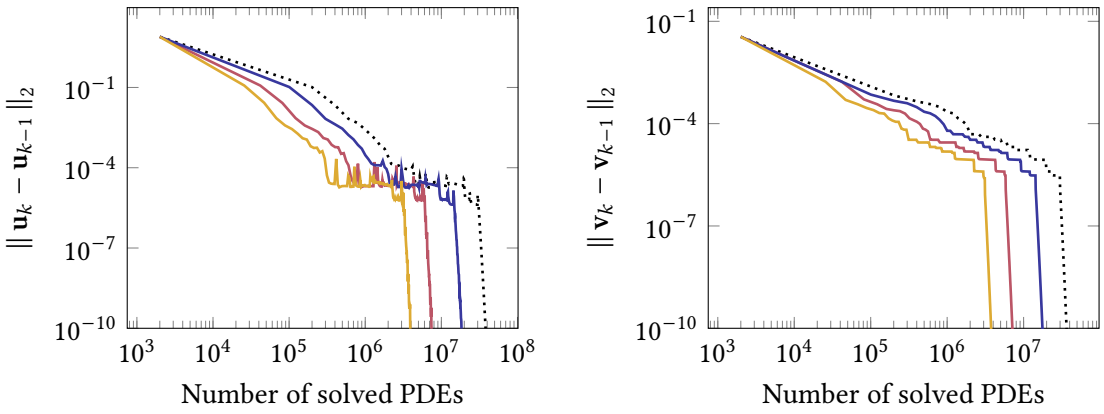
[2]Recall that, if $k \geq M^{1/3}$, the probability for choosing an index is $\max\left\{q_k, \left(1 - Mk^{-3}\right)^{1/S}\right\}$. Hence, the expected number of frozen indices gradually decreases beyond the iteration threshold of $M^{1/3}$.

The stopping criterion is satisfied in all presented examples in an iteration $k < 19{,}000$, hence the probability p is in fact used throughout the whole algorithm. One can see that, even for $q = 0.01$, i.e. if we expect to select 10 indices in each iteration, the algorithm converges. Table 6.2 shows how the choice of q affects the percentage of required PDE solves.

	# solved PDEs	
q	absolute	relative
-	$3.16 \cdot 10^6$	100%
0.5	$1.57 \cdot 10^6$	49.68%
0.2	$5.78 \cdot 10^5$	18.28%
0.1	$3.69 \cdot 10^5$	11.67%
0.05	$1.89 \cdot 10^5$	5.97%
0.01	$6.99 \cdot 10^4$	2.21%

Table 6.2: Number of solved PDEs for $\beta = 0.01$ with CGF and index selection rule №1 in relation to the choice of q

The first row contains the number of solved PDEs without CGF. Apparently, the number of required PDE solves decreases with q and is actually quite accurate q times the number of PDE solves without CGF. Only for $q = 0.01$, the number of selected indices is presumably so low that it needs to be compensated by a higher number of iterations (and hence by a higher number of PDE solves). However, this choice still results in a saving of almost 98% in the number of solved PDEs.



(a) Convergence of primal variable

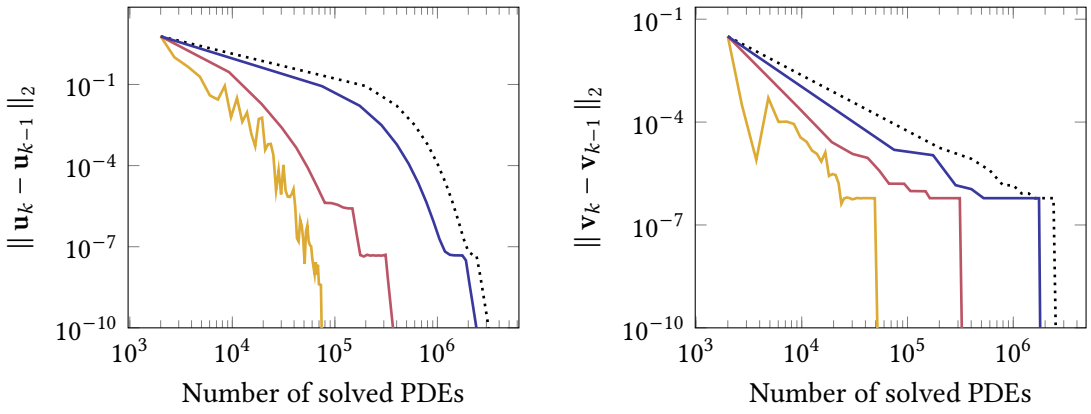
(b) Convergence of dual variable

Figure 6.6: Convergence behavior for $\beta = 0.99$ with CGF and index selection rule №1 for $q = 0.5$, $q = 0.2$, $q = 0.2$ and B_k, and without CGF (dotted)

As a comparison, we also show the corresponding results for $\beta = 0.99$ in Figure 6.6. As it turns out in this case, we can not choose q as small as above and still obtain convergence. This can be explained by the observation we made earlier, stating that $(\bar{\mathbf{v}}_{k+1})_j = 0$ for many $j \in \{1, \dots, S\}$ and $k \in \mathbb{N}$ if β is close to 1. In this case, only a few columns of $K'(\mathbf{u}_k)^*$ are multiplied with a non-zero coordinate in Line 11, and thus the probability q cannot be chosen too small in order to still achieve sufficient progress in the primal iterates. Therefore, we only present the graphs for

$q \in \{0.5, 0.2\}$ and additionally show how even more PDE solves can be saved for $q = 0.2$ if we also use the set B_k from (6.3.114). It turns out that the number of solved PDEs can be reduced by 80.5% if we use $q = 0.2$, and by 89.7% if we incorporate B_k as well.

In order to investigate the effects of using index selection rule №2 as described in Lemma 6.2.2, we define the sequence $(q_k)_{k\in\mathbb{N}}$ by $q_k := \log(k)a^{-1}$ for a fixed $a > 0$ and all $k \in \mathbb{N}$, and use the Julia function `StatsBase.sample` to sample a number of min $\{S, \max\{1, \lceil q_k S \rceil\}\}$ indices from $\{1, \dots, S\}$ (without replacement) in each iteration $k \in \mathbb{N}$. Figure 6.7 shows the convergence behavior for $\beta = 0.01$ and $a \in \{10, 10^2, 10^3\}$, if we use equal sample weights (i.e. every index is equally likely to be selected). The corresponding number of samples per iteration for this example is depicted in Figure 6.8.



(a) Convergence of primal variable

(b) Convergence of dual variable

Figure 6.7: Convergence behavior for $\beta = 0.01$ with CGF and index selection rule №2 for $a = 10$, $a = 10^2$, $a = 10^3$, and without CGF (dotted)

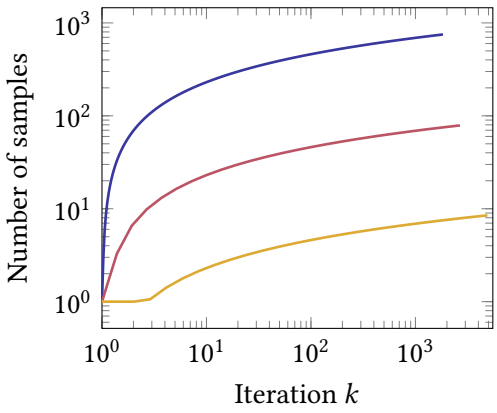


Figure 6.8: Number of samples per iteration for $\beta = 0.01$ and index selection rule №2 for $a = 10$, $a = 10^2$, and $a = 10^3$

We can see that this index selection rule can lead to similar savings in the number of solved PDEs as index selection rule №1. Although a direct comparison of Figure 6.5 and Figure 6.7 is not meaningful since the results depend strongly on how the sequence $(q_k)_{k\in\mathbb{N}}$ is defined, one can still notice two advantages of the second rule. First, we can exactly control the number of selected indices per iteration and not just the *average* number, as in rule №1. Second, rule №2 does not require the indices to be selected equally likely, which is why information about the scenarios could be used to define individual probabilities, e.g. in an adaptive approach.

6.4 Steady Burgers' Equation

In this chapter, we use Algorithm 4.1 to solve a problem constrained by the steady Burgers' equation presented in [KS16, Section 6.2]. This equation was named after the Dutch physicist Johannes Martinus Burgers and is a one-dimensional nonlinear model for convection-diffusion phenomena [Vol00]. Unlike the previous section, we do not prove that the assumptions are satisfied for this particular problem, but we focus on the discretization and numerical results.

6.4.1 Problem Formulation

We consider the optimal control of the steady Burgers' equation with uncertain coefficients. Let $\mathcal{U} := L^2(\Omega)$ be the space of control variables with $\Omega := (0,1)$ and $\mathcal{V} := L^2_{\mathbb{P}}(\Xi)$ the space of random variables with a separable probability space $(\Xi, \mathcal{A}, \mathbb{P})$. Let further $\beta \in (0,1)$ be the probability level of CVaR and $\alpha := 10^{-3}$ the weight of the penalty term. The problem described in [KS16, Section 6.2] can be formulated as

$$\min_{u \in \mathcal{U}} \quad \frac{1}{2}\mathrm{CVaR}_\beta\left[\int_\Omega \left(y\left(\hat{\xi}(\cdot), x; u\right) - 1\right)^2 \,\mathrm{d}x\right] + \frac{\alpha}{2}\int_\Omega u(x)^2 \,\mathrm{d}x, \tag{6.4.1}$$

where $\hat{\xi}\colon \Xi \to \Lambda := [-1,1]^4$ is a four-dimensional random vector with uniform density $\rho \equiv 2^{-4}$, $(\Sigma, \mathcal{F}, \mathbb{P})$ is a probability space, and $y := y(u) \in L^2_\rho\left(\Lambda; H^1(\Omega)\right)$ solves the weak form of

$$-\nu(\lambda)\partial_{xx} y(\lambda, x) + y(\lambda, x)\partial_x y(\lambda, x) = f(\lambda, x) + u(x), \qquad \forall (\lambda, x) \in \Lambda \times \Omega, \tag{6.4.2a}$$

$$y(\lambda, 0) = d_0(\lambda), \quad y(\lambda, 1) = d_1(\lambda), \qquad \forall \lambda \in \Lambda. \tag{6.4.2b}$$

The random coefficient ν (also referred to as *viscosity parameter*), the values of f, and the boundaries d_0 and d_1 are given by

$$\nu(\lambda) := 10^{\lambda_1 - 2}, \quad f(\lambda, x) := \frac{\lambda_2}{100}, \quad d_0(\lambda) := 1 + \frac{\lambda_3}{1000}, \quad \text{and} \quad d_1(\lambda) := \frac{\lambda_4}{1000}, \tag{6.4.3}$$

for all $\lambda \in \Lambda$ and $x \in \Omega$. Apart from the domain Ω, which has changed from $(-1,1)$ to $(0,1)$, the objective function takes the same form as in Section 6.3. Therefore, we can adopt the notation from (6.3.6) and (6.3.7) and write problem (6.4.1) as

$$\min_{u \in \mathcal{U}} \quad \mathrm{CVaR}_\beta(K(u)) + G(u). \tag{6.4.4}$$

Since this definition also includes the set $\mathcal{U}_{ad}$ as defined in (6.3.5), the resulting problem is slightly different from [KS16, Section 6.2].

6.4.2 PDE Discretization

The PDE (6.4.2) is discretized using a finite difference method (FDM, see [Del15, Section 2.4]) and $N \in \mathbb{N}$ inner grid points $x_1, \ldots, x_N \in \Omega = (0,1)$, i.e.

$$0 =: x_0 < x_1 < \cdots < x_N < x_{N+1} := 1. \tag{6.4.5}$$

We assume that the grid is uniform and denote the distance between two adjacent points by Δx, i.e. $\Delta x := x_i - x_{i-1}$ for all $i \in \{1, \ldots, N+1\}$. Since Ω is one-dimensional, it follows from the

Sobolev Embedding Theorem [Hin+09, Theorem 1.14] that $H^1(\Omega) \hookrightarrow\hookrightarrow C^0(\overline{\Omega})$. Therefore, the solution to (6.4.2) is continuous and can hence be approximated pointwise.

Let $S \in \mathbb{N}$ and $\{\xi^j \in \Lambda \mid j \in \{1, \dots, S\}\}$ be the set of samples from the random variable $\hat{\xi}$. In order to keep the notation simple, we consider only one arbitrary scenario $j \in \{1, \dots, S\}$ for now. Therefore, we can drop the dependence on ξ^j. So as not to worry about the inhomogeneous Dirichlet boundary conditions (6.4.2b) when discretizing the differential operators, we first introduce the function

$$\tilde{y} \colon \Omega \to [d_0, d_1], \; x \mapsto \tilde{y}(x) := (d_1 - d_0)\, x + d_0. \tag{6.4.6}$$

We can then write the solution $y(u)\left(\xi^j\right) \in H^1(\Omega)$ to (6.4.2) for the given scenario $\xi^j \in \Lambda$ and a control $u \in \mathcal{U}$ as

$$y(u) := y_0(u) + \tilde{y}, \tag{6.4.7}$$

where $y_0(u) \in H_0^1(\Omega)$ satisfies

$$e\big(y_0(u), u\big) := -\nu \partial_{xx}\big(y_0(u) + \tilde{y}\big) + \big(y_0(u) + \tilde{y}\big)\partial_x\big(y_0(u) + \tilde{y}\big) - f - u = 0 \tag{6.4.8}$$

as well as, due to the definition of $H_0^1(\Omega)$, the homogeneous Dirichlet boundary conditions

$$y_0(u)(0) = y_0(u)(1) = 0. \tag{6.4.9}$$

Since we know the derivatives of $\tilde{y}$, we can simplify (6.4.8) to get

$$e\big(y_0(u), u\big) = -\nu \partial_{xx} y_0(u) + \big(y_0(u) + \tilde{y}\big)\partial_x y_0(u) + \big(y_0(u) + \tilde{y}\big)\,(d_1 - d_0) - f - u = 0. \tag{6.4.10}$$

The function values at the grid points will be denoted by

$$y_{0,i} := y_0(u)(x_i), \qquad \tilde{y}_i := \tilde{y}(x_i), \tag{6.4.11}$$

$$f_i := f(x_i), \qquad \text{and} \quad u_i := u(x_i) \tag{6.4.12}$$

for all $i \in \{1, \dots, N\}$. The respective vector notations are

$$\mathbf{y}_0 := (y_{0,1}, \dots, y_{0,N})^\top \in \mathbb{R}^N, \qquad \tilde{\mathbf{y}} := (\tilde{y}_1, \dots, \tilde{y}_N)^\top \in \mathbb{R}^N, \tag{6.4.13}$$

$$\mathbf{f} := (f_1, \dots, f_N)^\top \in \mathbb{R}^N, \qquad \text{and} \quad \mathbf{u} := (u_1, \dots, u_N)^\top \in \mathbb{R}^N. \tag{6.4.14}$$

We will also write $\mathbf{y}_0(\mathbf{u})$ instead of $\mathbf{y}_0$ to emphasize the dependence on the discretized control $\mathbf{u}$. Together with the matrices $D_1, D_2 \in \mathbb{R}^{N \times N}$ given by

$$D_1 := (2\Delta x)^{-1} \begin{pmatrix} 0 & 1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & -1 & 0 \end{pmatrix} \qquad \text{and} \qquad D_2 := (\Delta x)^{-2} \begin{pmatrix} -2 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{pmatrix}, \tag{6.4.15}$$

we can use the first and second order central difference discretization to approximate equation (6.4.10) by

$$\begin{aligned} 0 = e\big(y_0(u), u\big) &\approx \tilde{e}\,(\mathbf{y}_0, \mathbf{u}) \\ &:= -\nu D_2 \mathbf{y}_0 + \operatorname{diag}(\mathbf{y}_0 + \tilde{\mathbf{y}})\, D_1 \mathbf{y}_0 + (\mathbf{y}_0 + \tilde{\mathbf{y}})(d_1 - d_0) - \mathbf{f} - \mathbf{u}. \end{aligned} \tag{6.4.16}$$

This equation can be solved for every (discretized) control $\mathbf{u} \in \mathbb{R}^N$ by applying Newton's method [Del15, Section 4.2] to the function

$$\varphi \colon \mathbb{R}^N \to \mathbb{R}^N, \quad \mathbf{y}_0 \mapsto \varphi(\mathbf{y}_0) := \tilde{e}(\mathbf{y}_0, \mathbf{u}). \tag{6.4.17}$$

The Jacobian of φ at $\mathbf{y}_0 \in \mathbb{R}^N$ is given by

$$J_\varphi(\mathbf{y}_0) := -\nu D_2 + \operatorname{diag}(D_1 \mathbf{y}_0) + \operatorname{diag}(\mathbf{y}_0 + \tilde{\mathbf{y}})\, D_1 + (d_1 - d_0)\mathrm{Id}, \tag{6.4.18}$$

where $\mathrm{Id} \in \mathbb{R}^{N \times N}$ is the identity matrix. Together with a starting vector $\mathbf{y}_0^0 \in \mathbb{R}^N$ and a tolerance $\varepsilon' > 0$ for the stopping criterion, this results in the following algorithm, which must be run separately for each scenario.

Algorithm 6.3 (Newton's Method)

1 Initialize $k := 0$.
2 **repeat**
3 Solve $J_\varphi(\mathbf{y}_0^k)\delta\mathbf{y}_0^k = \varphi(\mathbf{y}_0^k)$ for $\delta\mathbf{y}_0^k$.
4 $\mathbf{y}_0^{k+1} := \mathbf{y}_0^k + \delta\mathbf{y}_0^k$
5 Update $k \leftarrow k + 1$.
6 **until** $\left\| \varphi\left(\mathbf{y}_0^k\right) \right\|_2 < \varepsilon'$

6.4.3 Function Discretization

Except for the PDE, this problem is identical to the one described in Section 6.3.1. Therefore, we can use the discretizations of $\operatorname{prox}_{\mathrm{CVaR}_\beta^*}$, $\operatorname{prox}_G^{T_k}$, and CVaR_β described in Section 6.3.4. However, due to the different PDE and its discretization method, we need to derive approximations for $K(u)$, $K'(u)^*$, and $G(u)$ for every control $u \in \mathcal{U}$.

The function $K \colon \mathcal{U} \to \mathbb{R}^S$ can be discretized by using the trapezoidal rule for every $j \in \{1, \dots, S\}$, which yields

$$K(u)_j \approx K(\mathbf{u})_j := \frac{\Delta x}{4}\left[\left(d_0\left(\xi^j\right) - 1\right)^2 + \left(d_1\left(\xi^j\right) - 1\right)^2 + 2\sum_{i=1}^{N}\left(y_{i,j} - 1\right)^2\right], \tag{6.4.19}$$

where $y_{i,j} := y\left(\xi^j, x_i; u\right)$ for all $i \in \{1, \dots, N\}$.

The j-th component of the Fréchet derivative $K'(u)_j \colon \mathcal{U} \to \mathbb{R}$ for $u \in \mathcal{U}$ is given by

$$h \mapsto K'(u)_j h = \left\langle y\left(\xi^j, \cdot; u\right) - 1, y'(u)\, h \right\rangle_{\mathcal{U}} = \left\langle y'(u)^*\left(y(u) - 1\right), h \right\rangle_{\mathcal{U}} \tag{6.4.20}$$

for every $h \in \mathcal{U}$, where $y'(u) := y'\left(\xi^j; u\right) : \mathcal{U} \to H^1(\Omega)$ is the Fréchet derivative of $\mathcal{U} \ni u' \mapsto y(u')\left(\xi^j\right)$ in u and $y'(u)^* \colon H^1(\Omega) \to \mathcal{U}$ its adjoint operator. The derivative can be determined by differentiating equation (6.4.10) with respect to u, yielding

$$\frac{\mathrm{d}}{\mathrm{d}u} e\big(y_0(u), u\big) = e_{y_0}\big(y_0(u), u\big) y'(u) + e_u\big(y_0(u), u\big) = 0, \tag{6.4.21}$$

where e_{y_0} and e_u denote the partial derivatives of e. Since $e_u\big(y_0(u), u\big) \equiv -\mathrm{Id}$, we can equivalently write

$$y'(u) = e_{y_0}\big(y_0(u), u\big)^{-1} \approx J_\varphi\left(\mathbf{y}_0\right)^{-1}, \tag{6.4.22}$$

where $J_\varphi\left(\mathbf{y}_0\right)$ is the Jacobian from (6.4.18). The adjoint operator can then be approximated by transposition, i.e.

$$y'(u)^* \approx \left(J_\varphi\left(\mathbf{y}_0\right)^\top\right)^{-1}, \tag{6.4.23}$$

where we have used that the transpose of the inverse is equal to the inverse of the transpose of a matrix. In order to determine the function $K'(u)^* \colon \mathbb{R}^S \to \mathcal{U}$, we note that for all $h \in \mathcal{U}$ and $\mathbf{v} = (v_1, \ldots, v_S) \in \mathbb{R}^S$ we have

$$\begin{aligned}
\langle K'(u)^* \mathbf{v}, h\rangle_{\mathcal{U}} &= \langle \mathbf{v}, K'(u) h\rangle_{\mathbb{R}^S} \\
&\overset{(6.4.20)}{=} \sum_{j=1}^{S} v_j \left\langle y'\left(\xi^j, \cdot; u\right)^* \left(y\left(\xi^j, \cdot; u\right) - 1\right), h\right\rangle_{\mathcal{U}} \\
&= \left\langle \sum_{j=1}^{S} v_j y'\left(\xi^j, \cdot; u\right)^* \left(y\left(\xi^j, \cdot; u\right) - 1\right), h\right\rangle_{\mathcal{U}}
\end{aligned} \tag{6.4.24}$$

due to the bilinearity of $\langle \cdot, \cdot \rangle_{\mathcal{U}}$. Therefore, we can write the approximation as

$$\begin{aligned}
K'(u)^* \mathbf{v} &= \sum_{j=1}^{S} v_j y'\left(\xi^j, \cdot; u\right)^* \left(y\left(\xi^j, \cdot; u\right) - 1\right) \\
&\overset{(6.4.23)}{\approx} \sum_{j=1}^{S} v_j \left(J_\varphi\big(\mathbf{y}_0(\mathbf{u})\big)^\top\right)^{-1} \left(\mathbf{y}_0(\mathbf{u}) + \tilde{\mathbf{y}} - 1\right) && (6.4.25) \\
&= K'(\mathbf{u})^* \mathbf{v} && (6.4.26)
\end{aligned}$$

where $K'(\mathbf{u})^* \colon \mathbb{R}^S \to \mathbb{R}^N$ is the matrix in $\mathbb{R}^{N \times S}$ given by

$$K'(\mathbf{u})^* := \begin{pmatrix} \left(J_\varphi\big(\mathbf{y}_0(\mathbf{u})\big)_1\right)^{-\top} \left(\mathbf{y}_0(\mathbf{u}) + \tilde{\mathbf{y}} - 1\right) & \cdots & \left(J_\varphi\big(\mathbf{y}_0(\mathbf{u})\big)_S\right)^{-\top} \left(\mathbf{y}_0(\mathbf{u}) + \tilde{\mathbf{y}} - 1\right) \end{pmatrix}. \tag{6.4.27}$$

Note that, since $\mathbf{y}_0$ is the approximate solution of $\tilde{e}\left(\mathbf{y}_0, \mathbf{u}\right) = 0$ (see equation (6.4.16)), it still depends on the sample ξ^j (even if this is not explicitly mentioned here).

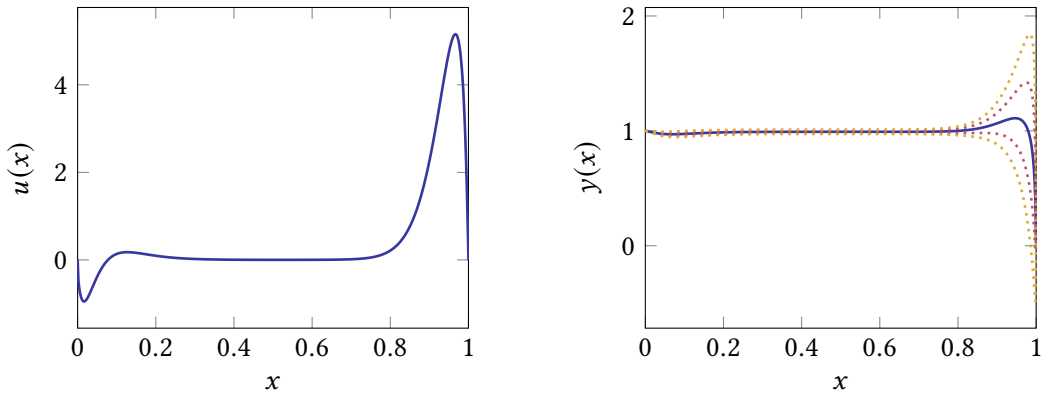
Finally, the function G can be discretized by using the trapezoidal rule for the integral in (6.3.6), which yields

$$\begin{aligned}
G(\mathbf{u}) &= \frac{\alpha}{2} \sum_{i=1}^{N+1} \Delta x \frac{u_{i-1}^2 + u_i^2}{2} \\
&= \frac{\alpha}{2} \Delta x \sum_{i=1}^{N} u_i^2
\end{aligned} \tag{6.4.28}$$

for every $\mathbf{u} \in \mathcal{U}_{ad} = \left\{ z \in \mathbb{R}^N \;\middle|\; u_a(x_i) \le z_i \le u_b(x_i) \text{ for all } i \in \{1, \ldots, N\} \right\}$.

6.4.4 Numerical Results

In this section, we present some numerical results using the Julia code provided in [Ang22, SBE.jl]. As in Section 6.3.6, we define the bounds within $\mathcal{U}_{ad}$ by -10 and 10, use the starting vectors $\mathbf{u}_0 = (0, \dots, 0) \in \mathbb{R}^N$ and $\mathbf{v}_0 = (0, \dots, 0) \in \mathbb{R}^S$, and set $\tilde{\gamma}_G := 0$. Furthermore, we use the same rule to determine the primal step size for a given dual step size $\sigma > 0$.

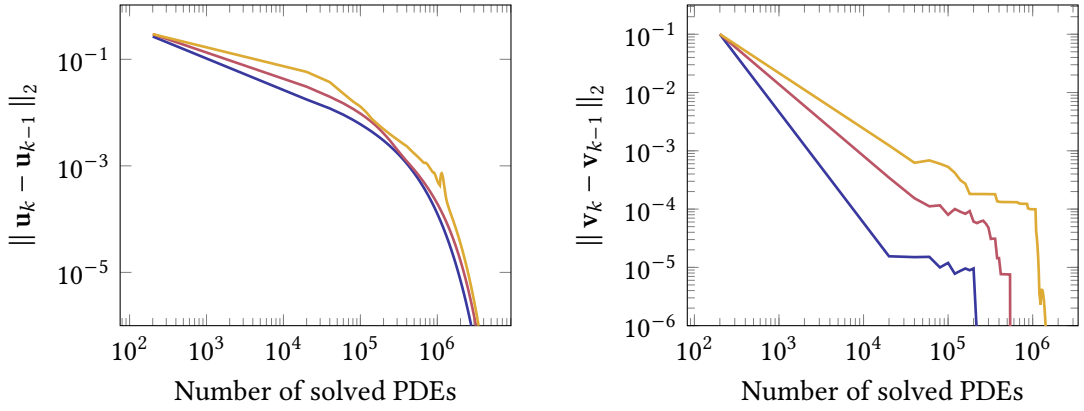


(a) Optimal control

(b) Optimal state (mean ± one and two standard deviations)

Figure 6.9: Example of optimal control and state for $\beta = 0.5$

A first impression of the problem is given in Figure 6.9, where the optimal control and state of (6.4.1) are shown for a probability level of $\beta = 0.5$. In this and all following examples, we use $S = 100$ scenarios and $N = 512$ grid points. The solution was computed using constant scalar step sizes with $\sigma = 0.1$, a stopping criterion tolerance of $\varepsilon = 10^{-6}$, and a tolerance for Newton's method of $\varepsilon' = 10^{-8}$.



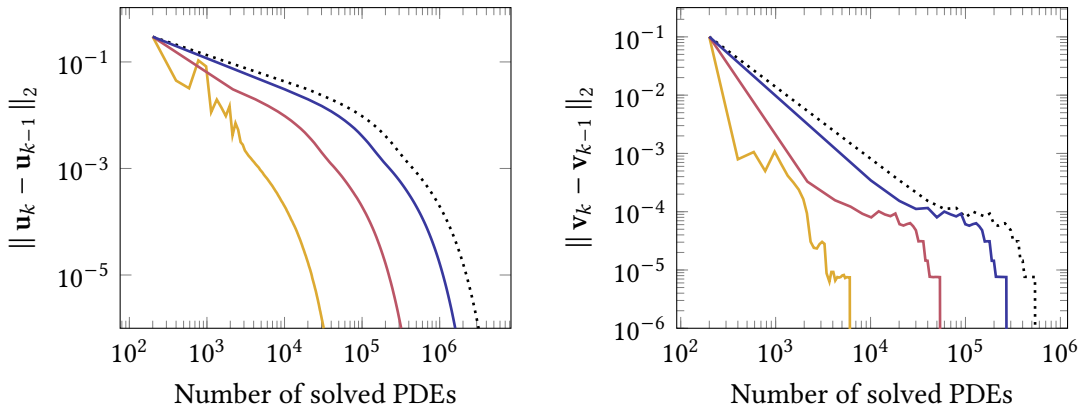
(a) Convergence of primal variable

(b) Convergence of dual variable

Figure 6.10: Convergence behavior for $\beta = 0.01$, $\beta = 0.5$, and $\beta = 0.99$

The convergence behavior without CGF is presented in Figure 6.10, where $N = 256$ grid points and constant scalar step sizes with $\sigma = 0.1$ were used to compute the solution for the probability levels $\beta \in \{0.01, 0.5, 0.99\}$. The stopping criterion tolerance for this and all following examples is $\varepsilon = 10^{-6}$ with a tolerance for Newton's method of $\varepsilon' = 10^{-8}$. As in Section 6.3.6, we can observe that a higher probability level does lead to a slower convergence. In the following analysis, we consider the case $\beta = 0.5$.

For the investigation of index selection rule №1, we use the same constant sequence of probabilities $q_k := q$ for some $q \in [0, 1]$ and all $k \in \mathbb{N}$ and $M := 10^{20}$ as in Section 6.3.6. Figure 6.11 shows the convergence behavior for $\beta = 0.5$ and $q \in \{0.5, 0.1, 0.01\}$. The required number of PDE solves in relation to the approach is shown in Table 6.3. As in Table 6.2, we can also observe here that the number of required PDE solves is pretty close to the product of q and the number of PDE solves without CGF.



(a) Convergence of primal variable

(b) Convergence of dual variable

Figure 6.11: Convergence behavior for $\beta = 0.5$ with CGF and index selection rule №1 for $q = 0.5$, $q = 0.1$, $q = 0.01$, and without CGF (dotted)

We already noticed in Section 6.3.6 that the use of index selection rule №2 leads to similar savings in the number of solved PDEs, which is why we do not consider index selection rule №2 in this example at all.

	# solved PDEs	
q	absolute	relative
-	$5.98 \cdot 10^6$	100%
0.5	$1.59 \cdot 10^6$	49.95%
0.2	$6.36 \cdot 10^5$	20.00%
0.1	$3.19 \cdot 10^5$	10.04%
0.05	$1.60 \cdot 10^5$	5.03%
0.01	$3.21 \cdot 10^4$	1.01%

Table 6.3: Number of solved PDEs for $\beta = 0.5$ with CGF and index selection rule №1 in relation to the choice of q

Chapter 7

Conclusion

In this thesis, we considered a non-convex optimization problem that is constrained by a partial differential equation (PDE) with uncertain coefficients. This problem was described in Chapter 3 after a short summary of some elementary but necessary results from the involved mathematical fields in Chapter 2. We chose the Conditional Value-at-Risk (CVaR), which is non-smooth, to measure the risk of the random field PDE solution and to include it into the objective function. The well-known *Chambolle-Pock method* does not require the objective function to be differentiable, which is why we used it as a template in Chapter 4 to develop a stochastic primal-dual proximal splitting method that solves the problem at hand. For the randomization of this method, we proposed the so-called *component-wise gradient freezing* or *CGF*, which was motivated by randomized coordinate descent methods and requires that only a subset of the coordinates of an occurring gradient is recalculated in each iteration. Furthermore, we used a special feature of the CVaR resulting in a proximal operator that is simply the metric projection onto the bounded probability simplex. We presented an algorithm that computes this projection efficiently and proved its convergence in Chapter 5. However, as the main part of this work, we proved the almost sure weak convergence of the proposed stochastic primal-dual proximal splitting method under some abstract assumptions in Chapter 4 and specified the results for the case of scalar and deterministic step sizes. Furthermore, we implemented the algorithm in Julia and used this code to provide insight into the performance of the algorithm in Chapter 6 using two examples. We discovered that a remarkable reduction of the iteration costs in terms of PDE solves of about 99% can be achieved in some cases by using CGF.

Future Research

Finally, we present some ideas about where future research might build on this work:

- In Section 4.4.4, we proved a lemma providing a primal step size bound which sets a limit on how far the next iterate can escape from a given neighborhood around a critical point. This lemma could be used to relax the requirement on G having a bounded domain in Assumption 3.1.2. It was needed in Lemma 4.1.1 to show that the primal iterates stay within a closed ball around the critical point. However, the idea in [CV20, Lemma 4.6] can possibly be adapted to also work in the present case. This way, one could get rid of the box constraints in (6.3.5).
- In Section 4.4, we considered the abstract convergence proof of the previous section and developed assumptions that are easier to verify in the case of scalar and deterministic step sizes. Firstly, the restriction of a fixed dual step size could be dropped. Secondly, the requirement of scalar and deterministic step sizes could be weakened by allowing step size operators, which may also be random in addition [Val19].

- We described in Section 5.1 the discretization of the probability space based on Monte Carlo sampling and the sample average approximation introduced in Section 2.5.2. However, additionally to CGF, a sparse grid [GG98] could result in a greatly reduced number of solved PDEs, as demonstrated in [KS16]. Another approach is called *adaptive stochastic collocation*, which is an interpolation-based technique that produces decoupled systems of deterministic PDEs [Kou+13].
- Apart from the discretization of the probability space, one could also employ a more sophisticated discretization of the PDEs in space. For example, Kouri and Surowiec [KS16] used continuous piecewise linear finite elements built on a piecewise uniform mesh that is denser in those areas where optimal control and state of the problem constrained by the steady Burgers' equation deflect strongly (see Figure 6.9).
- The index selection rules proposed in Section 6.2 are rather simple because we required the independence of the sequence $(A_k)_{k\in\mathbb{N}}$ in order to satisfy Assumption 4.4.10 (Reduction of CGF). However, there may be a way to remove this condition of independence so that more sophisticated rules for selecting the index sets could be used. For example, an adaptive sample size [BBN18; Bei+20; Cha+18] could further reduce the number of necessary PDE solves.

Bibliography

[ACL06] S. Alexander, T. F. Coleman, and Y. Li. "Minimizing CVaR and VaR for a portfolio of derivatives". In: *Journal of Banking & Finance* 30.2 (2006), pp. 583–605. DOI: 10.1016/j.jbankfin.2005.04.012.

[AFC19] A. Alacaoglu, O. Fercoq, and V. Cevher. *On the convergence of stochastic primal-dual hybrid gradient*. 2019. arXiv: 1911.00799.

[An+17] Y. An et al. "Robust treatment planning with conditional value at risk chance constraints in intensity-modulated proton therapy". In: *Medical Physics* 44.1 (2017), pp. 28–36. DOI: 10.1002/mp.12001.

[And+01] F. Andersson et al. "Credit risk optimization with Conditional Value-at-Risk criterion". In: *Mathematical Programming* 89.2 (2001), pp. 273–291. DOI: 10.1007/PL00011399.

[Ang18] S. Angerhausen. *Non-Smooth Optimization of the Conditional Value-at-Risk*. Master's Thesis. 2018. DOI: 10.5281/zenodo.6418547.

[Ang22] S. Angerhausen. *A Stochastic Primal-Dual Proximal Splitting Method for Risk-Averse Optimal Control of PDEs: Julia Code (v1.0.1)*. 2022. DOI: 10.5281/zenodo.7121224.

[Art+99] P. Artzner et al. "Coherent measures of risk". In: *Mathematical Finance* 9.3 (1999), pp. 203–228. DOI: 10.1111/1467-9965.00068.

[BBN18] R. Bollapragada, R. Byrd, and J. Nocedal. "Adaptive Sampling Strategies for Stochastic Optimization". In: *SIAM Journal on Optimization* 28.4 (2018), pp. 3312–3343. DOI: 10.1137/17M1154679.

[BC17] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. 2nd ed. Cham: Springer International Publishing, 2017, pp. 1–469. DOI: 10.1007/978-3-319-48311-5.

[BCH15] R. I. Boţ, E. R. Csetnek, and C. Hendrich. "Inertial Douglas–Rachford splitting for monotone inclusion problems". In: *Applied Mathematics and Computation* 256 (2015), pp. 472–487. DOI: 10.1016/j.amc.2015.01.017.

[Bei+20] F. Beiser et al. *Adaptive sampling strategies for risk-averse stochastic optimization with constraints*. 2020. arXiv: 2012.03844.

[BH15] R. I. Boţ and C. Hendrich. "Convex risk minimization via proximal splitting methods". In: *Optimization Letters* 9.5 (2015), pp. 867–885. DOI: 10.1007/s11590-014-0809-8.

[Bjö15] A. Björck. *Numerical Methods in Matrix Computations*. Cham: Springer, 2015. DOI: 10.1007/978-3-319-05089-8.

[BL11] J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. 2nd ed. New York: Springer, 2011. DOI: 10.1007/978-1-4614-0237-4.

[Boy+11] S. Boyd et al. "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers". In: *Foundations and Trends in Machine Learning* 3.1 (2011), pp. 1–122. DOI: 10.1561/2200000016.

[BPP16] W. A. Bukhsh, A. Papakonstantinou, and P. Pinson. "A Robust Optimisation Approach using CVaR for Unit Commitment in a Market with Probabilistic Offers". In: *2016 IEEE International Energy Conference (ENERGYCON)*. 2016, pp. 1–6. DOI: 10.1109/ENERGYCON.2016.7514076.

[Bre11] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. New York: Springer Science and Business Media, 2011. DOI: 10.1007/978-0-387-70914-7.

[BT97a] D. Bertsekas and J. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Belmont: Athena Scientific, 1997. ISBN: 1-886529-01-9.

[BT97b] D. Bertsimas and J. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997. ISBN: 1-886529-19-1.

[Ceg12] A. Cegielski. *Iterative Methods for Fixed Point Problems in Hilbert Spaces*. Berlin Heidelberg: Springer, 2012. DOI: 10.1007/978-3-642-30901-4.

[Cha+18] A. Chambolle et al. "Stochastic Primal-Dual Hybrid Gradient Algorithm with Arbitrary Sampling and Imaging Applications". In: *SIAM Journal on Optimization* 28.4 (2018), pp. 2783–2808. DOI: 10.1137/17M1134834.

[CHL08] K.-W. Chang, C.-J. Hsieh, and C.-J. Lin. "Coordinate Descent Method for Large-scale L2-loss Linear Support Vector Machines". In: *Journal of Machine Learning Research* 9.45 (2008), pp. 1369–1398. URL: http://jmlr.org/papers/v9/chang08a.html.

[Cla20] C. Clason. *Introduction to Functional Analysis*. Cham: Birkhäuser, 2020. DOI: 10.1007/978-3-030-52784-6.

[Cla21] C. Clason. *Introduction to Finite Element Methods*. Lecture Notes. 2021. arXiv: 1709.08618v2.

[Cla90] F. H. Clarke. *Optimization and Nonsmooth Analysis*. Society for Industrial and Applied Mathematics, 1990. DOI: 10.1137/1.9781611971309.

[CMP14] T. C. Y. Chan, H. Mahmoudzadeh, and T. G. Purdie. "A robust-CVaR optimization approach with application to breast cancer therapy". In: *European Journal of Operational Research* 238.3 (2014), pp. 876–885. DOI: 10.1016/j.ejor.2014.04.038.

[CMV19] C. Clason, S. Mazurenko, and T. Valkonen. "Acceleration and global convergence of a first-order primal-dual method for nonconvex problems". In: *SIAM Journal on Optimization* 29.1 (2019), pp. 933–963. DOI: 10.1137/18M1170194.

[CMV21] C. Clason, S. Mazurenko, and T. Valkonen. "Primal–Dual Proximal Splitting and Generalized Conjugation in Non-smooth Non-convex Optimization". In: *Applied Mathematics & Optimization* 84.2 (2021), pp. 1239–1284. DOI: 10.1007/s00245-020-09676-1.

[CP07] P. L. Combettes and J.-C. Pesquet. "A Douglas–Rachford Splitting Approach to Nonsmooth Convex Variational Signal Recovery". In: *IEEE Journal of Selected Topics in Signal Processing* 1.4 (2007), pp. 564–574. DOI: 10.1109/JSTSP.2007.910264.

[CP11a] A. Chambolle and T. Pock. "A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging". In: *Journal of Mathematical Imaging and Vision* 40.1 (2011), pp. 120–145. DOI: 10.1007/s10851-010-0251-1.

[CP11b] P. L. Combettes and J.-C. Pesquet. "Proximal Splitting Methods in Signal Processing". In: *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Ed. by H. H. Bauschke et al. New York, NY: Springer New York, 2011, pp. 185–212. DOI: 10.1007/978-1-4419-9569-8_10.

[CP15] P. L. Combettes and J. C. Pesquet. “Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping”. In: *SIAM Journal on Optimization* 25.2 (2015), pp. 1221–1248. DOI: 10.1137/140971233.

[CP16] A. Chambolle and T. Pock. “An introduction to continuous optimization for imaging”. In: *Acta Numerica* 25 (2016), pp. 161–319. DOI: 10.1017/S096249291600009X.

[CR97] G. H.-G. Chen and R. T. Rockafellar. “Convergence Rates in Forward-Backward Splitting”. In: *SIAM Journal on Optimization* 7.2 (1997), pp. 421–444. DOI: 10.1137/S1052623495290179.

[CV17] C. Clason and T. Valkonen. “Primal-Dual Extragradient Methods for Nonlinear Nonsmooth PDE-Constrained Optimization”. In: *SIAM Journal on Optimization* 27.3 (2017), pp. 1314–1339. DOI: 10.1137/16M1080859.

[CV20] C. Clason and T. Valkonen. *Introduction to Non-Smooth Analysis*. 2020. arXiv: 2001.00216v3.

[CW05] P. L. Combettes and V. R. Wajs. “Signal Recovery by Proximal Forward-Backward Splitting”. In: *Multiscale Modeling & Simulation* 4.4 (2005), pp. 1168–1200. DOI: 10.1137/050626090.

[CY11] Y. Chen and X. Ye. *Projection Onto a Simplex*. 2011. arXiv: 1101.6081.

[Dan12] D. Daners. “A Short Elementary Proof of Σ 1/k 2 = π 2 /6”. In: *Mathematics Magazine* 85.5 (2012), pp. 361–364. DOI: 10.4169/math.mag.85.5.361.

[Del15] J. C. De los Reyes. *Numerical PDE-Constrained Optimization*. Cham: Springer, 2015. DOI: 10.1007/978-3-319-13395-9.

[DL14] C. Dang and G. Lan. *Randomized First-Order Methods for Saddle Point Optimization*. 2014. arXiv: 1409.8625.

[DM13] P. Drabek and J. Milota. *Methods of Nonlinear Analysis*. 2nd ed. Basel: Birkhäuser, 2013. DOI: 10.1007/978-3-0348-0387-8.

[dOli21] W. de Oliveira. “Risk-Averse Stochastic Programming and Distributionally Robust Optimization Via Operator Splitting”. In: *Set-Valued and Variational Analysis* 29.4 (2021), pp. 861–891. DOI: 10.1007/s11228-021-00600-5.

[DS09] J. Duchi and Y. Singer. “Efficient Online and Batch Learning Using Forward Backward Splitting”. In: *Journal of Machine Learning Research* 10.99 (2009), pp. 2899–2934. URL: http://jmlr.org/papers/v10/duchi09a.html.

[Duc+08] J. Duchi et al. “Efficient Projections onto the L1-Ball for Learning in High Dimensions”. In: *Proceedings of the 25th International Conference on Machine Learning*. ICML ’08. New York, NY, USA: Association for Computing Machinery, 2008, pp. 272–279. DOI: 10.1145/1390156.1390191.

[EB92] J. Eckstein and D. P. Bertsekas. “On the Douglas—Rachford splitting method and the proximal point algorithm for maximal monotone operators”. In: *Mathematical Programming* 55.1 (1992), pp. 293–318. DOI: 10.1007/BF01581204.

[Eck89] J. Eckstein. “Splitting Methods for Monotone Operators with Applications to Parallel Optimization”. PhD thesis. Massachusetts Institute of Technology, 1989.

[Eva10] L. C. Evans. *Partial Differential Equations*. 2nd ed. Providence: American Mathematical Society, 2010. ISBN: 978-0-8218-4974-3.

[FB19] O. Fercoq and P. Bianchi. “A Coordinate-Descent Primal-Dual Algorithm with Large Step Size and Possibly Nonseparable Functions”. In: *SIAM Journal on Optimization* 29.1 (Jan. 2019), pp. 100–134. DOI: 10.1137/18m1168480.

[Gab83] D. Gabay. "Applications of the Method of Multipliers to Variational Inequalities". In: *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*. Ed. by M. Fortin and R. Glowinski. Vol. 15. Studies in Mathematics and Its Applications. Elsevier, 1983, pp. 299–331. DOI: 10.1016/S0168-2024(08)70034-1.

[GB17] P. Giselsson and S. Boyd. "Linear Convergence and Metric Selection for Douglas-Rachford Splitting and ADMM". In: *IEEE Transactions on Automatic Control* 62.2 (2017), pp. 532–544. DOI: 10.1109/TAC.2016.2564160.

[GDE20] E. B. Gutierrez, C. Delplancke, and M. J. Ehrhardt. *Convergence Properties of a Randomized Primal-Dual Algorithm with Applications to Parallel MRI*. 2020. arXiv: 2012.01255v3.

[GG98] T. Gerstner and M. Griebel. "Numerical integration using sparse grids". In: *Numerical Algorithms* 18.3 (1998), p. 209. DOI: 10.1023/A:1019129717644.

[GKT92] H. Goldberg, W. Kampowsky, and F. Tröltzsch. "On Nemitskij Operators in Lp-Spaces of Abstract Functions". In: *Mathematische Nachrichten* 155 (1992), pp. 127–140. DOI: 10.1002/mana.19921550110.

[GS21] C. Geiersbach and T. Scarinci. "Stochastic proximal gradient methods for nonconvex problems in Hilbert spaces". In: *Computational Optimization and Applications* 78.3 (2021), pp. 705–740. DOI: 10.1007/s10589-020-00259-y.

[Hin+09] M. Hinze et al. *Optimization with PDE Constraints*. Dordrecht: Springer, 2009. DOI: 10.1007/978-1-4020-8839-1.

[Hsi+08] C.-J. Hsieh et al. "A Dual Coordinate Descent Method for Large-Scale Linear SVM". In: *Proceedings of the 25th International Conference on Machine Learning*. ICML '08. New York, NY, USA: Association for Computing Machinery, 2008, pp. 408–415. DOI: 10.1145/1390156.1390208.

[HY10] B. He and X. Yuan. "Convergence Analysis of Primal-Dual Algorithms for Total Variation Image Restoration". In: *Optimization Online* (2010), pp. 1–22. URL: http://www.optimization-online.org/DB_HTML/2010/11/2790.html.

[HY12] B. He and X. Yuan. "Convergence Analysis of Primal-Dual Algorithms for a Saddle-Point Problem: From Contraction Perspective". In: *SIAM Journal on Imaging Sciences* 5.1 (2012), pp. 119–149. DOI: 10.1137/100814494.

[Hyt+17] T. Hytönen et al. *Analysis in Banach Spaces*. Vol. I. Cham: Springer International Publishing AG, 2017. DOI: 10.1007/978-3-319-69808-3.

[Kam+14] A. Kammerdiner et al. "Optimization of discrete broadcast under uncertainty using conditional value-at-risk". In: *Optimization Letters* 8.1 (2014), pp. 45–59. DOI: 10.1007/s11590-012-0542-0.

[Kes09] S. Kesavan. *Functional Analysis*. Gurgaon: Hindustan Book Agency, 2009. DOI: 10.1007/978-93-86279-42-2.

[Kle13] A. Klenke. *Probability Theory - A Comprehensive Course*. 2nd ed. London: Springer London, 2013. DOI: 10.1007/978-1-4471-5361-0.

[Kou+13] D. P. Kouri et al. "A Trust-Region Algorithm with Adaptive Stochastic Collocation for PDE Optimization under Uncertainty". In: *SIAM Journal on Scientific Computing* 35.4 (2013), A1847–A1879. DOI: 10.1137/120892362.

[KS16] D. Kouri and T. Surowiec. “Risk-Averse PDE-Constrained Optimization Using the Conditional Value-At-Risk”. In: *SIAM Journal on Optimization* 26.1 (2016), pp. 365–396. DOI: 10.1137/140954556.

[KS18a] D. P. Kouri and T. M. Surowiec. “Existence and optimality conditions for risk-averse PDE-constrained optimization”. In: *SIAM-ASA Journal on Uncertainty Quantification* 6.2 (2018), pp. 787–815. DOI: 10.1137/16M1086613.

[KS18b] D. P. Kouri and A. Shapiro. “Optimization of PDEs with Uncertain Inputs”. In: *Frontiers in PDE-Constrained Optimization*. Ed. by H. Antil et al. New York, NY: Springer New York, 2018, pp. 41–81. DOI: 10.1007/978-1-4939-8636-1_2.

[LD05] H.-J. Lüthi and J. Doege. “Convex risk measures for portfolio optimization and concepts of flexibility”. In: *Mathematical Programming* 104.2 (2005), pp. 541–559. DOI: 10.1007/s10107-005-0628-x.

[Lio71] J. L. Lions. *Optimal Control of Systems Governed by Partial Differential Equations*. Berlin: Springer, 1971. ISBN: 978-3-642-65026-0.

[MB21] A. N. Madavan and S. Bose. “A Stochastic Primal-Dual Method for Optimization with Conditional Value at Risk Constraints”. In: *Journal of Optimization Theory and Applications* 190.2 (2021), pp. 428–460. DOI: 10.1007/s10957-021-01888-x.

[MdP89] N. Maculan and G. G. de Paula. “A linear-time median-finding algorithm for projecting a vector on the simplex of Rn”. In: *Operations Research Letters* 8.4 (1989), pp. 219–222. DOI: 10.1016/0167-6377(89)90064-3.

[Min62] G. J. Minty. “Monotone (nonlinear) operators in Hilbert space”. In: *Duke Mathematical Journal* 29.3 (1962), pp. 341–346. DOI: 10.1215/S0012-7094-62-02933-2.

[MJV19] S. Mazurenko, J. Jauhiainen, and T. Valkonen. *Primal-dual block-proximal splitting for a class of non-convex problems*. 2019. arXiv: 1911.06284.

[MOA11] A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: Theory of Majorization and Its Applications*. 2nd ed. New York: Springer Science and Business Media, 2011. DOI: 10.1007/978-0-387-68276-1.

[Nes12] Y. Nesterov. “Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems”. In: *SIAM Journal on Optimization* 22.2 (2012), pp. 341–362. DOI: 10.1137/100802001.

[NW06] J. Nocedal and S. Wright. *Numerical Optimization*. 2nd ed. New York: Springer Science and Business Media, 2006. DOI: 10.1007/978-0-387-40065-5.

[Pet38] B. J. Pettis. “On Integration in Vector Spaces”. In: *Transactions of the American Mathematical Society* 44.2 (1938), pp. 277–304. DOI: 10.2307/1989973.

[Pfl00] G. C. Pflug. “Some Remarks on the Value-at-Risk and the Conditional Value-at-Risk”. In: *Probabilistic Constrained Optimization: Methodology and Applications*. Ed. by S. P. Uryasev. Boston, MA: Springer US, 2000, pp. 272–281. DOI: 10.1007/978-1-4757-3150-7_15.

[Poc+09] T. Pock et al. “An algorithm for minimizing the Mumford-Shah functional”. In: *2009 IEEE 12th International Conference on Computer Vision*. 2009, pp. 1133–1140. DOI: 10.1109/ICCV.2009.5459348.

[Roc70] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970. DOI: 10.1515/9781400873173.

[Roc76] R. T. Rockafellar. “Monotone Operators and the Proximal Point Algorithm”. In: *SIAM Journal on Control and Optimization* 14.5 (1976), pp. 877–898. DOI: 10.1137/0314056.

[RS06] A. Ruszczyński and A. Shapiro. "Optimization of Risk Measures". In: *Probabilistic and Randomized Methods for Design under Uncertainty*. Ed. by G. Calafiore and F. Dabbene. London: Springer London, 2006, pp. 119–157. DOI: 10.1007/1-84628-095-8_4.

[RS71] H. Robbins and D. Siegmund. *A convergence theorem for non negative almost supermartingales and some applications*. Ed. by J. S. Rustagi. New York: Academic Press, 1971, pp. 233–257. DOI: 10.1016/b978-0-12-604550-5.50015-8.

[RT14] P. Richtárik and M. Takáč. "Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function". In: *Mathematical Programming* 144.1 (2014), pp. 1–38. DOI: 10.1007/s10107-012-0614-z.

[RU00] R. T. Rockafellar and S. Uryasev. "Optimization of Conditional Value-at-Risk". In: *Journal of Risk* 2.3 (2000), pp. 21–41. DOI: 10.21314/JOR.2000.038.

[RU02] R. T. Rockafellar and S. Uryasev. "Conditional value-at-risk for general loss distributions". In: *Journal of Banking and Finance* 26.7 (2002), pp. 1443–1471. DOI: 10.1016/S0378-4266(02)00271-6.

[Rud91] W. Rudin. *Functional Analysis*. 2nd ed. Singapore: McGraw-Hill, 1991. ISBN: 978-0070542365.

[RUZ02] R. T. Rockafellar, S. P. Uryasev, and M. Zabarankin. *Deviation Measures in Risk Analysis and Optimization*. 2002. DOI: 10.2139/ssrn.365640.

[San+17] A. Santara et al. *RAIL: Risk-Averse Imitation Learning*. 2017. arXiv: 1707.06658.

[SDR14] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming - Modeling and Theory*. 2nd ed. Society for Industrial and Applied Mathematics, 2014. ISBN: 978-1-611973-42-6.

[ST06] R. Schultz and S. Tiedemann. "Conditional Value-at-Risk in Stochastic Programs with Mixed-Integer Recourse". In: *Mathematical Programming* 105.2 (2006), pp. 365–386. DOI: 10.1007/s10107-005-0658-4.

[TK09] A. Takeda and T. Kanamori. "A robust approach based on conditional value-at-risk measure to statistical learning problems". In: *European Journal of Operational Research* 198.1 (2009), pp. 287–296. DOI: 10.1016/j.ejor.2008.07.027.

[Trö10] F. Tröltzsch. *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*. Providence: American Mathematical Society, 2010. DOI: 10.1090/gsm/112.

[Val14] T. Valkonen. "A primal-dual hybrid gradient method for nonlinear operators with applications to MRI". In: *Inverse Problems* 30.5 (2014). DOI: 10.1088/0266-5611/30/5/055012.

[Val18] T. Valkonen. "Testing and Non-linear Preconditioning of the Proximal Point Method". In: *Applied Mathematics and Optimization* (2018). DOI: 10.1007/s00245-018-9541-6.

[Val19] T. Valkonen. "Block-proximal methods with spatially adapted acceleration". In: *ETNA - Electronic Transactions on Numerical Analysis* 51 (2019), pp. 15–49. DOI: 10.1553/etna_vol51s15.

[Val21] T. Valkonen. "First-Order Primal-Dual Methods for Nonsmooth Non-Convex Optimisation". In: *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging*. Springer International Publishing, 2021, pp. 1–42. DOI: 10.1007/978-3-030-03009-4_93-1.

[Vol00] S. Volkwein. “Application of the augmented Lagrangian-SQP method to optimal control problems for the stationary Burgers equation”. In: *Computational Optimization and Applications* 16.1 (2000), pp. 57–81. DOI: 10.1023/A:1008777520259.

[VP17] T. Valkonen and T. Pock. “Acceleration of the PDHGM on Partially Strongly Convex Functions”. In: *Journal of Mathematical Imaging and Vision* 59.3 (2017), pp. 394–414. DOI: 10.1007/s10851-016-0692-2.

[WB09] T. Wu and J. V. Blackhurst. *Managing Supply Chain Risk and Vulnerability - Tools and Methods for Supply Chain Decision Makers*. London: Springer London, 2009. DOI: 10.1007/978-1-84882-634-2.

[WC13] W. Wang and M. Á. Carreira-Perpiñán. *Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application*. 2013. arXiv: 1309.1541.

[Wri15] S. J. Wright. “Coordinate descent algorithms”. In: *Mathematical Programming* 151.1 (2015), pp. 3–34. DOI: 10.1007/s10107-015-0892-3.

[ZBO11] X. Zhang, M. Burger, and S. Osher. “A Unified Primal-Dual Algorithm Framework Based on Bregman Iteration”. In: *Journal of Scientific Computing* 46.1 (2011), pp. 20–46. DOI: 10.1007/s10915-010-9408-8.

[ZL15] Y. Zhang and X. Lin. “Stochastic Primal-Dual Coordinate Method for Regularized Empirical Risk Minimization”. In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, 2015, pp. 353–361. URL: https://proceedings.mlr.press/v37/zhanga15.html.

[ZS15] Z. Zhu and A. J. Storkey. “Adaptive Stochastic Primal-Dual Coordinate Descent for Separable Saddle Point Problems”. In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by A. Appice et al. Cham: Springer International Publishing, 2015, pp. 645–658. DOI: 10.1007/978-3-319-23528-8_40.