Technische Universität München
Fakultät für Mathematik
Lehrstuhl für Mathematische Optimierung

# Optimal Control under Uncertainty: Theory and Numerical Solution with Low-Rank Tensors

Sebastian Raphael Garreis

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

| | | |
|---|---|---|
| Vorsitzende: | | Prof. Dr. Caroline Lasser |
| Prüfer der Dissertation: | 1. | Prof. Dr. Michael Ulbrich |
| | 2. | Prof. Dr. Christian Clason<br>Universität Duisburg-Essen |
| | 3. | Prof. Dr. Matthias Heinkenschloss<br>Rice University, Houston, Texas, USA<br>(schriftliche Beurteilung) |

Die Dissertation wurde am 19. September 2018 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 5. Februar 2019 angenommen.

# Abstract

In this work, a class of optimal control problems under uncertainty constrained by semilinear, elliptic partial differential equations is analyzed. An inexact trust-region algorithm with a suitable error control procedure is presented to solve such problems adaptively. The state and adjoint equations are formulated in a tensor Banach space and solved using low-rank tensor methods. Numerical results show how the algorithms adapt to the problem data. The dissertation is concluded by an outlook to alternative risk measures, which yield risk-averse controls.

# Zusammenfassung

In dieser Arbeit wird eine Klasse von Optimalsteuerungsproblemen unter Unsicherheit analysiert, die semilineare, elliptische partielle Differentialgleichungen als Nebenbedingung haben. Ein inexaktes Trust-Region-Verfahren mit einer geeigneten Vorgehensweise zur Fehlerkontrolle wird vorgestellt, um solche Probleme adaptiv zu lösen. Die Zustands- und adjungierten Gleichungen werden in einem Tensor-Banachraum formuliert und mit Niedrigrangtensormethoden gelöst. Numerische Ergebnisse zeigen, wie sich die Algorithmen an die Problemdaten anpassen. Die Dissertation wird mit einem Ausblick auf alternative Risikomaße abgeschlossen, die risikoaverse Steuerungen liefern.

# Notation

The following notation is used in this thesis:

**Sets:**

| | |
|---|---|
| $\emptyset$ | empty set |
| $\lvert S\rvert$ | cardinality of the set $S$ |
| $2^S$ | power set of the set $S$ |
| $\operatorname{int}(S)$ | the topological interior of the set $S$ |
| $\overline{S}^{\lVert\cdot\rVert} = \operatorname{cl}(S)$ | the topological closure of the set $S$ w. r. t. the norm $\lVert\cdot\rVert$ |
| $S \times T$ | product set of the sets $S$ and $T$ |
| $\mathbb{N}$ | natural numbers: $\mathbb{N} = \{1, 2, 3, \ldots\}$ |
| $\mathbb{N}_0$ | non-negative integers: $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$ |
| $[n]$ | the first $n \in \mathbb{N}$ natural numbers: $[n] = \{1, 2, \ldots, n\}$ |
| $\mathbb{R}$ | real numbers |
| $\mathbb{C}$ | complex numbers |
| $\mathbb{K}$ | general field |
| $\mathbb{R}_{>0}$, $\mathbb{R}_{\geq 0}$, $\mathbb{R}_{<0}$, $\mathbb{R}_{\leq 0}$ | set of positive, non-negative, negative, non-positive real numbers, respectively |
| $\bar{\mathbb{R}}$ | extended real line $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\} = [-\infty, \infty]$ |
| $(a, b)$, $[a, b]$, $[a, b)$, $(a, b]$ | open, closed, half-open intervals, respectively, with endpoints $a, b \in \bar{\mathbb{R}}$, $a \leq b$, empty in the (half-)open case if $a = b$ |

**Vectors/matrices/tensors:**

| | |
|---|---|
| $S^n$ | set of column vectors with $n \in \mathbb{N}$ components from the set $S$ |
| $S^{m \times n}$ | set of $(m \times n)$-matrices with components from the set $S$ |
| $S^{n_1 \times \cdots \times n_d}$ | set of tensors of order $d \in \mathbb{N}$ ($d$-dimensional arrays) with dimensions $n_1, n_2, \ldots, n_d \in \mathbb{N}$ and entries from $S$ |
| $\mathbf{x}(K_1, K_2, \ldots, K_d)$ | indexing of a tensor $\mathbf{x} \in S^{n_1 \times \cdots \times n_d}$: modes $i \in [d]$ indexed by a list $K_i$ with elements from $[n_i]$ remain, modes indexed by a single index $K_i \in [n_i]$ are cut out |
| $\bullet$ | shorthand for $[n_i]$ (the set/list of all indices) when indexing a tensor |

| | |
|---|---|
| $\mathbf{x} \otimes \mathbf{y}$, $\lambda \otimes \mu$, $Y \otimes Z$ | Kronecker/outer product of the vectors/matrices/tensors $\mathbf{x}$ and $\mathbf{y}$, product of the measures $\lambda$ and $\mu$, tensor product of the Banach spaces $Y$ and $Z$ |
| $\mathbf{x} \odot \mathbf{y}$ | Hadamard (componentwise) product of the vectors/matrices/tensors $\mathbf{x}$ and $\mathbf{y}$ |
| $\mathbf{x} \oslash \mathbf{y}$ | componentwise quotient of the vectors/matrices/tensors $\mathbf{x}$ and $\mathbf{y}$ |
| $\mathbf{x}.^{\lambda}$ | componentwise exponentiation of the vector/matrix/tensor $\mathbf{x}$ with the exponent $\lambda \in \mathbb{R}$ |
| $\langle \mathbf{x}, \mathbf{y} \rangle_{s,t}$ | contraction of the tensors $\mathbf{x}$ and $\mathbf{y}$ along the modes $s$ and $t$ |
| $\langle \mathbf{x}, \mathbf{y} \rangle$ | inner product of the tensors $\mathbf{x}$ and $\mathbf{y}$ of the same size |
| $\|\mathbf{x}\|_{\mathsf{F}}$ | Frobenius norm of the tensor/matrix $\mathbf{x}$: $\|\mathbf{x}\|_{\mathsf{F}} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ |
| $\mathbb{1}$ | the vector/matrix/tensor only containing ones or the function which is constant one |
| $I$, $\mathsf{I}$ | identity mapping, identity matrix |
| $\mathsf{x}^{\top}$ | transpose of the vector or matrix $\mathsf{x}$ |

**Banach spaces and operators:**

| | |
|---|---|
| $\mathbb{R}^n$ | Euclidean space equipped with the inner product $(\mathsf{x}, \mathsf{y}) = \mathsf{x}^{\top}\mathsf{y}$ and norm $\|\cdot\| \equiv \|\cdot\|_2$ (see below) |
| $\|\cdot\|_p$ | $p$-norm on $\mathbb{R}^n$: $\|\mathsf{x}\|_p = \left(\sum_{k=1}^n |\mathsf{x}_k|^p\right)^{1/p}$ for $p \in [1, \infty)$, $\|\mathsf{x}\|_\infty = \max_{k \in [n]} |\mathsf{x}_k|$ |
| $Y$, $Y^*$, $\|\cdot\|_Y$ | a general Banach space, its dual space, the norm on $Y$ |
| $\langle \cdot, \cdot \rangle_{Y^*,Y}$ | dual pairing between $Y^*$ and $Y$, i. e., $\langle f, y \rangle_{Y^*,Y} = f(y)$ |
| $\langle \cdot, \cdot \rangle_{Y,Y^*}$ | dual pairing between $Y$ and $Y^*$, i. e., $\langle y, f \rangle_{Y,Y^*} = f(y)$ |
| $y_k \to y$, $y_k \rightharpoonup y$, $y_k \overset{*}{\rightharpoonup} y$ | Convergence of the sequence $(y_k)_{k \in \mathbb{N}} \subset Y$ to $y \in Y$ w. r. t. the strong, weak, weak* topology, respectively, as $k \to \infty$. The weak* topology is defined if $Y$ is the dual space of a given normed space. |
| $\mathcal{L}(Y, Z)$ | space of linear, bounded operators mapping from the Banach space $Y$ to the Banach space $Z$ |
| $\|\cdot\|_{\mathcal{L}(Y,Z)}$ | induced norm $\|A\|_{\mathcal{L}(Y,Z)} = \sup_{\|y\|_Y \leq 1} \|Ay\|_Z$ |
| $A : Y \to Z$ | linear, bounded operator $A$ mapping from $Y$ to $Z$ |
| $A^* : Z^* \to Y^*$ | adjoint operator $A^*$ mapping from the dual space $Z^*$ to the dual space $Y^*$ |
| $Y \hookrightarrow Z$ | continuous embedding of Banach spaces |
| $Y \hookrightarrow\hookrightarrow Z$ | compact embedding of Banach spaces |
| $\iota : Y \hookrightarrow Z$ | canonical embedding of Banach spaces $Y \subset Z$ |

| | |
|---|---|
| $U$, $(\cdot,\cdot)_U$ | a general Hilbert space, the inner product on $U$ |
| $\|\cdot\|_U$ | norm induced by the inner product, i. e., $\|u\|_U = \sqrt{(u,u)_U}$ |
| $N : Y \to Z$ | a general function mapping from the Banach space $Y$ to the Banach space $Z$ |
| $N' : Y \to \mathcal{L}(Y, Z)$ | its first derivative |
| $N_{y_1} : Y \to \mathcal{L}(Y_1, Z)$ | partial derivative of $N$ w. r. t. $y_1$ for $Y = Y_1 \times Y_2$ |
| $\frac{\partial}{\partial y_1}$ | partial derivative of an expression w. r. t. $y_1$ |
| $\frac{\mathrm{d}}{\mathrm{d}\, y_1}$ | total derivative of an expression w. r. t. $y_1$ |
| $N'' : Y \to \mathcal{L}(Y, \mathcal{L}(Y, Z))$ | second derivative of $N$ (etc.) |
| $N_{y_2 y_1} : Y \to \mathcal{L}(Y_2, \mathcal{L}(Y_1, Z))$ | partial second derivative for $Y = Y_1 \times Y_2 \times Y_3$ |
| $J : U \to \mathbb{R}$ | a general functional on the Hilbert space $U$ |
| $\nabla J : U \to U$ | its gradient (Riesz representative of $J'(u) \in U^* \cong U$) |
| $\nabla_{u_1} J : U \to U_1$ | partial gradient for $U = U_1 \times U_2$ |
| $\nabla^2 J : U \to \mathcal{L}(U, U)$ | Hessian of $J$ (linearization of the gradient) |
| $\nabla^2_{u_2 u_1} J : U \to \mathcal{L}(U_2, U_1)$ | partial Hessian of $J$ for $U = U_1 \times U_2 \times U_3$ |

**Spatial domain and functions:**

| | |
|---|---|
| $\Omega$, $x$ | spatial domain $\Omega \subset \mathbb{R}^n$, point $x \in \Omega$ |
| $1_\Omega : \mathbb{R}^n \to \mathbb{R}$ | indicator function of the set $\Omega$: $1_\Omega(x) = 1$ if $x \in \Omega$ and $1_\Omega(x) = 0$ if $x \notin \Omega$ |
| $\partial\Omega$ | boundary of the domain $\Omega$ |
| $\lambda$ | Lebesgue measure on $\Omega$ |
| $f : \Omega \to \mathbb{R}^m$ | a function mapping from $\Omega \subset \mathbb{R}^n$ to $\mathbb{R}^m$ |
| $\mathrm{D}^j f : \Omega \to \mathbb{R}^m$ | (weak) $k$-th partial derivative of $f$, where $j \in \mathbb{N}_0^n$ is a multi-index |

**Probability:**

| | |
|---|---|
| $(\Xi, \mathcal{F}, \mathbb{P})$, $\xi$ | probability space $\Xi$ with $\sigma$-algebra $\mathcal{F}$ and probability measure $\mathbb{P}$, random value $\xi \in \Xi$ |
| $X : \Xi \to \mathbb{R}$ | a real-valued random variable |
| $\mathbb{E}$ | expectation: $\mathbb{E}[X] := \int_\Xi X(\xi)\, \mathrm{d}\mathbb{P}$ |
| Var | variance: $\mathrm{Var}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2]$ |
| Cov | covariance: $\mathrm{Cov}[X, \tilde{X}] = \mathbb{E}[(X - \mathbb{E}[X])(\tilde{X} - \mathbb{E}[\tilde{X}])]$ |
| $\mathrm{CVaR}_\beta$ | conditional value-at-risk (expected shortfall) with quantile parameter $\beta \in (0, 1)$ |

**Function spaces:**

| | |
|---|---|
| $\mathcal{C}(\Omega)$ | space of continuous functions on $\Omega$ |
| $\mathcal{C}^k(\Omega)$ | space of $k$-times continuously differentiable functions on $\Omega$, $k \in \{1, 2, \ldots, \infty\}$ |
| $\mathcal{C}_C^k(\Omega)$ | space of $k$-times continuously differentiable functions on $\Omega$ with compact support |
| $L^p(\Omega)$ | Banach space of equivalence classes of measurable, $p$-integrable ($p \in [1, \infty)$) or essentially bounded ($p = \infty$) functions on $\Omega$ |
| $\lVert \cdot \rVert_{L^p(\Omega)}$ | norm on $L^p(\Omega)$: $\lVert f \rVert_{L^p(\Omega)} = \left(\int_\Omega \lvert f(x) \rvert^p \,\mathrm{d}x\right)^{1/p}$ for $p \in [1, \infty)$, $\lVert f \rVert_{L^\infty(\Omega)} = \operatorname{ess\,sup}_{x \in \Omega} \lvert f(x) \rvert$ |
| $(\cdot, \cdot)_{L^2(\Omega)}$ | inner product on $L^2(\Omega)$: $(f, g)_{L^2(\Omega)} = \int_\Omega f(x) g(x) \,\mathrm{d}x$ |
| $W^{k,p}(\Omega)$ | Sobolev space of $k$-times weakly differentiable $L^p(\Omega)$-functions with $L^p(\Omega)$-derivatives |
| $\lVert \cdot \rVert_{W^{k,p}(\Omega)}$ | norm on $W^{k,p}(\Omega)$: $\lVert f \rVert_{W^{k,p}(\Omega)} = \left(\sum_{\lvert j \rvert \leq k} \lVert \mathrm{D}^j f \rVert_{L^p(\Omega)}^p\right)^{1/p}$ for $p \in [1, \infty)$ and $\lVert f \rVert_{W^{k,\infty}(\Omega)} = \sum_{\lvert j \rvert \leq k} \lVert \mathrm{D}^j f \rVert_{L^\infty(\Omega)}$ |
| $H^k(\Omega)$ | shorthand for the Hilbert space $W^{k,2}(\Omega)$ |
| $(\cdot, \cdot)_{H^k(\Omega)}$ | inner product on $H^k(\Omega)$: $(f, g)_{H^k(\Omega)} = \sum_{\lvert j \rvert \leq k} (\mathrm{D}^j f, \mathrm{D}^j g)_{L^2(\Omega)}$ |
| $H_0^1(\Omega) = \overline{\mathcal{C}_C^\infty(\Omega)}^{\lVert \cdot \rVert_{H^1(\Omega)}}$ | Hilbert space of $H^1(\Omega)$-functions with zero boundary data in the sense of traces |
| $(\cdot, \cdot)_{H_0^1(\Omega)}$ | inner product on $H_0^1(\Omega)$ defined by $(f, g)_{H_0^1(\Omega)} = (\nabla f, \nabla g)_{L^2(\Omega)^n}$, inducing the norm $\lVert \cdot \rVert_{H_0^1(\Omega)}$ on $H_0^1(\Omega)$ and a seminorm on $H^1(\Omega)$ |
| $H^{-1}(\Omega)$ | dual space of $H_0^1(\Omega)$: $H^{-1}(\Omega) = H_0^1(\Omega)^*$ |
| $L_{\mathbb{P}}^p(\Xi; Y)$ | Bochner space of strongly $\mathbb{P}$-measurable, $p$-integrable ($p \in [1, \infty)$) or essentially bounded ($p = \infty$) functions from $\Xi$ to the Banach space $Y$ |
| $\lVert \cdot \rVert_{L_{\mathbb{P}}^p(\Xi;Y)}$ | norm $\lVert \boldsymbol{y} \rVert_{L_{\mathbb{P}}^p(\Xi;Y)} = \left(\int_\Xi \lVert \boldsymbol{y}(\xi) \rVert_Y^p \,\mathrm{d}\mathbb{P}\right)^{1/p}$ for $p \in [1, \infty)$, $\lVert \boldsymbol{y} \rVert_{L_{\mathbb{P}}^\infty(\Xi;Y)} = \operatorname{ess\,sup}_{\xi \in \Xi} \lVert \boldsymbol{y}(\xi) \rVert_Y$ |
| $\mathcal{P}_{\tilde{d}}(\Xi)$ | space of polynomials of *total* degree at most $\tilde{d} \in \mathbb{N}_0$ on $\Xi \subset \mathbb{R}^m$ |
| $\mathcal{P}_d(\Xi)$ | space of polynomials of *coordinate* degree at most $d \in \mathbb{N}_0^m$ on $\Xi \subset \mathbb{R}^m$ |

# Contents

# 1. Introduction

Many phenomena in physics or engineering applications, such as the flight of an aircraft, the distribution of temperature in a heated system, the diffusion of a liquid or a gas in a medium, or a car crash, can be modeled mathematically by differential equations, in particular partial differential equations (PDEs). If analytical solutions to these equations are not available, numerical methods relying on, e. g., finite element (FE) or wavelet discretizations can be used to simulate such systems. Based on these simulations, certain components of the system, for instance, the shape of the wings of the aircraft, the applied heat distribution, the locations of the points where the liquid is inserted into the medium, or the design of the car, can be optimized to meet certain requirements or to minimize a cost functional. In the mentioned examples, one could aim to maximize the lift of the aircraft while keeping the drag below a given threshold, to attain a desired heat or liquid distribution, or to minimize the deformation of the occupants' space in the car while having a small enough deceleration to not injure them heavily. The described problems can be formulated as *optimal control problems* or *optimization problems with PDE constraints*. The mentioned goals appear in the objective functions of them while further requirements may be posed as constraints.

Often, the input parameters of the simulations, such as the wind speed or material properties, are not known to high accuracy a priori or are uncertain by nature. Hence, they can only be assumed to follow a given or estimated probability distribution or to belong to a set. It can be important to investigate how this uncertainty influences output quantities of the simulations. For instance, one could be interested in estimating the distribution of the lift given the distribution of the wind speed or in computing a lower and an upper bound for the heat distribution knowing bounds for the thermal conductivity and the specific heat capacity of the material. *Uncertainty quantification* has become an important field of research in recent years, see, e. g., [106] for an overview or the SIAM/ASA Journal on Uncertainty Quantification, which appeared in 2013 for the first time. A natural next step is to control the systems under uncertain influences optimally, a field of research called *optimal control under uncertainty*, which is the focus of this thesis.

In particular, we consider the problem of selecting a deterministic control for a system described by a semilinear, elliptic PDE with uncertain input parameters, which follow given probability distributions. We develop necessary theory for a class of such problems and investigate efficient numerical algorithms for solving them. The stochastic space is discretized by a stochastic Galerkin method with a full tensor product basis, resulting in exponentially many unknowns in the discretized system. For example, if we have $m \in \mathbb{N}$ uncertain parameters and the dependence of a quantity on each of them is discretized using $n \in \mathbb{N}$ basis functions, respectively, the full tensor product basis consists of $n^m$ functions. To overcome the curse of dimensionality when having many parameters, the respective coefficients in tensor form, i. e., in form of a multi-dimensional array, are represented in a modern low-rank tensor format. These formats can reduce the storage and computational complexity drastically as already

observed in our previous work [46], which is the basis for some parts of this dissertation, but they offer only a limited set of efficiently implementable operations. Therefore, it is a main motivation of this thesis to investigate and develop algorithms which can be implemented using low-rank tensors. The complexity of low-rank tensor arithmetics depends strongly on the required tensor ranks. Truncation to smaller ranks is often necessary to make the algorithms efficient. This yields rounding errors, which have to be controlled suitably during the optimization process to achieve global convergence of the algorithm. Additionally, we control the errors resulting from stochastic and FE discretization and balance all error contributions.

In recent years, the mentioned efficient low-rank tensor formats, namely the Tensor Train format [88] and the hierarchical Tucker format [56], and numerical algorithms for them with good complexity have been developed, see [54, 67, 51] for an overview. Similar ideas have been used in quantum physics before, see, e. g., [83, 64] and the references therein. These formats and algorithms are an important ingredient of the methods developed in this thesis. For instance, they allow the numerical solution of PDEs with many random inputs in a reasonable amount of time. A more detailed introduction and overview of low-rank tensor methods and applications is given in Chapter 2.

## General Problem Setting

For formulating a general problem setting of optimal control under uncertainty matching the problems considered in this work, let $(\Xi, \mathcal{F}, \mathbb{P})$ be a complete probability space. An element $\xi \in \Xi$ of this space stands for uncertain parameters. Let the control space $U$, the state space $Y$, and the image space $Z$ be a Banach spaces. The system to be controlled is under the influence of uncertainty and is described by the state equation

$$E : Y \times U \times \Xi \to Z, \quad E(y(\xi), u, \xi) = 0 \text{ for almost every (a. e.) } \xi \in \Xi, \tag{1.1}$$

where $y(\xi) \in Y$ is the uncertain state. The control $u \in U$ is *deterministic*, i. e., it does not depend on the uncertain inputs and shall be chosen prior to the observation of the uncertainty. Additionally, it is required to belong to a set of admissible controls $U_{\text{ad}} \subset U$. We formulate the optimal control problem

$$\min_{y(\cdot) \in Y, u \in U} \mathcal{R}[J_1(y(\cdot), u, \cdot)] + J_2(u) \;\; \text{s. t.} \;\; E(y(\xi), u, \xi) = 0 \text{ for a. e. } \xi \in \Xi, \; u \in U_{\text{ad}}, \tag{1.2}$$

where $J_1 : Y \times U \times \Xi \to \mathbb{R}$ is the state-dependent part of the objective function, e. g., a tracking-type term, and $J_2 : U \to \mathbb{R}$ is a purely deterministic part used to, e. g., regularize the control. We assume that $J_1(y(\cdot), u, \cdot)$ is always measurable w. r. t. $\xi$ and hence a *random variable* with values in $\mathbb{R}$. A *risk measure* $\mathcal{R} : \mathcal{X} \to \mathbb{R}$ has to be applied to obtain a real-valued objective function which can be minimized. The domain $\mathcal{X} \subset \{f : \Xi \to \mathbb{R}\}$ is a suitable set of random variables, such as $\mathcal{X} = L^1_{\mathbb{P}}(\Xi)$. The risk measure $\mathcal{R}$ should return a typical value of the random variable, e. g., the expectation or a quantile.

Provided that the state equation $E(y(\xi), u, \xi) = 0$ admits a unique solution $y(u)(\xi) \in Y$ for every $u \in U_{\text{ad}}$ and almost every $\xi \in \Xi$, (1.2) can be reduced to an optimization problem

over the deterministic control variable only:

$$\min_{u \in U} \mathcal{R}[\hat{J}_1(u, \cdot)] + J_2(u) \quad \text{s.t.} \quad u \in U_{\text{ad}}, \tag{1.3}$$

where $\hat{J}_1(u, \xi) := J_1(y(u)(\xi), u, \xi)$. An algorithm for solving problems of the form (1.3) is presented and analyzed in Chapter 4. It can deal with inexact objective function and gradient evaluations as well as an inexact projection onto the feasible set $U_{\text{ad}}$.

## Modeling Optimal Control under Uncertainty

We want to stress that the formulation (1.2) does clearly not cover all imaginable formulations of optimal control problems under uncertainty. State constraints are excluded and the control may not be chosen dependent on the concrete realization of $\xi$. For an overview of different variants of modeling optimal control problems under uncertainty we refer to [4, Sec. 2]. The setup (1.2) yields "robust deterministic controls", cf. [4, Sec. 2], i. e., controls which are chosen before observing the random inputs, but account for the uncertainty in the system. Further works following this approach are [69, 70, 71] and our paper [46].

In contrast to that, optimal controls for specific realizations of the random parameters $\xi$ are considered in [25, 4]. Having computed the set of these controls in an offline phase allows to quickly select the optimal control online after observing the concrete values of the parameters. Alternatively, this can be understood as an uncertainty quantification task. Knowing how the optimal control is computed, one can estimate its distribution from the distribution of the parameters. Another alternative is to work with the distribution of the uncertain state $y(\xi)$ instead of only inserting it into a deterministic objective function. For instance, in [108] the stochastic moments of the state are fit to desired values. Similarly, the variance of the state is penalized in [20].

If we consider the formulation (1.2), an important modeling aspect is the choice of the risk measure $\mathcal{R}$. A natural choice is the expectation $\mathcal{R} \equiv \mathbb{E}$ used in [24, 69, 70, 68], our work [46], and in the main part of this thesis (Chapters 3, 5, 6, and 8). This approach is called *risk-neutral control*, see [103, Sec. 6.4] and [3], and does not account for rare but possibly harmful or costly events or the variability of the cost functional, but intends to have a small cost on average.

More general risk measures allowing for *risk-averse control*, see [103, Chap. 6] and [71, 3, 72], and the underlying theory are presented in Chapter 9. In particular, one can choose the smooth *mean-variance risk measure* $\mathcal{R} \equiv \mathbb{E} + \lambda \operatorname{Var}$ for some $\lambda > 0$ [3], which is discussed in more detail in Section 9.2. Other risk-averse risk measures such as the *conditional value-at-risk (CVaR)*—also called tail expectation—enjoy desirable theoretical properties which the mean-variance risk measure lacks, but are nonsmooth. Hence, it is suitable to apply smoothing techniques which retain important properties of the risk measure and enable the use of gradient-based optimization methods [71]. An interior point solution technique for such risk measures is discussed in Section 9.3.

## Numerical Solution of PDEs with Uncertain Inputs

A key ingredient of an algorithm for solving (1.2) is a suitable numerical solution method for the state equation $E(y(\xi), u, \xi) = 0$, which is a PDE with uncertain inputs in our context. Overviews over numerical methods for elliptic PDEs with random inputs are given in [102, 53], where the equations are formulated in the Bochner space $L^2_{\mathbb{P}}(\Xi; Y)$, where $Y$ is a Hilbert space, and in [7], which describes stochastic collocation techniques for linear, elliptic PDEs. In contrast to that, we consider the optimal control of *semilinear*, elliptic PDEs under uncertainty, which requires a more sophisticated analysis. In particular, it is necessary to consider the state space $L^p_{\mathbb{P}}(\Xi; Y)$ with $p > 3$ in our setting, see Chapter 3.

One approach to discretize the parameter space is a sampling-based method. This means that certain realizations of the random parameters $\xi$ are inserted and the respective deterministic PDE is solved. The ensemble of deterministic solutions is then used to approximate the full solution, its distribution function, or stochastic moments. An advantage of such methods is that they are often non-intrusive meaning that deterministic black-box solvers can be reused for the computation. Monte Carlo-type methods, e. g., multilevel Monte Carlo [16, 30, 107, 4] and quasi-Monte Carlo [84, 80], are very popular. They can be applied under fairly general circumstances and feature a typically rather slow convergence speed, which does not depend on the number of uncertain parameters. If the stochastic space $\Xi$ has tensor product structure, adaptive sparse grids can be used for quadrature or interpolation [24, 25, 69, 68]. Under smoothness assumptions on the integrand, they feature exponential convergence [86, 47], but often need a number of collocation points which increases exponentially in the number of parameters [45]. This can lead to a high computational cost if systems with many parameters shall be solved. Furthermore, some negative quadrature weights may appear, which can make discretized optimization problems non-convex or ill-posed although their continuous counterparts do not have these properties. Additionally, it is not possible then to make a smooth reformulation of optimization problems with certain nonsmooth risk measures such as the CVaR, where the nonsmoothness is moved to pointwise bound constraints. The solution of many similar deterministic PDEs in every collocation point can be sped up by using reduced basis methods [42, 28], proper orthogonal decomposition [25], or multigrid techniques [24]. The references [4, 24, 25, 69, 68, 28] indeed focus on optimal control and not only on solution techniques for stochastic PDEs.

Another option is a stochastic Galerkin discretization of the PDE with uncertain inputs [38, 39, 22, 23], where polynomials are chosen to discretize the dependence on the uncertain parameters. To make this efficient, either one chooses a small enough, "sparse" polynomial basis or a large basis with a data-sparse representation of the coefficients, e. g., a tensor product basis with a low-rank tensor representation as in [40]. In the latter case, one can benefit from the good complexity properties of modern low-rank tensor formats w. r. t. the number of parameters, but has to take errors due to tensor truncation into account. The papers [38, 39, 22, 23, 40] deal with adaptive solution techniques for linear, elliptic PDEs with uncertain inputs based on a posteriori error estimates. In this work, we derive a mixture of those, extend it to semilinear, elliptic PDEs, and apply it in the context of optimal control. Hence, we do not focus on error estimates in the energy norm for example, but choose a suitable reference norm, see Chapter 7. To simplify the discretization of the nonlinearity, weighted Lagrange polynomials are chosen as bases of the spaces of polynomials instead of the

usual basis of polynomials with increasing degree, cf. [44] and see Chapter 6. This is related to a full-grid Gaussian quadrature formula, which provides high accuracy and only positive weights. As pointed out in the discussion about negative quadrature weights appearing in sparse grid quadrature formulas, the latter is helpful in an optimization context because smooth reformulations of certain nonsmooth problems are possible in the discrete setting. Stochastic Galerkin methods for control problems were already used in [63, 82, 78, 79, 46]. In combination with low-rank tensors—the option we follow in this work—they were also applied in [21], but no control constraints were posed and the paper puts more focus on efficient linear algebra and works with a fixed discretization, whereas we select the discretization adaptively during the optimization process.

## Optimization Algorithms

In context of optimal control of PDEs, different inexact trust-region type algorithms have been proposed to solve problems adaptively and control the appearing errors due to discretization or the inexact solution of equations in finite dimensions while guaranteeing global convergence. In [119] deterministic PDE-constrained optimization problems are solved by an inexact trust-region sequential quadratic programming (SQP) algorithm, which is applied to the non-reduced optimal control problem, where the state equation is viewed as an equality constraint, cf. (1.2). The sources of inexactness are an adaptive finite element discretization and the inexact solution of the linearized state equation by a conjugate gradient (CG) method. Error estimates are used to control the error which is required for the optimization method to converge. The paper [118] extends this approach to control constraints using an inexact projection onto the feasible set.

In [70] and its predecessor [69], a class of inexact trust-region algorithms formulated in Hilbert space is presented and applied to solve optimal control problems of PDEs under uncertainty. Since these algorithms are posed on the control space $U$, they aim for solving the reduced optimal control problem (1.3). The algorithms in [69] allow for inexact gradient evaluations. They are extended to inexact objective function evaluations in [70]. In the application, the inexactness due to the adaptive discretization of the stochastic space by sparse grids is controlled by the algorithm, but no further error sources are considered. Furthermore, the proposed algorithms cannot handle control-constrained problems.

We extend [70] to the control-constrained case in Chapter 4 by using a possibly inexact projection onto the feasible set as done in [118], but allow for arbitrary mappings approximating the exact projection instead of working with the exact projection on a discrete subspace. We then apply it to the optimal control problem under uncertainty and control all appearing errors simultaneously, namely the FE and the stochastic Galerkin discretization error as well as the algebraic error coming from a low-rank tensor solver used to solve the discretized stochastic PDE. In the paper [46], we applied a semismooth Newton method having fast local convergence properties to solve such problems with a fixed discretization. This method does not feature global convergence and the errors due to the used low-rank tensor arithmetics were not controlled in [46]. Now, we use it as a solver for the trust-region subproblems. The trust-region framework guarantees global convergence and controls all errors appropriately.

## Outline of the Thesis

The rest of this thesis is structured as follows: An introduction to tensors including the used notation, operations, low-rank formats, and available algorithms is given in Chapter 2, where Section 2.1 focuses on finite-dimensional tensors. As already mentioned, one motivation and central question of the dissertation is how low-rank tensor methods can be used to solve optimal control problems under uncertainty. Therefore, we review tensor products of Hilbert spaces and certain Banach spaces in Section 2.2.

This concept is important for the formulation of the class of stochastic PDEs considered in Chapter 3 as well as its analysis and discretization in tensor form. We analyze semilinear, elliptic PDEs with uncertain inputs $\xi \in \Xi$, where the nonlinearity is a superposition operator induced by a $\mathcal{C}^2$-function $\varphi : \mathbb{R} \to \mathbb{R}$ fulfilling suitable growth conditions. Two formulations are presented: Either a deterministic PDE can be solved for a. e. realization $\xi \in \Xi$ or a weak formulation w. r. t. the inputs $\xi$ can be considered by integrating over the parameter space $\Xi$. The latter formulation is essential to apply a stochastic Galerkin discretization. The solution of the PDE is then represented as an element of a suitable Bochner space $L^p_{\mathbb{P}}(\Xi; Y)$ with $p \in (3, \infty)$. This requirement on $p$ is necessary for well-definedness and differentiability properties of the appearing superposition operator. The connection between the Bochner space $L^p_{\mathbb{P}}(\Xi; Y)$ and a tensor product of Banach spaces is established to derive a discretization in tensor form later. We show the equivalence of the pointwise and the weak formulation. In fact, they both have the same unique solution. At this point, it is important to show that the solution constructed pointwise for a. e. $\xi$ is indeed measurable w. r. t. $\xi$ and has the required integrability properties due to a priori estimates, i. e., it is contained in the space $L^p_{\mathbb{P}}(\Xi; Y)$. We formulate a class of risk-neutral optimal control problems with a tracking-type objective function and analyze it. The integrability properties of the state are used to apply the dominated convergence theorem at some points, e. g., to show the existence of a solution to the optimal control problem (Theorem 3.19) and the differentiability of the control-to-state mapping (Theorem 3.26). An adjoint representation of the gradient and Hessian of the reduced objective function is derived. The theory from, e. g., [60, Sec. 1.6] cannot be applied at this point since the linearized state equation operator is not boundedly invertible as a mapping from the state space $L^p_{\mathbb{P}}(\Xi; Y)$ to its dual. It is shown that a pointwisely defined adjoint state can be used to compute the gradient and a meaningful interpretation of the adjoint equation in the space $L^p_{\mathbb{P}}(\Xi; Y)$ is presented.

The reduced optimal control problem, which is an optimization problem formulated in a Hilbert space $U$ with a smooth objective function and a nonempty, closed, convex constraint set $U_{\mathrm{ad}} \subset U$, shall be solved adaptively. Due to discretization errors and inexact state and adjoint equation solves, the objective function and its gradient can only be evaluated inexactly in many relevant cases. Possibly, additional errors appear if the projection onto the feasible set $U_{\mathrm{ad}}$ is computed inexactly. Tailored to this situation, an inexact trust-region algorithm suitable for smooth problems with control constraints is formulated in Chapter 4. It can deal with the mentioned inexact evaluations and features global convergence provided the errors are controlled as demanded by the algorithm. The respective required error bounds are up to fixed, but possibly unknown constants, which typically appear in error estimates due to discretization. Global convergence is proven and it is discussed in detail how the required error tolerances can be met and implemented. In particular, the computation of

a generalized Cauchy point by a projected linesearch with an inexact, but suitably refined projection is discussed. A semismooth Newton method in function space is proposed for the solution of the trust-region subproblems.

The presented trust-region algorithm requires to evaluate the objective function, gradient, and criticality measure of the reduced optimal control problem to a certain accuracy. In Chapter 5, these error bounds are transferred to error bounds on the state, adjoint state, and the projection onto the set of admissible controls. It turns out that—depending on the problem data—the state and adjoint state errors need not be controlled in the $L^p_{\mathbb{P}}(\Xi; Y)$-norm, but a weaker norm such as the $L^2_{\mathbb{P}}(\Xi; Y)$-norm can be sufficient in certain cases.

Chapter 6 describes the discretization of the model problem presented in Chapter 3 and in particular of the PDE with uncertain inputs. We use linear finite elements for the space discretization and a stochastic Galerkin discretization with weighted Lagrange polynomials in tensor product form for the parameter space. It is shown how the discretized equations are formulated within the tensor calculus and how they can be solved by low-rank tensor solvers. The choice of the trust-region model in the discrete setting and the implementation of the semismooth Newton method for box constraints is explained in detail.

As derived in Chapter 5, error control in $L^2_{\mathbb{P}}(\Xi; Y)$ of the state and adjoint state can be sufficient to obtain a globally convergent algorithm. Based on the presented discretization, an a posteriori error estimate is discussed in Chapter 7. It splits into error contributions stemming from the FE discretization error, the stochastic Galerkin error for each random parameter and the algebraic error coming from the iterative low-rank tensor solver. The implementation of the evaluation of this error estimate with low-rank tensors is described. This is crucial because the solutions of the discretized PDEs are represented by low-rank tensors, the full representation of which, i. e., a full array of real numbers should never be computed explicitly due to efficiency reasons. Based on this, the PDEs with uncertain inputs can be solved adaptively.

In Chapter 8, numerical results of the described solution technique are shown for different setups. It can be seen how the algorithms adapt to the problem data, especially in terms of mesh refinement and polynomial grades chosen for the discretization of each parameter. Furthermore, we compare the optimal control obtained from deterministic optimization with the robust one.

Chapter 9 shows how different, risk-averse risk measures can be incorporated into the optimal control problem since the previous parts of the thesis deal with risk-neutral control. A general introduction to risk measures is given. Then, including the mean-variance risk measure into the objective function is discussed in theory. At the end, we investigate a log-barrier reformulation for problems with risk-averse, nonsmooth risk measures which are convex combinations of the mean and the CVaR. It is examined how solving the respective barrier problem, which is an approximation of the original problem, affects the underlying risk measure. At the end, numerical results of an implementation with a fixed discretization and low-rank tensors are presented. It is observed that the control can be chosen such that very large cost function values can be excluded with high probability.

The thesis is concluded in Chapter 10, where we also point out future research perspectives and a few aspects not covered in this work.

## Notation

In analogy to [46], the following notation is used throughout the thesis: Deterministic functions, e. g., the control $u$, and the corresponding function spaces and operators are written in italic font. If we want to emphasize that a function such as the state $\boldsymbol{y}$ belongs to some Bochner space of the form $L^p_{\mathbb{P}}(\Xi; Y)$, i. e., it also depends on the uncertain parameters $\xi$, we use bold and italic font. Functions from finite-dimensional (discretized) spaces are written in roman font (u and $\mathbf{y}$), and the coefficients representing them, such as the vector $\mathsf{u}$ and the tensor $\mathbf{y}$, are in sans-serif font. The same holds for the respective operators.

# 2. Tensors

One motivation of this work is the research on low-rank tensor methods carried out in the numerical linear algebra community in the past years. Tracing back to, e. g., even the 1920s, where one can find the roots of the canonical polyadic decomposition [61], low-rank tensor methods recently got attention within the scientific computing community because of their ability to break the curse of dimensionality. Especially new low-rank formats such as Tensor Train (TT) [88] and Hierarchical Tucker (HT) [56] have the properties of providing stable low-rank approximations and scaling well w. r. t. the tensor dimension. Older formats such as the canonical polyadic decomposition (CP) [93] and the Tucker format [109] lack either the first or the second property, respectively. Suitable numerical algorithms within the formats have been developed during the last two decades, approximately.

In this chapter, we give an introduction to finite-dimensional tensors and low-rank formats, which we will use in our numerical algorithms, and to tensor Hilbert spaces, which are suitable for the formulation of parametric problems in function spaces. Certain examples of tensor Banach spaces needed in Chapter 3 are also discussed.

## 2.1. Finite-Dimensional Tensors and Low-Rank Formats

After discretization, the state of the systems considered in this work can be represented as a real tensor, i. e., a multi-dimensional array of real numbers. To avoid the curse of dimensionality we will represent it in a low-rank format. Both—finite-dimensional tensors and the corresponding low-rank formats—are the topic of this section, which is very similar to the description in our paper [46].

### 2.1.1. Basics of Finite-Dimensional Tensors and Notation

In this thesis, we denote a $d$-dimensional array of numbers in the field $\mathbb{R}$ by a *finite-dimensional tensor* $\mathbf{x} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, cf. [76, Sec. 1], where $\mathbb{C}$ is used as underlying field. The array dimension $d$ is called the *order* of the tensor and the numbers $1, \dots, d$ are the *modes*. We denote the *dimension of the $i$-th mode* by $n_i$ and write $\mathbf{x}(k_1, \dots, k_d)$ for the $(k_1, \dots, k_d)$-component of $\mathbf{x}$ for readability purposes instead of using a subscript notation, where $k_i \in [n_i]$ and $i \in [d]$, writing $[m] := \{1, \dots, m\}$ here and throughout. The tensors of all ones will be denoted by $\mathbb{1}$ if it is clear which space is used. *Reshaping* a tensor as a vector in a certain order, e. g., reverse lexicographical as in [76], is denoted by $\mathrm{vec}(\mathbf{x}) \in \mathbb{R}^{\prod_{i=1}^{d} n_i}$ and reshaping as a matrix by $\mathbf{x}^{(t)} \in \mathbb{R}^{(\prod_{i \in t} n_i) \times (\prod_{j \notin t} n_j)}$, where $t \subset [d]$ is the set of modes that become the rows of the resulting matrix. Analogously we write $\mathrm{ten}(x)$ for reshaping a vector or matrix $x$ back into a tensor, when the dimensions are clear. For the *extraction* of parts of a tensor we also allow indexing with ordered lists and write "$\bullet$" for all indices of a dimension in ascending order: $\mathbf{x}(\bullet, 4, (1, 5), \bullet, \dots, \bullet) \in \mathbb{R}^{n_1 \times 2 \times n_4 \times \cdots \times n_d}$ is obtained from $\mathbf{x}$ by

fixing the second index at 4 and taking the first and the fifth component of the third mode and all components in the rest of the modes. Note that modes that are indexed by a single index only are cut out in the result while modes indexed by a list remain. This is especially relevant when indexing by a list containing only one index: $\mathbf{x}(2, \bullet, \ldots, \bullet) \in \mathbb{R}^{n_2 \times \cdots \times n_d}$, but $\mathbf{x}((2), \bullet, \ldots, \bullet) \in \mathbb{R}^{1 \times n_2 \times \cdots \times n_d}$. Sometimes, the notation becomes shorter if sets instead of lists are used for indexing. The respective set then stands for the list containing all its elements in ascending order. In the $i$-th mode of a tensor, one could replace "$\bullet$" by "$[n_i]$" for example. For a compact notation, we also allow indexing by index vectors, i. e., $\mathbf{x}(k) = \mathbf{x}(k_1, \ldots, k_d)$ for some $k \in \bigtimes_{i=1}^{d} [n_i]$.

As tensors form a vector space, all vector space operations are defined for tensors. Additionally, *componentwise operations* are useful, especially for function-related tensors: $\mathbf{x} \odot \mathbf{y}$ denotes multiplication, $\mathbf{x} \oslash \mathbf{y}$ division and $\mathbf{x}.^{\lambda}$ exponentiation by a scalar $\lambda \in \mathbb{R}$. The componentwise application of a function $f : \mathbb{R} \to \mathbb{R}$ is written as $f(\mathbf{x})$.

**Definition 2.1** ($i$-mode matrix product)**.** Let $\mathbf{x} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ be a tensor, $i \in [d]$ a mode and $\mathsf{A} : \mathbb{R}^{n_i} \to \mathbb{R}^m$ a linear operator or a matrix.
Then the $i$-mode matrix product $\mathsf{A} \circ_i \mathbf{x} \in \mathbb{R}^{n_1 \times \cdots \times n_{i-1} \times m \times n_{i+1} \times \cdots \times n_d}$ is defined by

$$(\mathsf{A} \circ_i \mathbf{x})(k_1, \ldots, k_d) := (\mathsf{A}\mathbf{x}(k_1, \ldots, k_{i-1}, \bullet, k_{i+1}, \ldots, k_d))(k_i)$$

for all $(k_1, \ldots, k_d) \in [n_1] \times \cdots \times [n_{i-1}] \times [m] \times [n_{i+1}] \times \cdots \times [n_d]$.

**Remark 2.2.** Equivalently, the $i$-mode matrix product is given by $\mathsf{A} \circ_i \mathbf{x} = \mathrm{ten}\big(\mathsf{A}\mathbf{x}^{(\{i\})}\big)$, cf. [76, Sec. 4.1], using reshaping and the standard product of matrices. The definition means that we take all vectors that result from fixing all indices of the tensor $\mathbf{x}$ except the $i$-th one, apply $\mathsf{A}$ and then use the resulting vectors to form $\mathsf{A} \circ_i \mathbf{x}$.

We view tensors of order 1 as column vectors and order-2-tensors as matrices, where the first index is for the rows. The contraction of two tensors is a further important operation. It generalizes inner and outer products as well as matrix multiplication.

**Definition 2.3** (tensor contraction)**.** Let $\mathbf{x} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ and $\mathbf{z} \in \mathbb{R}^{\tilde{n}_1 \times \cdots \times \tilde{n}_{\tilde{d}}}$ be tensors and let $s = (s_1, \ldots, s_p)$, $t = (t_1, \ldots, t_p)$ with $s_i \in [d]$, $t_i \in [\tilde{d}]$ be ordered lists of modes that shall be contracted. We require $n_{s_i} = \tilde{n}_{t_i}$ for all $i \in [p]$. Furthermore, let $\bar{s}_i$ ($i \in [d-p]$) and $\bar{t}_i$ ($i \in [\tilde{d}-p]$) be the remaining, untouched modes in ascending order.
Then the *contraction* $\langle \mathbf{x}, \mathbf{z} \rangle_{s,t} \in \mathbb{R}^{n_{\bar{s}_1} \times \cdots \times n_{\bar{s}_{d-p}} \times \tilde{n}_{\bar{t}_1} \times \cdots \times \tilde{n}_{\bar{t}_{\tilde{d}-p}}}$ of $\mathbf{x}$ and $\mathbf{z}$ along the modes $s$ and $t$ is defined as

$$\langle \mathbf{x}, \mathbf{z} \rangle_{s,t}(k_{\bar{s}_1}, \ldots, k_{\bar{s}_{d-p}}, \ell_{\bar{t}_1}, \ldots, \ell_{\bar{t}_{\tilde{d}-p}}) = \sum_{k_{s_1}, \ldots, k_{s_p}=1}^{n_{s_1}, \ldots, n_{s_p}} \mathbf{x}(k_1, \ldots, k_d)\mathbf{z}(\ell_1, \ldots, \ell_{\tilde{d}})|_{\ell_{t_i} = k_{s_i} \, \forall \, i \in [p]}$$

componentwise for all indices $k_{\bar{s}_i} \in [n_{\bar{s}_i}]$ ($i \in [d-p]$) and $\ell_{\bar{t}_j} \in [\tilde{n}_{\bar{t}_j}]$ ($j \in [\tilde{d}-p]$). If $s$ or $t$ are sets rather than ordered lists, they stand for the lists of elements in ascending order as in the case of indexing.

Here, similar to a matrix product, a component of the resulting tensor is obtained by fixing indices in the untouched modes and computing an inner product of the resulting tensors of

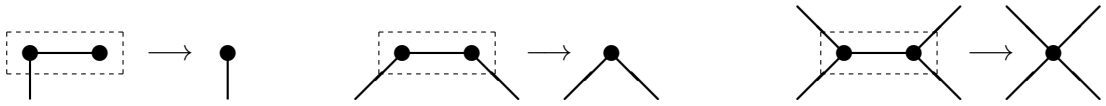


Figure 2.1.: Examples for tensor network diagrams: matrix-vector-product resulting in a vector (left), matrix-matrix-product resulting in a matrix (middle), two tensors of order 3 contracted along one mode resulting in a tensor of order 4 (right)

same size. Tensor contraction can be nicely visualized by tensor network diagrams [76, Sec. 5.1]. The resulting tensor of the contraction is represented by a network, where each node stands for a tensor in the contraction and the edges connected to it are the tensor modes. If two nodes are connected by an edge, the respective tensors are contracted along the respective modes. The "free" edges indicate the modes of the resulting tensor. Some simple examples of tensor networks are shown in Figure 2.1. Note that for a full description of the tensor contraction one should, e. g., display the mode numbers next to the edges, which we skip here as long as it is clear or not essential.

We sometimes need the following special cases of tensor contraction:

**Definition 2.4** (special cases of tensor contraction)**.** Let $\mathsf{x}, \mathsf{y} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ and $\mathsf{z} \in \mathbb{R}^{\tilde{n}_1 \times \cdots \times \tilde{n}_{\tilde{d}}}$ be tensors.

- The *outer product*[1] $\mathsf{x} \otimes \mathsf{z} := \langle \mathsf{x}, \mathsf{z} \rangle_{\emptyset,\emptyset} \in \mathbb{R}^{n_1 \times \cdots \times n_d \times \tilde{n}_1 \times \cdots \times \tilde{n}_{\tilde{d}}}$ of the tensors $\mathsf{x}$ and $\mathsf{z}$ is obtained as

$$(\mathsf{x} \otimes \mathsf{z})\,(k_1, \ldots, k_d, \ell_1, \ldots, \ell_{\tilde{d}}) = \mathsf{x}(k_1, \ldots, k_d)\mathsf{z}(\ell_1, \ldots, \ell_{\tilde{d}}).$$

- An *elementary tensor* or *rank-1-tensor* $\bigotimes_{i=1}^{d} \mathsf{u}^i \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ is an outer product of vectors:

$$\Big(\bigotimes_{i=1}^{d} \mathsf{u}^i\Big)(k_1, \ldots, k_d) = \prod_{i=1}^{d} \mathsf{u}^i_{k_i},$$

  where $\mathsf{u}^i_{k_i}$ is the $k_i$-th component of the vector $\mathsf{u}^i \in \mathbb{R}^{n_i}$.

- The *inner product* $\langle \mathsf{x}, \mathsf{y} \rangle := \langle \mathsf{x}, \mathsf{y} \rangle_{[d],[d]} \in \mathbb{R}$ of two tensors of same size is

$$\langle \mathsf{x}, \mathsf{y} \rangle = \sum_{k_1=1}^{n_1} \cdots \sum_{k_d=1}^{n_d} \mathsf{x}(k_1, \ldots, k_d)\mathsf{y}(k_1, \ldots, k_d).$$

- The *Frobenius norm* of a tensor $\mathsf{x}$ is $||\mathsf{x}||_{\mathsf{F}} := \sqrt{\langle \mathsf{x}, \mathsf{x} \rangle}$.

We conclude this section with the statement that the $i$-mode matrix product is also a special case of contraction since the tensor $\mathsf{A} \circ_i \mathsf{x}$ is obtained as contraction of the second order tensor $\mathsf{A}$ with the tensor $\mathsf{x}$ along the modes (2) and $(i)$.

[1] Note that the vectorization/matricization of the outer product of two tensors coincides with the standard Kronecker product (also denoted by "$\otimes$") of vectorizations/matricizations of them.

### 2.1.2. Low-Rank Tensor Formats

As the amount of data of a full tensor is in general very high, namely of order $\mathcal{O}(n_{\max}^d)$ with $n_{\max} := \max_{i\in[d]} n_i$, it is important to find a representation or approximation of it that requires substantially less memory. One approach is to approximate a tensor by one of low rank. Different notions of tensor rank have been developed in the past. For instance, in the *canonical polyadic (CP) decomposition*, cf. [54, Chap. 7], a rank-$R$-tensor $\mathbf{x}$ is written as the sum of $R$ elementary tensors:

$$\mathbf{x} = \sum_{j=1}^{R} \bigotimes_{i=1}^{d} \mathsf{u}^{i,j} \quad \text{with} \quad \mathsf{u}^{i,j} \in \mathbb{R}^{n_i} \text{ for } i \in [d],\ j \in [R].$$

This representation requires storage of order $\mathcal{O}(Rn_{\max}d)$. We note that this is linear in $d$ instead of exponential in the case of storing the full tensor. But one should take into account that $R$ can grow if a good approximation of a given tensor is desired. Furthermore, this format has certain drawbacks as the possible ill-posedness of the best approximation of a given tensor by a tensor of rank at most $R$ [33]. Concretely, this means that there exists a rank-3-tensor of order $d \geq 3$, which can be approximated arbitrarily well by a rank-2-tensor [54, Prop. 9.10].

A widely used approach for the matrix case $d = 2$ is a low rank approximation by a truncated singular value decomposition (SVD). The quadratic error (squared Frobenius norm) made by this approximation is bounded by the sum of the truncated, squared singular values [54, Lem. 2.30] and can thus be estimated easily. The SVD approximation is optimal w. r. t. this error in the set of matrices of a fixed maximum rank. A generalization of this approximation technique to $d$-dimensional tensors is therefore desirable, but cannot be available in the CP format due to the ill-posedness of the approximation problem. Thus, a different approach works with subspaces of the $\mathbb{R}^{n_i}$, as done in the *Tucker format* or *tensor subspace representation* [54, Chap. 8]: Here, for representing a tensor $\mathbf{x} \in \bigotimes_{i=1}^{d} \mathbb{R}^{n_i}$ (more on this tensor product of Hilbert spaces in Section 2.2), bases of $r_i$-dimensional subspaces of the $\mathbb{R}^{n_i}$ are selected, stored as matrices $\mathsf{U}_i \in \mathbb{R}^{n_i \times r_i}$ and a basis $\bigotimes_{i=1}^{d} \mathsf{U}_i$ of the tensor product of subspaces is formed. One can combine the basis elements linearly with coefficients stored in the so-called *core tensor* $\mathbf{c} \in \mathbb{R}^{r_1 \times \cdots \times r_d}$ to obtain the tensor $\mathbf{x}$. Equivalently, this representation of $\mathbf{x}$ can be written as a contraction of the tensors $\mathsf{U}_i$ and $\mathbf{c}$ w. r. t. a tensor network as depicted in Figure 2.2. Actually, the CP format is a special case of this network with $r_i = R$ for all $i \in [d]$, $\mathsf{U}_i$ containing exactly the vectors $\mathsf{u}^{i,j}$ and $\mathbf{c} \in \mathbb{R}^{R\times\cdots\times R}$ being diagonal in the sense that only the entries $\mathbf{c}(k_1, \ldots, k_d)$ with $k_1 = \ldots = k_d$ are allowed to be non-zero. For the Tucker format the *HOSVD* (higher order SVD, which uses the standard SVD in substeps) is available for truncating a tensor to lower Tucker rank giving a quasi-optimal result[2] and offering error control similar to the one in standard SVD [81]. We give an impression of SVD based tensor truncation in Subsection 2.1.3, using the example of conversion from TT to HT tensors. Unfortunately, defining $r_{\max} := \max_{i\in[d]} r_i$, the required storage for the Tucker format is $\sum_{i=1}^{d} n_i r_i + \prod_{i=1}^{d} r_i = \mathcal{O}(n_{\max} r_{\max} d + r_{\max}^d)$ and grows exponentially in $d$.

[2] Best approximation in the Frobenius norm up to a factor that depends on the order $d$ of the tensor ($\sqrt{d}$ for the Tucker format, see, e. g., [51]).

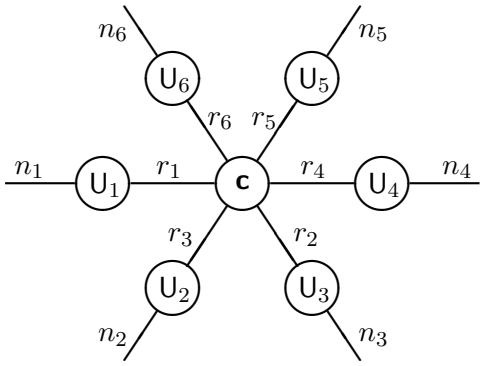



Figure 2.2.: Tensor network representing the Tucker format in the case $d = 6$. The numbers next to the edges are the respective mode dimensions.

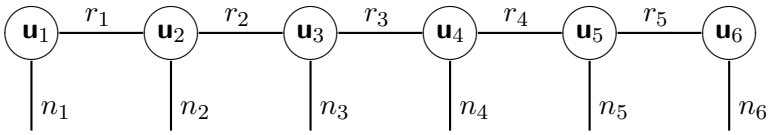



Figure 2.3.: Tensor network representing the Tensor Train format in the case $d = 6$ with the dimensions of the modes displayed next to the edges

Two more recent low-rank tensor formats, meeting the requirement of having a storage and arithmetic complexity scaling linearly in $n_{\max}$ and $d$, and for which a variant of a higher-order SVD is available, are the *hierarchical Tucker (HT) format* [56], and the *tensor train (TT) format* [88], also known as *matrix product states (MPS)* from the quantum physics community [115]. The latter works as follows:

The *TT rank* of a tensor $\mathbf{x} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ is a vector $r := (r_0, r_1, \ldots, r_d)$ with $r_0 = r_d = 1$ and $r_i \in [n_i]$ for $i \in \{2, \ldots, d-1\}$. In the TT format, $\mathbf{x} = \mathcal{TT}(\mathbf{u})$ is represented by a $d$-tuple $\mathbf{u} := (\mathbf{u}_1, \ldots, \mathbf{u}_d) \in \bigtimes_{i=1}^{d} \mathbb{R}^{r_{i-1} \times n_i \times r_i}$ of order-3-tensors:

$$\mathbf{x}(k_1, \ldots, k_d) = \mathbf{u}_1(\bullet, k_1, \bullet)\mathbf{u}_2(\bullet, k_2, \bullet) \cdots \mathbf{u}_d(\bullet, k_d, \bullet) \in \mathbb{R}^{1\times 1} \cong \mathbb{R} \quad \forall\, k_i \in [n_i],\ i \in [d].$$

Note that this is a simple product of matrices, resulting in a scalar value because the first and the last matrix of the chain are a row and a column vector, respectively. The tensor $\mathbf{x}$ can also be written as a contraction of the so-called core tensors $\mathbf{u}_i$ as shown in Figure 2.3. The storage requirement of a tensor in the TT format is exactly $\sum_{i=1}^{d} r_{i-1} n_i r_i = \mathcal{O}(n_{\max} r_{\max}^2 d)$ with $r_{\max} := \max_{i \in \{0,1,\ldots,d\}} r_i$ and scales linearly in $n_{\max}$ and $d$. As in the case of the Tucker format, the low-rank approximation problem is well-posed for TT tensors and an SVD-based truncation to smaller TT rank, the so-called *TT-SVD*, is available and gives a quasi-optimal result up to the factor $\sqrt{d-1}$ [88, Cor. 2.4]. An error control mechanism based on the truncated singular values allows to choose the ranks such that a prescribed approximation accuracy can be guaranteed.

In the related *hierarchical Tucker (HT) decomposition* [56], the idea of writing a tensor as a contraction of smaller ones is generalized in a certain way: The contraction is performed according to a more general, binary dimension tree $\mathcal{T}$, which has one root node $[d]$. Each node

$t \in \mathcal{T}$ containing more than one index splits up into two children $t_\mathrm{l}$, $t_\mathrm{r} \neq \emptyset$ with $t = t_\mathrm{l} \cup t_\mathrm{r}$, where we assume w. l. o. g. that $t_\mathrm{l}$ contains only smaller indices than the ones in $t_\mathrm{r}$. The leaves of the tree are sets of only one index so that we have $d$ leaves and $2d-1$ nodes overall. To simplify the notation we call the leaf nodes $t_1, \dots, t_d$, the parent node of $t_1$ and $t_2$ would be called $t_{12}$ and the root node is $t_{12\dots d}$. The tree tells us how to form a tensor starting with subspaces of the $\mathbb{R}^{n_i}$ and recursively selecting subspaces of tensor products of subspaces. For each leaf node $t_i$ ($i \in [d]$) a matrix $\mathsf{U}_i \in \mathbb{R}^{n_i \times r_i}$ is stored, which contains vectors that span a linear subspace $V_i$ of $\mathbb{R}^{n_i}$ as in the Tucker format. One can use an orthogonal basis here for example. Now, the full tensor is represented by these leaf matrices and additional transfer matrices $\mathsf{B}_t \in \mathbb{R}^{r_\mathrm{l} r_\mathrm{r} \times r_t}$ ($r_t \leq r_\mathrm{l} r_\mathrm{r}$, $r_{[d]} = 1$) assigned to the non-leaf nodes $t$, as follows: For computing the full tensor, which is normally never done explicitly, one would recursively go through the tree from the leaves to the root. At each parent node $t$ a generating system of the tensor product of the two subspaces at the children $t_\mathrm{l}$ and $t_\mathrm{r}$ would be built combining all vectors, i. e., the matrix $\mathsf{U}_\mathrm{r} \otimes \mathsf{U}_\mathrm{l} \in \mathbb{R}^{n_\mathrm{l} n_\mathrm{r} \times r_\mathrm{l} r_\mathrm{r}}$ would be computed. This can also be viewed as a tensor; the Kronecker product is a possible matricization of it. Then, by multiplying with the transfer matrix $\mathsf{B}_t$, which can also be viewed as a three-dimensional tensor $\mathbf{b}_t \in \mathbb{R}^{r_\mathrm{l} \times r_\mathrm{r} \times r_t}$, a generating system of a subspace would be obtained. It can be represented by the matrix $\mathsf{U}_t = (\mathsf{U}_\mathrm{r} \otimes \mathsf{U}_\mathrm{l})\, \mathsf{B}_t \in \mathbb{R}^{n_\mathrm{l} n_\mathrm{r} \times r_t}$, which would contain vectors generating a subspace of $\bigotimes_{i \in t} V_i$. Arriving at the root node one would obtain a $\mathbb{R}^{\prod_{i=1}^{d} n_i}$-vectorization of the represented tensor. On the other hand we can view the represented tensor to be a contraction of the leaf matrices $\mathsf{U}_i$ and transfer tensors $\mathbf{b}_t$ according to a tensor network defined by the tree $\mathcal{T}$ as shown in Figure 2.4, where a typical balanced tree is used. The required storage for the hierarchical Tucker representation is

- $\mathcal{O}(n_\mathrm{max} r_\mathrm{max} d)$ for the basis matrices $U_i$ at each of the $d$ leaf nodes and
- $\mathcal{O}(r_\mathrm{max}^3 d)$ for the transfer matrices $B_t$ at the non-leaf nodes with $r_\mathrm{max} := \max_{t \in \mathcal{T}} r_t$,

and sums to $\mathcal{O}(n_\mathrm{max} r_\mathrm{max} d + r_\mathrm{max}^3 d)$. This is linear in $d$, but again, for a good approximation the hierarchical Tucker rank, i. e., the vector of ranks at the nodes of $\mathcal{T}$, can be required to be chosen dependent on $d$ and $n_\mathrm{max}$. Truncation to a lower rank is well-posed and can be done by the SVD-based *hierarchical SVD (HSVD)* [50] with error control and quasi-optimal approximation up to the factor $\sqrt{2d-3}$.

### 2.1.3. Available Algorithms

Due to their good complexity, we want to use TT or HT tensors for all high-dimensional computations. Many relevant algorithms are available within (MATLAB) toolboxes, which we present in the following. In addition, we discuss some iterative algorithms used for computing quantities which cannot be computed directly with low-rank tensors in a reasonable amount of time. For instance, the componentwise application of a function $f : \mathbb{R} \to \mathbb{R}$ to a low-rank tensor could *in theory* be computed by building the full array, applying $f$ to its components and truncating the result to obtain a low-rank tensor again. But this approach is highly prohibitive because even the storage required for the full tensor could easily exceed any available storage, at least if the order $d$ is large enough and $n_i \geq 2$ for all $i \in [d]$. Therefore, it is important to compute or approximate every tensor within a computation in a low-rank

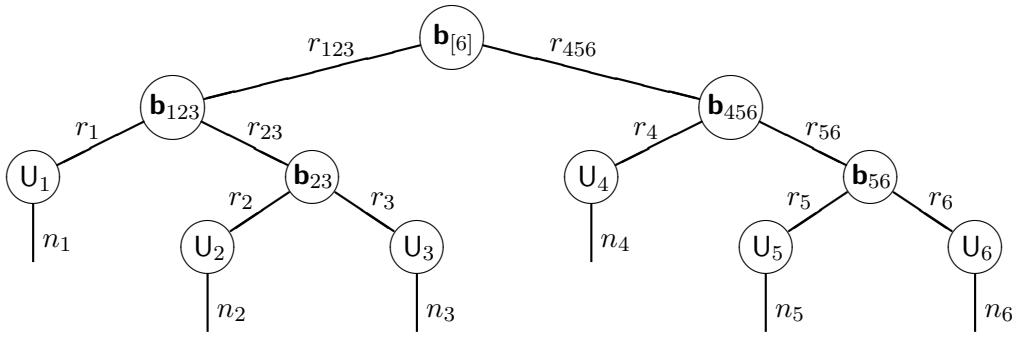



Figure 2.4.: Tensor network representing the Hierarchical Tucker format in the case $d = 6$ with a balanced tree. The dimensions of the modes are displayed next to the edges.

format without ever forming a full tensor explicitly. If there are no direct algorithms available, iterative algorithms operating on the low-rank tensor representation or inexact versions of well-known iterative algorithms formulated on the tensor space $\mathbb{R}^{n_1 \times \cdots \times n_d}$ with truncation in between should be chosen.

### Toolboxes and Conversion between Formats

The `TT-Toolbox` [89] provides efficient implementations of TT tensors and important operations in MATLAB and PYTHON. These are, e. g., elementwise addition and multiplication, extraction of elements of the tensor, implementations of linear operators, and truncation to a lower rank by TT-SVD. Truncation is crucial in practice to avoid infeasible rank growth because when using the standard implementations for summing or multiplying two TT tensors elementwise, the ranks sum or multiply, respectively. Clearly, truncation comes with the drawback of additional errors introduced in the computation. Therefore, it is an important aspect of this thesis how those errors have to be controlled to obtain convergence of the final algorithms, which use tensor computations in substeps.

Due to the similar structure, comparable algorithms are available for HT tensors. The `htucker` toolbox [76] implements tensors in this format and—similarly to the `TT-Toolbox`—efficient versions of the most important operations, e. g., extraction of parts of the tensor, application of linear operators to the $i$-th mode, contraction, tensor orthogonalization and truncation to a lower rank by HSVD [50]. For elementwise addition and multiplication, exact and special, truncated versions are available such that the explosion of ranks can be avoided. HT tensors can be orthogonalized such that the matrices $\mathsf{U}_t$ at all nodes $t$ except for the root node form an orthogonal basis of the respective tensor subspace. This procedure can simplify certain computations and make some tensor algorithms more stable. All mentioned operations can be performed in a reasonable amount of time, which is linear in $n_{\max}$ and $d$ and polynomial in $r_{\max}$, in particular at most $\mathcal{O}(n_{\max} d r_{\max}^2 + d r_{\max}^4)$, but often less. The same holds for tensors in TT format.

For the algorithms developed in this thesis, operations and subsolvers that currently are not all available within a single format or toolbox are needed. Thus, we use HT tensors with a linear, TT-like dimension tree as provided by `htucker`, see Figure 2.5.

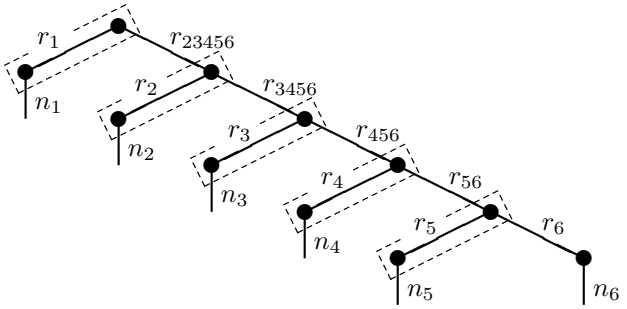



Figure 2.5.: Tensor network representing the Hierarchical Tucker format in the case $d = 6$ with a linear tree, which makes the conversion to TT tensors simple. By performing the marked contractions, a TT network is obtained.

This allows to convert between the HT and TT formats, cf. [54, Sec. 12.2.2], so that both the `TT-Toolbox` and the `htucker` toolbox can be used simultaneously. By performing the marked contractions in Figure 2.5, a TT network is obtained. Conversely, the TT cores $\mathbf{u}_1, \dots \mathbf{u}_{d-1}$ can be split into the basis matrices $\mathsf{U}_1, \dots \mathsf{U}_{d-1}$ and the respective transfer tensors in order to convert from the TT to the HT format: In fact, we aim for finding $r_i \in [n_i]$, $\mathsf{U}_i \in \mathbb{R}^{n_i \times r_i}$ and $\mathbf{b}_{i,\dots,d} \in \mathbb{R}^{r_i \times r_{i+1},\dots,d \times r_{i,\dots,d}}$ such that $\langle \mathsf{U}_i, \mathbf{b}_{i,\dots,d} \rangle_{2,1} = \tilde{\mathbf{u}}_i \in \mathbb{R}^{n_i \times r_{i+1},\dots,d \times r_{i,\dots,d}}$ holds, where $\tilde{\mathbf{u}}_i$ is a reshaped version of $\mathbf{u}_i$. Writing this in a matricized form, we have

$$\mathsf{U}_i \, \mathbf{b}^{(r_i)}_{i,\dots,d} = \tilde{\mathbf{u}}^{(n_i)}_i \in \mathbb{R}^{n_i \times r_{i+1},\dots,d r_{i,\dots,d}}$$

with $\mathbf{b}^{(r_i)}_{i,\dots,d} \in \mathbb{R}^{r_i \times r_{i+1},\dots,d r_{i,\dots,d}}$. Clearly, one could choose $r_i = n_i$, $\mathsf{U}_i = \mathsf{I}_{n_i}$ (identity matrix), and $\mathbf{b}_{i,\dots,d} = \tilde{\mathbf{u}}_i$ to obtain the result, but this approach is not suitable due to the possibly (for large $n_i$) large ranks $r_i = n_i$ so that the large matrices $\mathsf{U}_i \in \mathbb{R}^{n_i \times n_i}$ have to be stored. Low-rank tensor implementations use dense linear algebra because the representing small tensors and matrices cannot be expected to be sparse so that one cannot benefit from the sparsity of the matrix $\mathsf{I}_{n_i}$ here. A smaller rank $r_i$ can be obtained by using an SVD of $\tilde{\mathbf{u}}^{(n_i)}_i = \mathsf{V}_i \Sigma_i \mathsf{W}_i^\top$, where $\mathsf{V}_i \in \mathbb{R}^{n_i \times R_i}$ and $\mathsf{W}_i \in \mathbb{R}^{r_{i+1},\dots,d r_{i,\dots,d} \times R_i}$ are orthogonal matrices, and $\Sigma = \operatorname{diag}((\sigma^i_1, \dots, \sigma^i_{R_i})) \in \mathbb{R}^{R_i \times R_i}$, where $R_i = \min\{n_i,\ r_{i+1},\dots,d \, r_{i,\dots,d}\}$, i. e., we use the economic variant of SVD or *thin SVD* [49, Sec. 2.5.4]. The singular values shall be ordered: $\sigma^i_1 \geq \sigma^i_2 \geq \dots \geq \sigma^i_{R_i}$. Truncating this decomposition to rank $r_i \leq R_i$, we set $\mathsf{U}_i := \mathsf{V}_i(\bullet, [r_i])$ and $\mathbf{b}_{i,\dots,d} := \operatorname{ten}(\Sigma_i([r_i], [r_i]) \, \mathsf{W}_i(\bullet, [r_i])^\top)$ to obtain an approximate decomposition of the TT core tensor $\mathbf{u}_i$. The accuracy of this approximation can be controlled easily since the squared Frobenius error is given by the sum of the truncated, squared singular values.

### Additional Arithmetic Operations

When designing optimization algorithms for constrained problems in tensor space, one often needs more componentwise operations than only addition, subtraction, and multiplication. Standard algorithms for constrained problems typically require the computation of special componentwise functions such as penalty terms, projections, indicators, e. g., for a semi-smooth Newton method, or the reciprocal $\mathbf{x}.^{-1}$, which is needed to compute the derivative of a log-barrier term in an interior point method. One option is to approximate such quantities

by applying Newton's method to a suitable equation. In the case of the elementwise reciprocal $\mathbf{y} = \mathbf{x}.^{-1}$, one can consider the equation $\mathbf{x} - \mathbf{y}.^{-1} = 0$ and solve it for $\mathbf{y}$, see also [43, Sec. 4.4]. The resulting Newton equation given the current iterate $\mathbf{y}_k$ is $\mathbf{y}_k.^{-2}(\mathbf{y}_{k+1} - \mathbf{y}_k) = -\mathbf{x} + \mathbf{y}_k.^{-1}$, giving the Newton update $\mathbf{y}_{k+1} = \mathbf{y}_k \odot (2 \cdot \mathbb{1} - \mathbf{x} \odot \mathbf{y}_k)$. This iteration can be performed with available operations within the low-rank format; especially no componentwise division is needed. It can be shown to converge if one chooses the first iterate $\mathbf{y}_0$ such that $\operatorname{sgn}(\mathbf{y}_0) = \operatorname{sgn}(\mathbf{x})$ and $|\mathbf{y}_0| \leq |\mathbf{x}|.^{-1}$ (componentwise). One can choose, e. g., $\frac{1}{\|\mathbf{x}\|_\mathsf{F}^2}\mathbf{x}$ for general $\mathbf{x}$ or $\frac{1}{\|\mathbf{x}\|_\mathsf{F}}\mathbb{1}$ for $\mathbf{x} > 0$. This iteration was already proposed to approximate the inverse of structured matrices [55, Sec. 4.1] and is called Newton-Schulz method [101]. When implementing it with low-rank tensors—as done in the example section of `htucker`—the ranks of the iterates have to be bounded by truncation. This causes the implementation to be an *inexact* version of Newton's method, which could only be expected to converge if the error caused by the truncation was controlled suitably, see [55, Sec. 2]. In practice, this approach is sometimes unfeasible because the required ranks are too large. If truncation is performed always to a fixed rank, one typically obtains only a solution of a certain accuracy, which can be quantified by $\|\mathbb{1} - \mathbf{x} \odot \mathbf{y}\|_\mathsf{F}$. In order to guarantee convergence of this truncated iteration for $\mathbf{x} > 0$ one would have to ensure that all iterates fulfill $\mathbf{y}_k > 0$ elementwise. By error control in the maximum norm $\|\cdot\|_\infty$ this goal could be achieved. Such a rounding procedure is currently not available to our knowledge. Using the equivalence of norms in finite-dimensional spaces is not suitable here because of the magnitude of the equivalence constant: We have $\|\mathbb{1}\|_\mathsf{F} = \sqrt{n_1 \cdots n_d}$ for $\mathbf{x} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, whereas $\|\mathbb{1}\|_\infty = 1$, but can only ensure $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_\mathsf{F}$ in general, i. e., it could be necessary to decrease the Frobenius norm of the error further although the maximum norm is small enough already.

A similar iteration, namely $\mathbf{y}_{k+1} = \frac{1}{2}\mathbf{y}_k \odot (3 \cdot \mathbb{1} - \mathbf{y}_k.^2 \odot \mathbf{x})$, can be employed to compute $\mathbf{x}.^{-1/2}$ for $\mathbf{x} > 0$. This gives then access to $|\mathbf{x}| \approx \mathbf{x}.^2 \odot (\mathbf{x}.^2 + \varepsilon \cdot \mathbb{1}).^{-1/2}$, where the addition of $\varepsilon \cdot \mathbb{1}$ with a small $\varepsilon > 0$ makes the algorithm more stable. Using the componentwise absolute value, quantities such as quadratic penalty terms or projections onto boxes can be approximated with low-rank tensors and used in an optimization algorithm.

In our experiments [46] it was found that the Newton-Schulz iteration performs often well in practice, but does not give satisfying results sometimes if the entries of the positive tensor $\mathbf{x} > 0$ are too close to 0. Then we computed $\mathbf{y} = \mathbf{x}.^{-1}$ by solving the linear system $\mathbf{x} \odot \mathbf{y} = \mathbb{1}$ by an iterative method working on the low-rank tensor representation, such as ALS, AMEn or optimization on manifolds, see below.

The iterative methods for computing componentwise functions of low-rank tensors in a low-rank format described above are all designed for a specific problem. This has the advantage that they can be analyzed very well, but comes with the drawback that a separate algorithm has to be designed for each function. A more general approach for computing componentwise functions is given by sampling some entries of the resulting $\mathbf{y}$, i. e., computing $\mathbf{y}(k) = f(\mathbf{x}(k))$ for some indices $k$ in some index set $K \subset \bigtimes_{i=1}^{d} [n_i]$. Then the full tensor in the low-rank format is computed by some tensor completion method. Such methods aim to find a tensor of a prescribed or adaptively selected low rank which contains the sampled entries, sometimes at least approximately. Suitable algorithms are cross approximation [91, 15, 14] or tensor completion using optimization algorithms such as ALS [62] or optimization on the manifold of tensors of fixed TT or HT rank [73, 104, 32].

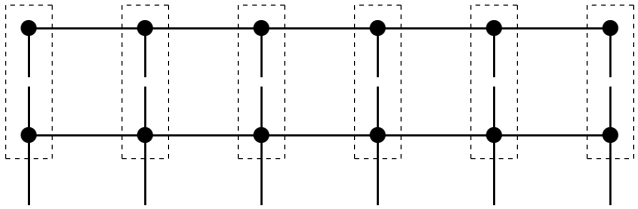


Figure 2.6.: Product of a TT matrix (below) and a TT tensor/vector (above) represented as a contraction. If the modes along which one contracts in the resulting network are "vectorized", a TT network is obtained. The TT ranks of the result are thus obtained by multiplying the TT ranks of the TT vector and the TT matrix. See also [12].

### Solution of Linear Equations

In many applications, the solution of linear systems involving low-rank tensors is a key tool. They arise, e. g., when discretizing PDEs in high space or stochastic dimension. It is then important that the corresponding linear operators can at least be applied efficiently to low-rank tensors. For instance, they can be of the following forms:

- Linear operators in the CP format: Given $d \in \mathbb{N}$ and $R \in \mathbb{N}$, let $\mathsf{A}_i^{(j)} : \mathbb{R}^{n_i} \to \mathbb{R}^{m_i}$ be linear operators or matrices for $i \in [d]$, $j \in [R]$. Consider the operator

$$\mathbf{A} : \mathbb{R}^{n_1 \times \cdots \times n_d} \to \mathbb{R}^{m_1 \times \cdots \times m_d},\ \mathbf{A} = \sum_{j=1}^{R} \bigotimes_{i=1}^{m} \mathsf{A}_i^{(j)},\ \mathbf{A}(\mathsf{x}_1 \otimes \cdots \otimes \mathsf{x}_m) = \sum_{j=1}^{R} \bigotimes_{i=1}^{m} \mathsf{A}_i^{(j)} \mathsf{x}_i,$$

  defined by its action on elementary tensors. Operators of this type can be efficiently applied to low-rank tensors of any type using $i$-mode matrix products and summation:

$$\mathbf{A}\mathbf{x} = \sum_{j=1}^{R} \mathsf{A}_m^{(j)} \circ_m \cdots \mathsf{A}_1^{(j)} \circ_1 \mathbf{x}.$$

- Linear operators in special low-rank formats, such as HTD for linear operators [76] or `tt_matrix` [89]. Operators of this type can easily be applied to HT or TT tensors giving HT or TT tensors as a result, respectively, essentially by performing a suitable contraction. This is visualized in Figure 2.6 for the case of a product of a TT matrix and a TT tensor (vector).

- Componentwise multiplication by a tensor $\mathbf{y} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$: $\mathbf{x} \mapsto \mathbf{y} \odot \mathbf{x}$ is linear and can be computed by the respective algorithms.

In general, such linear operators between tensor spaces cannot be inverted easily except for rank-1-operators: If the matrices $\mathsf{A}_i \in \mathbb{R}^{n_i \times n_i}$ $(i \in [d])$ are invertible, it holds that

$$\Big(\bigotimes_{i=1}^{m} \mathsf{A}_i\Big)^{-1} = \bigotimes_{i=1}^{m} \mathsf{A}_i^{-1}$$

so that the inverse of a rank-1-operator can be applied to low-rank tensors via $i$-mode matrix products. This fact can be used to construct good preconditioners for operators which are perturbations of rank-1-operators. For general linear systems, however, iterative methods have to be applied to compute the solution $\mathbf{x} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ of $\mathbf{A}\mathbf{x} = \mathbf{b}$ with $\mathbf{b} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ and $\mathbf{A} \in \mathcal{L}(\mathbb{R}^{n_1 \times \cdots \times n_d}, \mathbb{R}^{n_1 \times \cdots \times n_d})$, which we assume to be a symmetric operator for simplicity. There are several options to do this:

- Similarly to the Newton-Schulz method, one can apply standard iterative methods such as the preconditioned conjugate gradient (PCG) method [75] formulated on the space $\mathbb{R}^{n_1 \times \cdots \times n_d}$ but implemented with low-rank tensors only and round the iterates in between to avoid infeasible rank growth. See also [11] for a review of such methods, and [9, 10, 13] for concrete examples. To obtain provable convergence of such an inexact scheme, the errors caused by truncation typically have to be controlled suitably so that a guaranteed rank bound for the iterates of the method cannot be provided even though the solution $\mathbf{x}$ of the system may have low rank. If a suitable soft thresholding procedure is used for rounding to a fixed tensor rank, convergence results of fixed point methods can be established [11].

- Alternative approaches guess the rank $r$ of the solution $\mathbf{x}$ and formulate the equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ as a least squares problem, which is then solved on the set of tensors of rank at most $r$ by optimization over the representing tensors. In the case of the TT format, $\mathbf{x} = \mathcal{TT}(\mathbf{u}_1, \ldots, \mathbf{u}_d)$ holds and one would solve

$$\min_{\mathbf{u}} \tfrac{1}{2} \|\mathbf{A}\,\mathcal{TT}(\mathbf{u}) - \mathbf{b}\|_{\mathsf{F}}^2.$$

  Since the map $\mathbf{u} \mapsto \mathcal{TT}(\mathbf{u})$ is only multilinear and therefore *nonlinear* in $\mathbf{u}$, this is a general nonlinear and especially non-quadratic problem. It can be solved locally by block-wise optimization over the TT cores for example. If all cores except $\mathbf{u}_i$ are fixed, a quadratic function in $\mathbf{u}_i$ is obtained, a stationary point of which can be found by solving a linear system. This is known as alternating linear scheme (ALS) [62, 98]. The idea of block-wise optimization can also be applied to more general, nonlinear optimization problems in tensor space.

- When optimizing over the TT cores, the problem appears that the TT representation of a rank-$r$-tensor is not unique. In fact, certain rescalings by orthogonal matrices do not change the represented tensor [98]. Along those rescalings, the objective function will be constant and its curvature will be zero so that typical convergence results requiring second-order sufficient conditions cannot be applied. One way to overcome this issue is to factor out the ambiguity of the representation and work on the manifold of tensors of fixed rank $r$. Riemannian optimization algorithms on manifolds are well understood [1] and were also developed especially for low-rank tensor manifolds [100, 73, 74, 104, 32, 113].

- Estimating the rank of a solution of a linear system is a very hard task in general. Therefore, ALS or methods operating on a fixed manifold may not give accurate enough results. Modifications of ALS such as MALS [62], DMRG [87, 90], or AMEn [36,

35] allow the adaption of ranks by essentially optimizing over the contraction of two "neighboring" TT cores at once and then splitting this optimized tensor into two cores again as described above when discussing the conversion from TT to HT tensors. At this point, a rank can be chosen such that, e. g., a prescribed accuracy is met.

Initially, we used the HT PCG method in our implementations with an adequate rank-1-preconditioner, which is a viable approach. In numerical experiments we found that AMEn yielded more accurate results with smaller ranks and better computing times [46]. One has to mention that—to our knowledge—the convergence theory of this method is still limited, but the method performs well in practice. In principle, it supports operators $\mathbf{A}$ in the `tt_matrix` format and was extended to operators in the so-called {d, R}-format in [35], which is essentially the CP format with sparse matrices. We extended it further to also have componentwise multiplication operators[3] and a rank-1-preconditioner in decomposed form, i. e., a rank-1-operator $\mathbf{T} \approx \mathbf{A}^{-1}$, which has to be given in the form $\mathbf{T} = \bigotimes_{i=1}^{d} \mathsf{P}_i \mathsf{P}_i^*$, where $\mathsf{P}_i \in \mathcal{L}(\mathbb{R}^{n_i}, \mathbb{R}^{n_i})$. There, we also allowed for function handles $\mathsf{R}_i$ so that also, e. g., inverses of triangular matrices or certain linear transformations, such as the fast Fourier transform, could be implemented efficiently.

### Determination of Maximum/Minimum Entries

For determining stepsizes or feasibility in constrained optimization, it can be necessary to compute or estimate the maximum or minimum entry of a given low-rank tensor. In [43] it is proposed to use the fact that the linear mapping $\mathbf{y} \mapsto \mathbf{x} \odot \mathbf{y}$ has exactly the components of $\mathbf{x}$ as eigenvalues. A normed basis of eigenvectors is given by the tensors $\bigotimes_{i=1}^{d} \mathsf{e}_{k_i}^i$ ($k \in \bigtimes_{i=1}^{m} [n_i]$), where $\mathsf{e}_{k_i}^i \in \mathbb{R}^{n_i}$ denotes the $k_i$-th unit vector. The element of $\mathbf{x}$ with maximum absolute value can thus be found by performing a power iteration with initial tensor $\mathbf{y}_0 := \frac{1}{\sqrt{n_1 \cdots n_d}} \mathbb{1}$. The iterates are given by $\mathbf{y}_j = \frac{1}{\|\mathbf{x}.^j\|_\mathsf{F}} \mathbf{x}.^j$. Hence, the iteration can be sped up by squaring each iterate so that only $\frac{1}{\|\mathbf{x}.^{2^j}\|_\mathsf{F}} \mathbf{x}.^{2^j}$ is computed for each $j \in \mathbb{N}$. This is implemented in the example section of `htucker`. The maximum entry of $\mathbf{x}$ is approximated by the sequence $\lambda_j := \frac{\langle \mathbf{x} \odot \mathbf{x}.^{2^j}, \mathbf{x}.^{2^j} \rangle}{\langle \mathbf{x}.^{2^j}, \mathbf{x}.^{2^j} \rangle}$ (Rayleigh quotient). It can be shown that $|\lambda_j| \leq \max_k |\mathbf{x}(k)|$ holds for all $j$. Since the ranks square in each iteration, truncation is necessary. Therefore, one can only expect to be able to compute a lower bound on the absolute value of the maximum entry with this procedure. Therefore, we also experimented with an upper bound, which can be obtained from this iteration: It holds $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_{2^j} = \langle \mathbf{x}.^{2^{j-1}}, \mathbf{x}.^{2^{j-1}} \rangle^{2^{-j}}$ for all $j \in \mathbb{N}$ and $\lim_{j \to \infty} \|\mathbf{x}\|_{2^j} = \|\mathbf{x}\|_\infty$. The approximation $\|\mathbf{x}\|_{2^j}$ can be computed from the repeatedly squared tensor if one handles arithmetic overflow suitably.

Numerical experiments [46] showed us that rounding can be a severe issue in this iteration so that we did not obtain accurate enough results. For tensors describing a function on a product set, we therefore chose the multilevel coordinate search (MCS) method [65] a non-rigorous global optimization routine for box-constrained problems where only function values are available, to optimize the represented function globally to approximate the maximum element of the tensor. This approach yielded better results in practice but is limited to function-related tensors.

[3] The implementation of this task was essentially done by Prof. Dr. Michael Ulbrich.

## 2.2. Tensor Products of Hilbert and Banach Spaces

In this subsection we introduce the basic tools and results for tensor products of Hilbert spaces, following [54] and [116, Sec. 3.4]. We will need this concept to formulate parametric PDE problems. Note that we only collect the most important facts about tensor products of Hilbert spaces and point out only a few generalizations to Banach spaces as far as it is necessary for the functional analytic setting of the semilinear, elliptic PDE with uncertain inputs discussed in Chapter 3.

**Definition 2.5** (Algebraic and topological tensor product of vector spaces[4])**.** Let $V$ and $W$ be vector spaces over a field $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. The *algebraic tensor product space* $V \otimes_a W$ of $V$ and $W$ is defined as the quotient vector space

$$V \otimes_a W := \operatorname{span}\{(v, w) : v \in V, w \in W\}/\mathcal{N}$$

with

$$\mathcal{N} := \operatorname{span}\Big\{\sum_{i=1}^{m}\sum_{j=1}^{n}\alpha_i\beta_j(v_i, w_j) - \big(\sum_{i=1}^{m}\alpha_i v_i, \sum_{j=1}^{n}\beta_j w_j\big) : \\ m, n \in \mathbb{N}, \alpha_i, \beta_j \in \mathbb{K}, v_i \in V, w_j \in W\Big\},$$

where the span of the set is taken in the algebraic sense, meaning that it consists of *finite* formal linear combinations of pairs $(v_i, w_i)$. This first set of formal linear combinations is normally referred to as the *free vector space* over the set $V \times W$. The equivalence class of a pair $(v, w)$ is denoted by $v \otimes w$ and called *elementary tensor*.
If a norm $||\cdot||$ is given on $V \otimes_a W$, the *topological tensor space* is defined as the closure of the algebraic tensor space with respect to this norm:

$$V \otimes_{||\cdot||} W := \overline{V \otimes_a W}^{||\cdot||}.$$

**Proposition 2.6** (Characterization of equivalence classes[5])**.** *Let $V$ and $W$ be vector spaces over a field $\mathbb{K}$ and let $v, \tilde{v} \in V$ and $w, \tilde{w} \in W$ be given. Consider the algebraic tensor space $V \otimes_a W$. The pair $(\tilde{v}, \tilde{w})$ belongs to the equivalence class $v \otimes w$ if and only if the following holds:*

$$(v = 0 \vee w = 0) \wedge (\tilde{v} = 0 \vee \tilde{w} = 0) \tag{2.1}$$

*or*

$$(v \neq 0 \wedge w \neq 0) \vee (\tilde{v} \neq 0 \wedge \tilde{w} \neq 0) \quad \text{and} \quad \exists\, c \in \mathbb{K}\backslash\{0\} \text{ s. t. } \tilde{v} = cv \text{ and } w = c\tilde{w}.$$

*Case* (2.1) *holds exactly for pairs belonging to the equivalence class* $0 \otimes 0$.

*Proof.* The proof of this proposition is given in the appendix (Proposition A.1). □

[4]This definition follows [116, Sec. 3.4] and [54, Sec. 3.2.2].
[5]This proposition follows [116, Exercise 3.11].

**Remark 2.7.** More basic facts and notations about tensor spaces are given in [54, Sec. 3.2]. We collect only some of them:

- Tensor spaces are again vector spaces [54, Def. 3.9].
- It holds that $V \otimes_a W = \text{span}\{v \otimes w : v \in V, w \in W\}$ [54, Eq. (3.11)].
- The tensor product of two finite-dimensional vector spaces $V$ and $W$ is again finite-dimensional. We have $V \otimes_a W = V \otimes_{||\cdot||} W =: V \otimes W$ [54, Notation 3.8(a)].
- The map $\otimes : V \times W \to V \otimes W$ is bilinear [54, Lem. 3.10].
- If $B$ is a basis of $V$ and $C$ is a basis of $W$, then $\{b \otimes c : b \in B, c \in C\}$ is a basis of $V \otimes_a W$. We obtain $\dim(V \otimes_a W) = \dim(V)\dim(W)$ [54, Lem. 3.11].
- If $V$, $W$ and $X$ are vector spaces over $\mathbb{K}$ and a bilinear map $\otimes : V \times W \to X$ fulfills that $\text{span}\{v \otimes w : v \in V, w \in W\} = X$ and that sets $B \subset V$ and $C \subset W$ of linearly independent vectors are mapped to a set $\{b \otimes c : b \in B, c \in C\} \subset X$ of linearly independent vectors, then $X$ is isomorphic to $V \otimes_a W$ [54, Prop. 3.12].

**Example 2.8.** Let $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ and let $S$ and $T$ be some suitable sets, e. g., $[n]$ for some $n \in \mathbb{N}$ to represent the vector space $\mathbb{K}^n$, $\mathbb{N}$ to represent vector spaces of sequences or some set $S \subset \mathbb{R}^n$ to represent more general function spaces. Let $V \subset \{v : S \to \mathbb{K}\}$ and $W \subset \{w : T \to \mathbb{K}\}$ be vector spaces of $\mathbb{K}$-valued functions on the given sets. Now we can consider the algebraic tensor product space $V \otimes_a W$ and make the identification of elementary tensors $v \otimes w \cong vw$, i. e., we say that the elementary tensor $v \otimes w$ is a pointwise product of functions. We get an isomorphism between the sets $X_1 := V \otimes_a W$ and $X_2 := \text{span}\{vw, v \in V, w \in W\} \subset \{x : S \times T \to \mathbb{K}\}$ with the usual equality that $x = y$ for $x, y \in X_2$ if and only if $x(s,t) = y(s,t)$ for all $s \in S$ and all $t \in T$. Therefore, we will frequently use the outer product $\otimes : V \times W \to V \otimes_a W, (v \otimes w)(s,t) := v(s)w(t)$ for function-related tensors, see also [34, Chap. I, Sec. 2.4].
Analogously, if $U$ is a Banach space over $\mathbb{K}$ and we consider $U \otimes_a V$, we can identify $u \otimes v \cong u\, v(\cdot)$, i. e.,

$$U \otimes_a V \cong \{y : S \to U \text{ s. t. } \exists\, n \in \mathbb{N}, (u_k, v_k) \in U \times V \text{ with } y(s) = \sum_{k=1}^{n} v_k(s) u_k \ \forall\, s \in S\},$$

see [34, Chap. I, Sec. 2.4]). This means that we can identify elements of the tensor product space $U \otimes_a V$ by $U$-valued functions on the set $S$.

We briefly discuss how linear maps from one tensor space to another can look like. It is important to observe that linear maps are uniquely determined by their action on a basis, in our case on certain elementary tensors. When having linear maps $\varphi : V \to W$ and $\psi : \tilde{V} \to \tilde{W}$ between $\mathbb{K}$-vector spaces $V$, $\tilde{V}$, $W$, and $\tilde{W}$, the *tensor product of these maps* is defined by

$$(\varphi \otimes \psi)(v \otimes \tilde{v}) := \varphi(v) \otimes \psi(\tilde{v}). \tag{2.2}$$

This tensor product is an elementary tensor in the space $\mathcal{L}(V,W) \otimes_a \mathcal{L}(\tilde{V},\tilde{W})$. It holds that $\mathcal{L}(V,W) \otimes_a \mathcal{L}(\tilde{V},\tilde{W}) \subset \mathcal{L}(V \otimes_a \tilde{V}, W \otimes_a \tilde{W})$ with equality if the spaces $V$ and $\tilde{V}$ are finite dimensional [54, Prop. 3.49].

Let now $V$ and $W$ be Hilbert spaces over $\mathbb{K} \in \{\mathbb{R},\mathbb{C}\}$ with the inner products $(\cdot,\cdot)_V$ and $(\cdot,\cdot)_W$, respectively. Elements of their algebraic tensor product $X_a = V \otimes_a W$ can be written as finite linear combination $x = \sum_{i=1}^m \alpha_i\, v_i \otimes w_i$ of elementary tensors. Therefore, an inner product on the algebraic tensor product space can be defined in a natural way [116, Sec. 3.4]: For $x = \sum_{i=1}^m \alpha_i\, v_i \otimes w_i \in X_a$ and $y = \sum_{j=1}^n \beta_j\, \tilde{v}_j \otimes \tilde{w}_j \in X_a$ the inner product is given by

$$(x,y)_X := \sum_{i=1}^m \sum_{j=1}^n \overline{\alpha_i}\beta_j (v_i,\tilde{v}_j)_V\, (w_i,\tilde{w}_j)_W, \tag{2.3}$$

where $\overline{\alpha_i}$ denotes the complex conjugate of $\alpha_i$. It can be shown that this is well-defined meaning that the value of the inner product does not depend on the representation of $x$ and $y$ and that this defines a scalar product, the so-called *induced scalar product* [54, Lem. 4.124]. The completion $X$ of $X_a$ with respect to the norm induced by this scalar product is called the *complete tensor product* of the Hilbert spaces $V$ and $W$ and is itself a Hilbert space. We denote it by $V \otimes W$.

We collect some facts about the complete tensor product $V \otimes W$ of two Hilbert spaces $V$ and $W$ and the induced inner product $(\cdot,\cdot)_{V\otimes W}$:

- If $B$ is an orthonormal basis of $V$ and $C$ is an orthonormal basis of $W$, the set $\{b \otimes c : b \in B, c \in C\}$ is an orthonormal basis of $V \otimes W$ [116, Thm. 3.12(b)].
- The norm $||x||_{V\otimes W} := \sqrt{(x,x)_{V\otimes W}}$ is a *reasonable crossnorm* [54, Def. 4.66, Lem. 4.67], i. e., $||v \otimes w||_{V\otimes W} = ||v||_V ||w||_W$ holds for all $v \in V$, $w \in W$ and $||\varphi \otimes \psi||_{(V\otimes W)^*} = ||\varphi||_{V^*}||\psi||_{W^*}$ holds for all $\varphi \in V^*$, $\psi \in W^*$ [54, Prop. 4.127].
- As a Hilbert space, the space $V \otimes W$ is reflexive. Its dual space is $(V \otimes W)^* = V^* \otimes W^*$ [54, Lem. 4.75].
- The norm $||\cdot||_{V\otimes W}$ is a *uniform crossnorm* [54, Def. 4.77], i. e., for any operators $A \in \mathcal{L}(V,V)$ and $B \in \mathcal{L}(W,W)$, the operator $A \otimes B$ belongs to $\mathcal{L}(V \otimes W, V \otimes W)$ and has the operator norm $||A \otimes B||_{\mathcal{L}(V\otimes W, V\otimes W)} = ||A||_{\mathcal{L}(V,V)}||B||_{\mathcal{L}(W,W)}$ [54, Prop. 4.127]. The same holds for the norm $||\cdot||_{(V\otimes W)^*}$, which is also a uniform and reasonable crossnorm [54, Prop. 4.80].

Note that we can extend all statements in this section to tensor products $\bigotimes_{i=1}^m V_i$ of $m \in \mathbb{N}$ vector or Hilbert spaces $V_i$, since the tensor product is associative. $\otimes : \bigtimes_{i=1}^m V_i \to \bigotimes_{i=1}^m V_i$ is then a multilinear map and the properties hold in an analogous sense, see also [54, Sec. 3.2.4].

**Example 2.9** (Tensor product of $L^2$-spaces)**.** Let $m \in \mathbb{N}$ open sets $\Xi_i \subset \mathbb{R}$ ($i \in [m]$) be given and equipped with the Borel $\sigma$-algebras and given probability measures $\mathbb{P}_i$, respectively. We consider the spaces $L^2_{\mathbb{P}_i}(\Xi_i)$ of $\mathbb{R}$-valued random variables with finite variance and their tensor product $\bigotimes_{i=1}^m L^2_{\mathbb{P}_i}(\Xi_i)$. Equipped with the scalar product $(v,\tilde{v})_{L^2_{\mathbb{P}_i}(\Xi_i)} := \int_{\Xi_i} v\tilde{v}\, \mathrm{d}\mathbb{P}_i$ these spaces are Hilbert spaces. According to Example 2.8 we identify elementary tensors

by the product of univariate functions: $(v^1 \otimes v^2 \otimes \ldots \otimes v^m)(\xi) := v^1(\xi_1)v^2(\xi_2)\cdots v^m(\xi_m)$ for every $\xi \in \Xi := \bigtimes_{i=1}^m \Xi_i \subset \mathbb{R}^m$. We compute the inner product of two elementary tensors:

$$\Big(\bigotimes_{i=1}^m v^i, \bigotimes_{i=1}^m \tilde{v}^i\Big)_{\bigotimes_{i=1}^m L^2_{\mathbb{P}_i}(\Xi_i)} = \prod_{i=1}^m \big(v^i, \tilde{v}^i\big)_{L^2_{\mathbb{P}_i}(\Xi_i)} = \prod_{i=1}^m \Big(\int_{\Xi_i} v^i \tilde{v}^i \,\mathrm{d}\mathbb{P}_i\Big) = \int_\Xi \bigotimes_{i=1}^m v^i \bigotimes_{i=1}^m \tilde{v}^i \,\mathrm{d}\mathbb{P}.$$

This holds due to Fubini's theorem and $\mathbb{P} := \otimes_{i=1}^m \mathbb{P}_i$ is the product measure of the measures $\mathbb{P}_i$ ($i \in [m]$). Due to linearity we see that this scalar product is exactly the $L^2_{\mathbb{P}}(\Xi)$-inner product $(\mathbf{v}, \tilde{\mathbf{v}})_{L^2_{\mathbb{P}}(\Xi)} = \int_\Xi \mathbf{v}\tilde{\mathbf{v}} \,\mathrm{d}\mathbb{P}$. Since any algebraic tensor has a finite norm in $L^2_{\mathbb{P}}(\Xi)$, we get $\bigotimes_{i=1}^m L^2_{\mathbb{P}_i}(\Xi_i) \subset L^2_{\mathbb{P}}(\Xi)$ with the used identification of the tensor product using that the closure of a subset of a set is contained in its closure. We want to show that the sets are in fact equal. Note that the simple functions on $\Xi$ that are obtained as finite linear combinations of tensor products of simple functions in one variable are dense in $L^2_{\mathbb{P}}(\Xi)$. Since they are also square integrable and therefore contained in the tensor product space ${}_a\bigotimes_{i=1}^m L^2_{\mathbb{P}_i}(\Xi_i)$, this algebraic tensor product is actually dense in $L^2_{\mathbb{P}}(\Xi)$ and its closure therefore equal to $L^2_{\mathbb{P}}(\Xi)$, cf. [34, Chap. I, Sec. 7].

This example raises the question if it is generalizable to $L^p$-spaces with $p \in [1, \infty]$. If it comes to tensor products of general Banach spaces instead of Hilbert spaces, the construction (2.3) is no longer valid since no inner product exists and a similar construction with norms is not possible. Instead, one has to define a suitable norm on the algebraic tensor product and complete it w. r. t. this norm. In general, one can always define the so-called *projective norm* [54, Sec. 4.2.4], which is the strongest norm[6] on $V \otimes_a W$ ensuring continuity of the tensor product mapping $V \times W \ni (v, w) \mapsto v \otimes w \in V \otimes_a W$, and the *injective norm* [54, Sec. 4.7.2], which is the weakest norm on $V \otimes_a W$ yielding a dual norm on $(V \otimes_a W)^*$ rendering the tensor product mapping $V^* \times W^* \ni (\varphi, \psi) \mapsto \varphi \otimes \psi \in V^* \otimes_a W^* \subset (V \otimes_a W)^*$ continuous. However, in the case of $L^p$ spaces, it is more suitable to work with the $L^p$ norm on the product set:

**Example 2.10** (Tensor product of $L^p$-spaces)**.**[7] Let $p \in [1, \infty)$ and consider the spaces $L^p_{\mathbb{P}_i}(\Xi_i)$ of $\mathbb{R}$-valued, $p$-integrable random variables and the algebraic tensor product space ${}_a\bigotimes_{i=1}^m L^p_{\mathbb{P}_i}(\Xi_i)$ in analogy to Example 2.9. Identifying functions in this space with functions on $\Xi$, we equip ${}_a\bigotimes_{i=1}^m L^p_{\mathbb{P}_i}(\Xi_i)$ with the $L^p_{\mathbb{P}}(\Xi)$ norm. Indeed, this norm is finite for any algebraic tensor since it is finite for any elementary tensor:

$$\Big\|\bigotimes_{i=1}^m v^i\Big\|^p_{L^p_{\mathbb{P}}(\Xi)} = \int_\Xi \Big|\prod_{i=1}^m v^i\Big|^p \,\mathrm{d}\mathbb{P} = \prod_{i=1}^m \int_{\Xi_i} |v^i|^p \,\mathrm{d}\mathbb{P}_i = \prod_{i=1}^m \|v^i\|^p_{L^p_{\mathbb{P}_i}(\Xi_i)} < \infty.$$

Therefore, we can complete the algebraic tensor space w. r. t. to the $L^p_{\mathbb{P}}(\Xi)$ norm to obtain $\bigotimes_{i=1}^m L^p_{\mathbb{P}_i}(\Xi_i) \subset L^p_{\mathbb{P}}(\Xi)$. By the simple functions argument from Example 2.9 we even obtain equality of the sets. Note that this does not work any longer for $p = \infty$ so that in general $\bigotimes_{i=1}^m L^\infty_{\mathbb{P}_i}(\Xi_i) \subsetneq L^\infty_{\mathbb{P}}(\Xi)$ holds, see [34, Chap. I, Exercise 7.2].

[6] See [54, Prop. 4.46] for the precise meaning of this statement.
[7] cf. [54, Example 4.40] and [34, Chap. I, Sec. 7]

Switching back to Hilbert spaces, we introduce a space, which is typically used for the functional analytic setting of linear, elliptic PDEs with uncertain coefficients.

**Example 2.11** (An anisotropic Sobolev space)**.** Let for some $n \in \mathbb{N}$ an open domain $\Omega \subset \mathbb{R}^n$ be given and consider the Sobolev space $H^1(\Omega)$ of weakly differentiable $L^2(\Omega)$ functions with $L^2(\Omega)$ derivatives. Moreover, let $(\Xi, \mathcal{F}, \mathbb{P})$ be a probability space, cf. Example 2.9. Now we consider the tensor product $H^1(\Omega) \otimes L^2_{\mathbb{P}}(\Xi)$. Again, we compute the induced inner product of two elementary tensors with $v, \tilde{v} \in H^1(\Omega)$, $w, \tilde{w} \in L^2_{\mathbb{P}}(\Xi)$:

$$\begin{aligned}(v \otimes w, \tilde{v} \otimes \tilde{w})_{H^1(\Omega)\otimes L^2(\Xi)} &= \Big(\int_\Omega v\,\tilde{v} + \nabla v \cdot \nabla \tilde{v}\,\mathrm{d}x\Big)\Big(\int_\Xi w\tilde{w}\,\mathrm{d}\mathbb{P}\Big) = \\ &= \int_\Xi \int_\Omega (v \otimes w)\,(\tilde{v} \otimes \tilde{w}) + \nabla_x(v \otimes w) \cdot \nabla_x(\tilde{v} \otimes \tilde{w})\,\mathrm{d}x\,\mathrm{d}\mathbb{P}\end{aligned}$$

Similarly as above this extends to the whole space, which can also be characterized analogously to an anisotropic Sobolev space [54, Sec. 4.2.3]. We identify $H^1(\Omega) \otimes L^2_{\mathbb{P}}(\Xi) = \{\boldsymbol{y} \in L^2(\Omega \times \Xi) : \partial_{x_i} \boldsymbol{y} \in L^2(\Omega \times \Xi)\ \forall\, i \in [n]\}$. The partial derivatives w. r. t. the space variables are weak derivatives. This space is also isomorphic to the Bochner space $L^2_{\mathbb{P}}(\Xi; H^1(\Omega))$ of $H^1(\Omega)$-valued random variables with finite variance [102, Thm. B.17], [34, Chap. 1, Sec. 7].

Again, we want to extend this example to Bochner spaces with general $p \in [1, \infty)$ to obtain the space of choice for the semilinear, elliptic PDE with uncertain coefficients discussed in Chapter 3. For a given Banach space $Y$ and a measure space $(\Omega, \mathcal{A}, \mu)$, the Bochner space $L^p_\mu(\Omega; Y)$ ($p \in [1, \infty]$) is the space of equivalence classes of strongly measurable functions $\boldsymbol{y} : \Xi \to Y$ such that the norm

$$\begin{aligned}\|\boldsymbol{y}\|_{L^p_{\mathbb{P}}(\Xi;Y)} &:= \Big(\int_\Omega \|\boldsymbol{y}(\omega)\|_Y^p\,\mathrm{d}\mu(\omega)\Big)^{\frac{1}{p}} \quad (p \in [1, \infty)), \\ \|\boldsymbol{y}\|_{L^\infty_{\mathbb{P}}(\Xi;Y)} &:= \operatorname*{ess\,sup}_{\omega\in\Omega} \|\boldsymbol{y}(\omega)\|_Y\end{aligned}$$

is finite. If $Y$ is a Hilbert space, $L^2_\mu(\Omega; Y)$ is a Hilbert space with inner product

$$(\boldsymbol{y}, \boldsymbol{v})_{L^2_\mu(\Omega;Y)} := \int_\Omega (\boldsymbol{y}(\omega), \boldsymbol{v}(\omega))_Y\,\mathrm{d}\mu(\omega).$$

More information about Bochner spaces can be found in, e. g., [66, Chaps. 1 and 2].

**Example 2.12** (Connection to Bochner spaces)**.** Let $Y$ be a Banach space, let $(\Xi, \mathcal{F}, \mathbb{P})$ be a probability space and let $p \in [1, \infty)$. Consider the algebraic tensor product $Y \otimes_a L^p_{\mathbb{P}}(\Xi)$. As seen in Example 2.8, we can identify an elementary tensor $y \otimes w$ with a $Y$-valued function $\boldsymbol{y}$ on $\Xi$ via $\boldsymbol{y}(\xi) = y\,w(\xi)$. Using this identification, the space $Y \otimes_a L^p_{\mathbb{P}}(\Xi)$ can be endowed with the $L^p_{\mathbb{P}}(\Xi; Y)$-norm, which is finite on the algebraic tensor product:

$$\|y \otimes w\|^p_{L^p_{\mathbb{P}}(\Xi;Y)} = \int_\Xi \|y\|_Y^p |w(\xi)|^p\,\mathrm{d}\mathbb{P} = \|y\|_Y^p\,\|w\|^p_{L^p_{\mathbb{P}}(\Xi)} < \infty.$$

By the argument that simple functions of the form $\sum_{k=1}^n y_k\,1_{A_k}(\cdot)$ with $y_k \in Y$, $A_k \in \mathcal{F}$

are dense in $L^p_{\mathbb{P}}(\Xi; Y)$, see [34, Chap. I, Sec. 7.2], we obtain equality of the completion of the algebraic tensor product and the Bochner space: $Y \otimes L^p_{\mathbb{P}}(\Xi) \cong L^p_{\mathbb{P}}(\Xi; Y)$.

We have seen in (2.2) that tensor products of linear maps $\varphi_i \in \mathcal{L}(V_i, W_i)$ are defined by their action on elementary tensors. By linearity, this definition extends to the algebraic tensor product space: $\bigotimes_{i=1}^m \varphi_i : {}_a\bigotimes_{i=1}^m V_i \to {}_a\bigotimes_{i=1}^m W_i \subset \bigotimes_{i=1}^m W_i$. If $\bigotimes_{i=1}^m \varphi_i$ is bounded, it can be uniquely extended to the space $\bigotimes_{i=1}^m V_i$ as it is defined on a dense subset. Note that it is important to know the norms on the tensor products of spaces to derive continuity of the tensor product of operators even though each operator $\varphi_i$ itself might be bounded.

# 3. A Class of Optimal Control Problems under Uncertainty

In this chapter, we present a class of risk-neutral optimal control problems under uncertainty with a semilinear, elliptic PDE as state equation. A suitable functional analytic setting in a Bochner space is developed such that all appearing functions are well-defined and the optimal control problem admits a solution. The connection to tensor products of Banach spaces as discussed in Example 2.12 is drawn. Furthermore, adjoint formulations of the derivatives of the reduced objective function are derived to enable the efficient, gradient-based solution of problems of this class.

## 3.1. Problem Formulation

Let $\xi \in \mathbb{R}^m$ be a vector of $m \in \mathbb{N}$ independently distributed, real-valued random variables. Note that we do not consider a more general probability space here, but involve directly the finite noise assumption as done in [70]. These finitely many random variables can also originate from a truncated Karhunen-Loève expansion [106, Thm. 11.4] of some random field. Then they are only uncorrelated, and independence has to be assumed or shown additionally; for example, it follows from uncorrelatedness if they are Gaussian. They will act as uncertain parameters in the systems we consider. For each parameter $\xi_i$ ($i \in [m]$) we have a corresponding sample space $\Xi_i \subset \mathbb{R}$ equipped with the Borel $\sigma$-algebra and a probability measure $\mathbb{P}_i$. Due to the independence of the random variables, the random vector $\xi$ gives rise to the product measure $\mathbb{P} := \bigotimes_{i=1}^m \mathbb{P}_i$ on $\Xi := \bigtimes_{i=1}^m \Xi_i$, cf. Example 2.9. We define the mean as $\bar{\xi} := \int_\Xi \xi \,\mathrm{d}\mathbb{P}$. For a function $\boldsymbol{v} : \Xi \to \mathbb{R}$, the expectation is defined by $\mathbb{E}[\boldsymbol{v}] := \int_\Xi \boldsymbol{v} \,\mathrm{d}\mathbb{P}$.

Now let $Y$ and $U$ be Hilbert spaces of deterministic functions and denote by $Y^*$ the dual space of $Y$. We consider parametrized, nonlinear PDEs of the form

$$A(\xi)y(\xi) + N(y(\xi), \xi) = B(\xi)u + b(\xi) \tag{3.1}$$

with strongly measurable functions $A : \Gamma \to \mathcal{L}(Y, Y^*)$, $N : Y \times \Xi \to Y^*$, $B : \Xi \to \mathcal{L}(U, Y^*)$ and $b : \Xi \to Y^*$, cf. [70]. Moreover, $y(\xi) \in Y$ is the parameter-dependent state and $u \in U$ the deterministic control. Writing

$$E(y, u, \xi) := A(\xi)y + N(y, \xi) - B(\xi)u - b(\xi)$$

we have a state equation of the form (1.1). We want to control the system prior to the observation of the parameters, only knowing their distribution. Typically, one wants to minimize an objective function depending on the state and the control, like the squared deviation $J_1(y(\xi), u, \xi) = \frac{1}{2}||Q(\xi)y(\xi) - \hat{q}(\xi)||_H^2$ of the state from a desired state $\hat{q}(\xi) \in H$,

skipping regularization of the control here, where $H$ is a Hilbert space and $Q : \Xi \to \mathcal{L}(Y, H)$ is strongly measurable. This quantity is a random variable. Therefore, we have to incorporate a risk measure into the objective function to handle the uncertainty in the system. This risk measure can be the expectation or the conditional value-at-risk (CVaR) for example.

In [69, 70] the authors view all parameter-dependent quantities in (3.1) as Banach space-valued random variables and formulate the problem in a suitable Bochner space. The state is a function $y : \Xi \to Y$, which is required to have finite variance, i. e., to be contained in the Hilbert space $L^2_{\mathbb{P}}(\Xi; Y)$. They assume the continuous dependence of the solution $y$ on the parameters $\xi$ to be able to use a stochastic collocation approach and to approximate stochastic integrals with sparse grids. In contrast to that we formulate (3.1) in the Bochner space $L^p_{\mathbb{P}}(\Xi; Y)$ and apply a stochastic Galerkin discretization later, where $p \in [2, \infty)$ has to be chosen appropriately, and with the identification $\boldsymbol{y}(x, \xi) = y(\xi)(x)$ as described in Section 2.2. The state space is $\boldsymbol{Y} := L^p_{\mathbb{P}}(\Xi; Y)$ and the image space is identified with $\boldsymbol{Y}^* = L^{p^*}_{\mathbb{P}}(\Xi; Y^*)$ with $p^* := \frac{p}{p-1} \in (1, 2]$. Then, for some $\boldsymbol{b} \in \boldsymbol{Y}^*$, $\boldsymbol{b}(\cdot, \xi)$ belongs to $Y^*$ and $\boldsymbol{b}$ is applied to $\boldsymbol{y} \in \boldsymbol{Y}$ via $\langle \boldsymbol{b}, \boldsymbol{y} \rangle_{\boldsymbol{Y}^*, \boldsymbol{Y}} = \int_\Xi \langle \boldsymbol{b}(\cdot, \xi), \boldsymbol{y}(\cdot, \xi) \rangle_{Y^*, Y} \, \mathrm{d}\mathbb{P}$. Problem (3.1) is formulated equivalently as

$$\boldsymbol{A}\boldsymbol{y} + \boldsymbol{N}(\boldsymbol{y}) = \boldsymbol{B}u + \boldsymbol{b} \tag{3.2}$$

with $\boldsymbol{y} \in \boldsymbol{Y}$, $\boldsymbol{A} \in \mathcal{L}(\boldsymbol{Y}, \boldsymbol{Y}^*)$, $\boldsymbol{N} : \boldsymbol{Y} \to \boldsymbol{Y}^*$, $\boldsymbol{B} \in \mathcal{L}(U, \boldsymbol{Y}^*)$ and $\boldsymbol{b} \in \boldsymbol{Y}^*$:

$$\begin{aligned} (\boldsymbol{A}\boldsymbol{y})(\cdot, \xi) &:= A(\xi)y(\xi), & (\boldsymbol{N}(\boldsymbol{y}))(\cdot, \xi) &:= N(y(\xi), \xi), \\ (\boldsymbol{B}u)(\cdot, \xi) &:= B(\xi)u, & \boldsymbol{b}(\cdot, \xi) &:= b(\xi). \end{aligned}$$

The space of the desired state $\hat{\boldsymbol{q}}$ is $\boldsymbol{H} := L^2_{\mathbb{P}}(\Xi, H)$, $\hat{\boldsymbol{q}}(\cdot, \xi) = \hat{q}(\xi)$, and the operator $\boldsymbol{Q} \in \mathcal{L}(\boldsymbol{Y}, \boldsymbol{H})$ is the tensor version of $Q$, i. e., $(\boldsymbol{Q}\boldsymbol{y})(\cdot, \xi) := Q(\xi)y(\xi)$.

Throughout, we assume that $\hat{\boldsymbol{q}} \in L^2_{\mathbb{P}}(\Xi; H)$ holds. If the operators $A$, $B$, the functional $b$ given after (3.1), and the operator $Q$ are well-defined and in particular regular enough w. r. t. the parameters, meaning that they are $q$-integrable with a large enough $q \in [1, \infty)$ or even essentially bounded, the tensor operators $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{Q}$, and the functional $\boldsymbol{b}$ are also well-defined:

**Proposition 3.1.** *Let $Y$ and $Z$ be Banach spaces and let $\Xi \subset \mathbb{R}^m$ be measurable and equipped with the probability measure $\mathbb{P}$. Let $p_Y, p_Z \in [1, \infty]$, $p_Y \geq p_Z$ be given and define $\boldsymbol{Y} := L^{p_Y}_{\mathbb{P}}(\Xi; Y)$ and $\boldsymbol{Z} := L^{p_Z}_{\mathbb{P}}(\Xi; Z)$). Assume that $A \in L^{r_A}_{\mathbb{P}}(\Xi, \mathcal{L}(Y, Z))$ holds for $r_A \in [1, \infty]$ such that $\frac{1}{r_A} + \frac{1}{p_Y} = \frac{1}{p_Z}$, i. e., $\int_\Xi \|A(\xi)\|^{r_A}_{\mathcal{L}(Y,Z)} \, \mathrm{d}\mathbb{P} \leq C^{r_A} < \infty$ for $r_A < \infty$ and $\|A(\xi)\|_{\mathcal{L}(Y,Z)} \leq C$ for $\mathbb{P}$-a. e. $\xi \in \Xi$ in the case $r_A = \infty$.*
*Then, the operator $\boldsymbol{A}$ defined by $(\boldsymbol{A}\boldsymbol{y})(\cdot, \xi) = A(\xi)y(\xi)$ for all $\xi \in \Xi$ belongs to $\mathcal{L}(\boldsymbol{Y}, \boldsymbol{Z})$ and in particular it holds that $\|\boldsymbol{A}\boldsymbol{y}\|_{\boldsymbol{Z}} \leq C\|\boldsymbol{y}\|_{\boldsymbol{Y}}$ for all $\boldsymbol{y} \in \boldsymbol{Y}$.*

*Proof.* This result is obtained by some standard estimates. In the case $q_Z < \infty$ we have

$$\begin{aligned} \|\boldsymbol{A}\boldsymbol{y}\|^{p_Z}_{\boldsymbol{Z}} &= \int_\Xi \|(\boldsymbol{A}\boldsymbol{y})(\cdot, \xi)\|^{p_Z}_Z \, \mathrm{d}\mathbb{P} = \int_\Xi \|A(\xi)y(\xi)\|^{p_Z}_Z \, \mathrm{d}\mathbb{P} \\ &\overset{\text{Hölder's inequality}}{\leq} \left\| \|A(\cdot)\|^{p_Z}_{\mathcal{L}(Y,Z)} \right\|_{L^{r_A/p_Z}_{\mathbb{P}}(\Xi)} \cdot \left\| \|y(\cdot)\|^{p_Z}_Y \right\|_{L^{p_Y/p_Z}_{\mathbb{P}}(\Xi)} \leq C^{p_Z} \|\boldsymbol{y}\|^{p_Z}_{\boldsymbol{Y}} \end{aligned}$$

using $\frac{p_Z}{r_A} + \frac{p_Z}{p_Y} = 1$. In the case $p_Z = \infty$, it follows that $p_Y = \infty$ and $r_A = \infty$. Thus,

$$\begin{aligned}\|\boldsymbol{Ay}\|_{\boldsymbol{Z}} &= \operatorname*{ess\,sup}_{\xi\in\Xi} \|(\boldsymbol{Ay})(\cdot,\xi)\|_Z = \operatorname*{ess\,sup}_{\xi\in\Xi} \|A(\xi)y(\xi)\|_Z \\ &\leq \big(\operatorname*{ess\,sup}_{\xi\in\Xi} \|A(\xi)\|_{\mathcal{L}(Y,Z)}\big)\,\big(\operatorname*{ess\,sup}_{\xi\in\Xi} \|y(\xi)\|_Y\big) \leq C\|\boldsymbol{y}\|_{\boldsymbol{Y}}.\end{aligned}$$

□

This proposition can be applied to the operator $\boldsymbol{A}$ from (3.2) by setting $Z = Y^*$, $p_Y = p \in [2,\infty)$, and $p_Z = p^* = \frac{p}{p-1} \in (1,2]$. We then need $r_A = \frac{p}{p-2}$ to obtain a bounded linear operator $A \in \mathcal{L}(\boldsymbol{Y}, \boldsymbol{Y}^*)$.

**Remark 3.2.**

- Note that in the case $p = 2$ we require the uniform boundedness of the operators $A(\xi)$ due to $r_A = \infty$; 2-integrability w. r. t. the parameters would not be sufficient.
- The same proposition applies in a similar manner to the operators $\boldsymbol{B}$ and $\boldsymbol{Q}$ and the functional $\boldsymbol{b}$. The operator $Q$, for example, must belong to $L^{2p/(p-2)}_{\mathbb{P}}(\Xi;\mathcal{L}(Y,H))$ for the operator $\boldsymbol{Q} \in \mathcal{L}(\boldsymbol{Y},\boldsymbol{H}))$ to be well-defined since we have $p_Y = p \in [2,\infty)$, $p_H = 2$.
- For the nonlinear operator $\boldsymbol{N}$, a similar estimation cannot be done; therefore it will be part of the assumptions that it is well-defined.

Using the expectation as risk measure, we obtain the regularized optimal control problem

$$\min_{\boldsymbol{y}\in\boldsymbol{Y},u\in U} \boldsymbol{J}(\boldsymbol{y},u) := \tfrac{1}{2}||\boldsymbol{Qy} - \hat{\boldsymbol{q}}||^2_{\boldsymbol{H}} + \tfrac{\gamma}{2}||u||^2_U \quad \text{s. t.} \quad \boldsymbol{E}(\boldsymbol{y},u) = 0, \quad u \in U_{\text{ad}} \tag{3.3}$$

with the definition $\boldsymbol{E}(\boldsymbol{y},u) := \boldsymbol{Ay} + \boldsymbol{N}(\boldsymbol{y}) - \boldsymbol{B}u - \boldsymbol{b}$ and a nonempty, closed and convex set of admissible controls $U_{\text{ad}} \subset U$. Note that the $\boldsymbol{H}$-inner product is the expectation of the $H$-inner product: $(\hat{\boldsymbol{q}},\boldsymbol{v})_{\boldsymbol{H}} = \int_\Xi (\hat{\boldsymbol{q}}(\cdot,\xi),\boldsymbol{v}(\cdot,\xi))_H \,\mathrm{d}\mathbb{P}$. Hence, the objective function can be written as the expectation of some deterministic objective function:

$$\boldsymbol{J}(\boldsymbol{y},u) = \int_\Xi J[\xi](\boldsymbol{y}(\xi),u)\,\mathrm{d}\mathbb{P} \quad \text{with} \quad J[\xi](y,u) := \tfrac{1}{2}\|Q(\xi)y - \hat{q}(\xi)\|^2_H + \tfrac{\gamma}{2}\|u\|^2_U. \tag{3.4}$$

Having

$$J_1(y,u,\xi) := \tfrac{1}{2}\|Q(\xi)y - \hat{q}(\xi)\|^2_H, \quad J_2(u) := \tfrac{\gamma}{2}\|u\|^2_U, \tag{3.5}$$

and $\mathcal{R} \equiv \mathbb{E}$ this fits into the general setting (1.2).

## 3.2. A Class of Semilinear, Elliptic PDEs with Uncertain Coefficients

In this section, we concretize the model problem and discuss a class of parameter-dependent, elliptic PDEs. For this purpose let $\Omega \subset \mathbb{R}^n$ ($n \in \{2,3\}$) be an open, bounded domain with Lipschitz boundary $\partial\Omega$. We choose $Y := H^1_0(\Omega)$ as state space, i. e., the Sobolev space of weakly differentiable $L^2(\Omega)$ functions with $L^2(\Omega)$ derivatives and zero boundary data in the

sense of traces. In the light of Poincaré's inequality [60, Thm. 1.13], we equip this space with the inner product $(v, \tilde{v})_{H_0^1(\Omega)} = \int_\Omega \nabla v \cdot \nabla \tilde{v} \,\mathrm{d}x$, which is not the standard $H^1(\Omega)$-inner product given by $(v, \tilde{v})_{H^1(\Omega)} = \int_\Omega v\,\tilde{v} + \nabla v \cdot \nabla \tilde{v} \,\mathrm{d}x$. $U := L^2(\Omega_u)$ shall be the control space, where $\Omega_u$ can be a measurable subset of $\Omega$ for example or—for finite-dimensional controls—$\Omega_u = [n_u]$ for some $n_u \in \mathbb{N}$. The control will act as distributed control on the system via a linear, bounded operator $D \in \mathcal{L}(L^2(\Omega_u), L^2(\Omega))$. Boundary control with, e. g., $\Omega_u$ being a subset of $\partial\Omega$ can also be handled by the algorithm presented later, but is not discussed here. The parameter-dependence of the PDE is due to an uncertain, space-dependent function $\kappa \in L^\infty(\Omega \times \Xi)$ and due to a parameter-dependent right-hand side offset $f(\xi) \in L^2(\Omega)$. The considered semilinear, elliptic PDE is

$$-\mathrm{div}\,(\kappa(\cdot, \xi)\nabla y) + \varphi(y) = Du + f(\xi) \quad (\text{in } \Omega), \qquad y = 0 \quad (\text{on } \partial\Omega), \tag{3.6}$$

where $\varphi : \mathbb{R} \to \mathbb{R}$ is possibly nonlinear.

**Assumption 3.3.** We require the following additional properties of the problem data:

- The function $\kappa \in L^\infty(\Omega \times \Xi)$ is uniformly bounded and positive, i. e., there exist constants $\underline{\kappa}, \overline{\kappa} \in (0, \infty)$, $\underline{\kappa} \le \overline{\kappa}$, such that $\underline{\kappa} \le \kappa(x, \xi) \le \overline{\kappa}$ holds for almost every $x \in \Omega$ and $\xi \in \Xi$, cf. [44].
- The function $\varphi : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable and monotonically increasing. Its second derivative fulfills the growth condition

  $$|\varphi''(t)| \le a''_{\varphi''} + c''_{\varphi''}|t|^{p-3} \tag{3.7}$$

  for all $t \in \mathbb{R}$ with constants $a''_{\varphi''}, c''_{\varphi''} \ge 0$ and the exponent $p \in (3, \infty)$ if $n = 2$ and $p \in (3, 6]$ if $n = 3$.
- It holds that $f \in L^{r_f}_{\mathbb{P}}(\Xi; L^2(\Omega))$ for some $r_f \in [p, \infty]$ with $p$ from (3.7).
- $D : U \to L^2(\Omega)$ is linear and bounded.

**Remark 3.4.**

- The growth condition (3.7) implies similar conditions for the function $\varphi$ itself and its first derivative by Lemma A.10 with the exponents $p - 1$ and $p - 2$, respectively.
- An additional dependence of $\varphi$ on the parameters $\xi$ and the space variable $x$ could be included easily (see Section A.4), but is skipped here to keep the presentation compact.
- The assumption on $\varphi$ is used to show that the induced Nemytskii operator is well-defined and twice continuously differentiable.

The weak formulation of (3.6) is

$$(\kappa(\cdot, \xi)\nabla y, \nabla v)_{L^2(\Omega)^n} + \int_\Omega \varphi(y)v \,\mathrm{d}x = (Du, v)_{L^2(\Omega)} + (f(\xi), v)_{L^2(\Omega)} \quad \forall\, v \in Y. \tag{3.8}$$

**Proposition 3.5.** *Under Assumption 3.3, the state equation* (3.8) *has a unique solution* $y(\xi) = S[\xi](u) \in Y \cap \mathcal{C}(\overline{\Omega})$. *It satisfies the estimate*

$$\|S[\xi](u)\|_{H_0^1(\Omega)} \leq \tfrac{C_\Omega}{\underline{\kappa}} \|Du + f(\xi) - \varphi(0)\|_{L^2(\Omega)}, \tag{3.9}$$

*where* $C_\Omega$ *is the constant from Poincaré's inequality depending only on* $\Omega$.

*Proof.* Existence is shown in, e. g., [60, Thm. 1.25, Remark 1.12]. The Sobolev embedding $H^1(\Omega) \hookrightarrow L^p(\Omega)$ implies that the nonlinear part of the equation is well-defined. The given a priori estimate can be proven by inserting $v = y$ into (3.8) and using $(\varphi(y) - \varphi(0))y \geq 0$ because of the monotonicity of $\varphi$. □

The map $S[\xi] : U \to Y$ is the parametrized control-to-state mapping. We want to emphasize that the uncertain parameters enter only in the realization $f(\xi)$ of the right-hand side in the estimate (3.9).

With $Y^* = H_0^1(\Omega)^* =: H^{-1}(\Omega)$ and the definitions

$$\begin{aligned} A(\xi) : Y \to Y^*,\ \langle A(\xi)y,\ v\rangle_{Y^*,Y} &= (\kappa(\cdot,\xi)\nabla y, \nabla v)_{L^2(\Omega)^n}, \\ N : Y \times \Xi \to Y^*,\ \langle N(y,\xi),\ v\rangle_{Y^*,Y} &= \int_\Omega \varphi(y)v \,\mathrm{d}x, \\ B(\xi) : U \to Y^*,\ \langle B(\xi)u,\ v\rangle_{Y^*,Y} &= (Du, v)_{L^2(\Omega)}, \\ b(\xi) \in Y^*,\ \langle b(\xi), v\rangle_{Y^*,Y} &= (f(\xi), v)_{L^2(\Omega)} \end{aligned} \tag{3.10}$$

for every $v \in H_0^1(\Omega)$, (3.8) can be written in the form of (3.1). Parameter dependence of the operator $B$ can be handled, but we consider only this fixed form for simplicity reasons.

Now, writing $\boldsymbol{y}(\cdot,\xi) = y(\xi)(\cdot)$, we use the state space $\boldsymbol{Y} = L^p_{\mathbb{P}}(\Xi; H_0^1(\Omega))$. We require $p$ to be chosen according to Assumption 3.3. Especially, if $\varphi$ is a truly nonlinear function, $p > 3$ is required to get a twice continuously differentiable Nemytskii operator. In the linear case $\varphi \equiv 0$, $p = 2$—as also allowed in Section 3.1—is sufficient and a better choice since Hilbert space theory can be applied then, see, e. g., [46]. We write (3.10) as an equation in $\boldsymbol{Y}$, in a weak sense also w. r. t. the parameters: $\boldsymbol{A}\boldsymbol{y} + \boldsymbol{N}(\boldsymbol{y}) = \boldsymbol{B}u + \boldsymbol{b}$ as in (3.2), where

$$\begin{aligned} \boldsymbol{A} : \boldsymbol{Y} \to \boldsymbol{Y}^*,\ \langle \boldsymbol{A}\boldsymbol{y}, \boldsymbol{v}\rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}} &= \int_\Xi (\kappa(\cdot,\xi)\nabla_x \boldsymbol{y}(\cdot,\xi), \nabla_x \boldsymbol{v}(\cdot,\xi))_{L^2(\Omega)^n} \,\mathrm{d}\mathbb{P}, \\ \boldsymbol{N} : \boldsymbol{Y} \to \boldsymbol{Y}^*,\ \langle \boldsymbol{N}(\boldsymbol{y}), \boldsymbol{v}\rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}} &= \int_\Xi \int_\Omega \varphi(\boldsymbol{y}(x,\xi))\boldsymbol{v}(x,\xi) \,\mathrm{d}x \,\mathrm{d}\mathbb{P} \\ \boldsymbol{B} : U \to \boldsymbol{Y}^*,\ \langle \boldsymbol{B}u, \boldsymbol{v}\rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}} &= \int_\Xi (Du, \boldsymbol{v}(\cdot,\xi))_{L^2(\Omega)} \,\mathrm{d}\mathbb{P} \\ \boldsymbol{b} \in \boldsymbol{Y}^*,\ \langle \boldsymbol{b}, \boldsymbol{v}\rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}} &= \int_\Xi (f(\xi), \boldsymbol{v}(\cdot,\xi))_{L^2(\Omega)} \,\mathrm{d}\mathbb{P} \quad \forall\ \boldsymbol{v} \in \boldsymbol{Y}. \end{aligned} \tag{3.11}$$

This equation has a unique solution in $\boldsymbol{Y}$, which we prove using the theory of monotone operators and pointwise considerations. First note that the operators in (3.11) are well-

defined. For the linear operator $\boldsymbol{A}$ this follows from Proposition 3.1, because $A$ belongs to $L^\infty_{\mathbb{P}}(\Xi;\mathcal{L}(Y,Y^*)) \subset L^{p/(p-2)}_{\mathbb{P}}(\Xi;\mathcal{L}(Y,Y^*))$:

$$\|A(\xi)\|_{\mathcal{L}(Y,Y^*)} = \sup_{y,v\in Y,\|y\|_Y=\|v\|_Y=1} (\kappa(\cdot,\xi)\nabla y,\nabla v)_{L^2(\Omega)^n} \le \|\kappa(\cdot,\xi)\|_{L^\infty(\Omega)} \le \overline{\kappa}$$

holds for a. e. $\xi \in \Xi$ by, e. g., Hölder's inequality and the Cauchy-Schwarz inequality on $\mathbb{R}^n$. This gives that $\int_\Xi \|A(\xi)\|^{p/(p-2)}_{\mathcal{L}(Y,Y^*)}\,\mathrm{d}\mathbb{P} \le \overline{\kappa}^{p/(p-2)}$ for any $p \in [2,\infty)$ and especially $p \in (3,\infty)$ by Proposition A.2 and thus that $\boldsymbol{A}$ is bounded: $\|\boldsymbol{A}\|_{\mathcal{L}(\boldsymbol{Y},\boldsymbol{Y}^*)} \le \overline{\kappa}$. Similarly, we compute bounds for the norms

$$\|B(\xi)\|_{\mathcal{L}(U,Y^*)} = \sup_{u\in U,v\in Y,\|u\|_U=1,\|v\|_Y=1} (Du,v)_{L^2(\Omega)} \le \|D\|_{\mathcal{L}(U,L^2(\Omega))}$$

resulting in $\|\boldsymbol{B}\|_{\mathcal{L}(U,\boldsymbol{Y}^*)} \le \|D\|_{\mathcal{L}(U,L^2(\Omega))}$, and

$$\|b(\xi)\|_{Y^*} = \sup_{v\in Y,\|v\|_Y=1} (f(\xi),v)_{L^2(\Omega)} \le \|f(\xi)\|_{L^2(\Omega)}$$

implying $\|\boldsymbol{b}\|_{\boldsymbol{Y}^*} \le \|f\|_{L^{p^*}_{\mathbb{P}}(\Xi;L^2(\Omega))}$ as in Proposition 3.1. As long as $r_f \ge p^*$ holds, which can be deduced from Assumption 3.3, this is bounded by Proposition A.2. Therefore, all appearing linear operators are bounded and also twice continuously differentiable.

**Proposition 3.6.** *Under the conditions on $\varphi$ and $p$ stated in Assumption 3.3, the nonlinear superposition operator $\boldsymbol{N} : \boldsymbol{Y} \to \boldsymbol{Y}^*$ as defined in* (3.11) *is well-defined, monotone and twice continuously Fréchet-differentiable. The derivatives are given by*

$$\langle \boldsymbol{N}'(\boldsymbol{y})\boldsymbol{v},\tilde{\boldsymbol{v}}\rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}} = \int_\Xi \int_\Omega \varphi'(\boldsymbol{y}(x,\xi))\boldsymbol{v}(x,\xi)\tilde{\boldsymbol{v}}(x,\xi)\,\mathrm{d}x\,\mathrm{d}\mathbb{P},$$

$$\langle [\boldsymbol{N}''(\boldsymbol{y})\boldsymbol{v}]\boldsymbol{w},\tilde{\boldsymbol{v}}\rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}} = \int_\Xi \int_\Omega \varphi''(\boldsymbol{y}(x,\xi))\boldsymbol{v}(x,\xi)\boldsymbol{w}(x,\xi)\tilde{\boldsymbol{v}}(x,\xi)\,\mathrm{d}x\,\mathrm{d}\mathbb{P}.$$

*Proof.* This statement is proven in Section A.4. Note that Lemma A.10 shows that Assumption A.11 is satisfied for the function $\varphi$. □

For a discussion of the more concrete case $\varphi(t) = t^3$ and $p = 4$ see Section A.5 in the appendix.

### Unique Solvability of the State Equation

Now we show that the equation

$$\tilde{\boldsymbol{N}}(\boldsymbol{y}) := \boldsymbol{A}\boldsymbol{y} + \boldsymbol{N}(\boldsymbol{y}) = \tilde{\boldsymbol{b}}, \quad \boldsymbol{y} \in \boldsymbol{Y} \tag{3.12}$$

has a unique solution $\boldsymbol{y} \in \boldsymbol{Y}$ for all $\tilde{\boldsymbol{b}} \in \boldsymbol{Y}^*$ with $\tilde{\boldsymbol{b}} = \boldsymbol{B}u + \boldsymbol{b}$, where $u \in U$ is given. For this purpose we use the following theorem, which is obtained by combining Theorem 26.A, Definition 25.2, and Definition 26.1 in [117]:

**Theorem 3.7** (Excerpt from the Minty-Browder theorem about monotone operators)**.** *Let $Y$ be a real, reflexive Banach space. Let $\tilde{N} : Y \to Y^*$ be*

- *strictly monotone, i. e., $\langle \tilde{N}(y) - \tilde{N}(\tilde{y}), y - \tilde{y} \rangle_{Y^*,Y} > 0$ holds for all $y \neq \tilde{y} \in Y$,*
- *coercive, i. e., $\lim_{\|y\|_Y \to \infty} \frac{\langle \tilde{N}(y), y \rangle_{Y^*,Y}}{\|y\|_Y} = \infty$, and*
- *hemicontinuous, i. e., the real function $t \mapsto \langle \tilde{N}(y + t\tilde{y}), v \rangle_{Y^*,Y}$ is continuous on $[0, 1]$ for all $y, \tilde{y}, v \in Y$.*

*Then, the equation*

$$\tilde{N}(y) = b, \quad y \in Y$$

*has a unique solution $y \in Y$ for all right-hand sides $b \in Y^*$. The inverse operator $\tilde{N}^{-1} : Y^* \to Y$ exists and is strictly monotone, i. e., $\langle \tilde{N}^{-1}(b) - \tilde{N}^{-1}(\tilde{b}), b - \tilde{b} \rangle_{Y,Y^*} > 0$ for all $b \neq \tilde{b} \in Y^*$ and bounded. If $\tilde{N}$ is additionally*

- *strongly monotone, i. e., there exists $c > 0$ such that $\langle \tilde{N}(y) - \tilde{N}(\tilde{y}), y - \tilde{y} \rangle_{Y^*,Y} \geq c\|y - \tilde{y}\|_Y^2$ holds for all $y, \tilde{y} \in Y$,*

*the inverse operator $\tilde{N}^{-1}$ is Lipschitz continuous with Lipschitz constant $c^{-1}$.*

**Remark 3.8.** Note that a strongly monotone operator $\tilde{N}$ is always

- monotone, i. e., $\langle \tilde{N}(y) - \tilde{N}(\tilde{y}), y - \tilde{y} \rangle_{Y^*,Y} \geq 0$ for all $y, \tilde{y} \in Y$,
- coercive, and
- strictly monotone.

**Corollary 3.9.** *Under Assumption 3.3, the state equation* (3.8) *has a unique solution $y(\xi) = S[\xi](u) \in Y$ for almost every $\xi \in \Xi$. The parametrized control-to-state mapping $S[\xi] : U \to Y$ is Lipschitz continuous with constant $\frac{C_\Omega \|D\|_{\mathcal{L}(U,L^2(\Omega))}}{\underline{\kappa}}$.*

*Proof.* The operator $A(\xi)$ defined in (3.10) is strongly monotone on $Y = H_0^1(\Omega)$ with constant $\underline{\kappa}$ for a. e. $\xi \in \Xi$. The nonlinear operator $N$ is well-defined by the growth condition (3.7) and the Sobolev embedding $H_0^1(\Omega) \hookrightarrow L^p(\Omega)$, which implies $\varphi(y) \in L^{p/(p-1)}(\Omega)$. It is monotone due to $(\varphi(y(x)) - \varphi(\tilde{y}(x)))(y(x) - \tilde{y}(x)) \geq 0$ for all $x \in \Omega$ and arbitrary $y, \tilde{y} \in Y$ because of the monotonicity of $\varphi$. Thus, the operator $\tilde{N}(\xi) \equiv A(\xi) + N : Y \to Y^*$ is strongly monotone with constant $\underline{\kappa}$. It is continuous (and therefore hemicontinuous) because $A(\xi)$ is bounded and $N$ is continuous as discussed in Section A.3. Hence, Theorem 3.7 can be applied to deduce the statements. □

**Remark 3.10.** Proposition 3.5 provides the same statement as Corollary 3.9 with the additional fact that the solution $y(\xi)$ is continuous on $\overline{\Omega}$.

When trying to verify the prerequisites from Theorem 3.7 for (3.12) it turns out that strict monotonicity and hemicontinuity of $\tilde{\boldsymbol{N}}$ on $\boldsymbol{Y} = L^p_{\mathbb{P}}(\Xi; H_0^1(\Omega))$ can be shown quite simply. Hemicontinuity follows from well-definedness and continuity, and strict monotonicity follows

from the strict monotonicity of $\boldsymbol{A}$ and the monotonicity of $\boldsymbol{N}$. Strict monotonicity of the operator $\boldsymbol{A}$ is equivalent to $\langle \boldsymbol{Ay}, \boldsymbol{y} \rangle_{\boldsymbol{Y}^*, \boldsymbol{Y}} > 0$ for all $\boldsymbol{y} \in \boldsymbol{Y} \setminus \{0\}$. This is true due to

$$\begin{aligned} \langle \boldsymbol{Ay}, \boldsymbol{y} \rangle_{\boldsymbol{Y}^*, \boldsymbol{Y}} &= \int_\Xi (\kappa(\cdot, \xi) \nabla_x \boldsymbol{y}(\cdot, \xi), \nabla_x \boldsymbol{y}(\cdot, \xi))_{L^2(\Omega)^n} \, \mathrm{d}\mathbb{P} \\ &\geq \underline{\kappa} \int_\Xi \|\boldsymbol{y}(\cdot, \xi)\|^2_{H_0^1(\Omega)} \, \mathrm{d}\mathbb{P} \geq 0. \end{aligned}$$

The last term is 0 if and only if $\|\boldsymbol{y}(\cdot, \xi)\|_{H_0^1(\Omega)} = 0$ for almost every $\xi \in \Xi$ such that for $\boldsymbol{y} \neq 0$ we get strict positivity. But the operator $\tilde{\boldsymbol{N}}$ is not coercive on $\boldsymbol{Y}$. This issue can be overcome by considering a larger space endowed with a weaker norm, see Section A.5 for a concrete choice of such space in the case $\varphi(t) = t^3$.

In the general case, we show existence and uniqueness of a solution by constructing it pointwise for a. e. $\xi \in \Xi$.

**Definition 3.11** (Control-to-state mapping)**.** Let $\boldsymbol{S} : U \to \boldsymbol{Y}$ defined by $\boldsymbol{S}(u)(x, \xi) := S[\xi](u)(x)$ for almost every $\xi \in \Xi$, where $S[\xi](u)$ is the weak solution of (3.8), i. e.,

$$\langle A(\xi) S[\xi](u) + N(S[\xi](u), \xi) - \tilde{b}(\xi), v \rangle_{Y^*, Y} = 0 \quad \forall\, v \in Y \tag{3.13}$$

with $\tilde{b}(\xi) := B(\xi)u + b(\xi)$.

We show that $\boldsymbol{S}$ is well-defined. Measurability of the weak solution $\boldsymbol{S}(u)$ w. r. t. $\xi$ can be shown using the following theorem:

**Theorem 3.12.** *Let $A : \Xi \to \mathcal{L}(Y, Y^*)$ be Bochner measurable and such that $A(\xi)$ is strongly monotone with constant $\underline{\kappa} > 0$ (independent of $\xi$) for almost every $\xi \in \Xi$. Let $b : \Xi \to Y^*$ also be Bochner measurable and let $N : Y \to Y^*$ be a deterministic, monotone and continuous operator.*
*Then, the function $\xi \mapsto y(\xi)$, which is defined almost everywhere on $\Xi$, where $y(\xi)$ is the unique solution of the equation $A(\xi)y + N(y) = b(\xi)$, is also Bochner measurable.*

*Proof.* W. l. o. g. we assume $N(0) = 0$. If this is not the case, we can subtract the value $N(0)$ from both sides of the equation without changing the required properties of the nonlinearity and the right-hand side. For the rest of the proof, we consider only random vectors $\xi \in \Xi_{\underline{\kappa}} := \{\xi \in \Xi : A(\xi) \text{ is strongly monotone with constant } \underline{\kappa}\}$. By assumption, the set $\Xi_{\underline{\kappa}} \subset \Xi$ has measure 1. Observe that the equation $A(\xi)y + N(y) = b(\xi)$ has a unique solution $y(\xi)$ fulfilling $\|y(\xi)\|_Y \leq \frac{1}{\underline{\kappa}} \|b(\xi)\|_{Y^*}$ by Theorem 3.7 in analogy to Corollary 3.9 for every $\xi \in \Xi_{\underline{\kappa}}$.

Due to the Bochner measurability, there exist sequences $(A_n)_{n \in \mathbb{N}}$, $A_n : \Xi \to \mathcal{L}(Y, Y^*)$ and $(b_n)_{n \in \mathbb{N}}$, $b_n : \Xi \to Y^*$ of simple functions with $\lim_{n \to \infty} \|A(\xi) - A_n(\xi)\|_{\mathcal{L}(Y, Y^*)} = 0$ for every $\xi \in \Xi_A \subset \Xi$ and $\lim_{n \to \infty} \|b(\xi) - b_n(\xi)\|_{Y^*} = 0$ for every $\xi \in \Xi_b \subset \Xi$, where $\mathbb{P}(\Xi_A) = 1 = \mathbb{P}(\Xi_b)$. Overall, we obtain

$$\lim_{n \to \infty} \|A(\xi) - A_n(\xi)\|_{\mathcal{L}(Y, Y^*)} + \|b(\xi) - b_n(\xi)\|_{Y^*} = 0$$

for all $\xi \in \Xi_{\underline{\kappa}} \cap \Xi_A \cap \Xi_b$, which is a set of measure 1. Due to convergence we can assume w. l. o. g. that also $A_n(\xi)$ is strongly monotone with constant $\frac{\underline{\kappa}}{2}$ for all $\xi \in \Xi_{\underline{\kappa}} \cap \Xi_A \cap \Xi_b$ and all $n \in \mathbb{N}$ by possibly restricting to a subsequence.

Now define the sequence $(y_n)_{n\in\mathbb{N}}$, $y_n : \Xi \to Y$, where $y_n(\xi)$ is the unique solution of the equation $A_n(\xi)y + N(y) = b_n(\xi)$ for $\xi \in \Xi_{\underline{\kappa}} \cap \Xi_A \cap \Xi_b$ and $y_n(\xi) = 0$ for $\xi \in \Xi \setminus (\Xi_{\underline{\kappa}} \cap \Xi_A \cap \Xi_b)$. The elements of this sequence fulfill $\|y_n(\xi)\|_Y \le \frac{2}{\underline{\kappa}}\|b(\xi)\|_{Y^*}$ for all $\xi \in \Xi$. They are also simple functions. If $A_n$ takes $N_A < \infty$ many values on $\mathcal{L}(Y, Y^*)$ and $b_n$ takes $N_b < \infty$ many values on $Y^*$, then $y_n(\xi)$ admits at most $N_A N_b + 1 < \infty$ many values.

Now we show that indeed $\lim_{n\to\infty} \|y_n(\xi) - y(\xi)\|_Y = 0$ holds for almost every $\xi \in \Xi$, namely for all $\xi \in \Xi_{\underline{\kappa}} \cap \Xi_A \cap \Xi_b$. By strong monotonicity of $A(\xi)$ and monotonicity of $N$ we get, skipping the argument $\xi$:

$$\begin{aligned}
0 &\le \underline{\kappa}\|y - y_n\|_Y^2 \\
&\le \langle A(y - y_n), y - y_n\rangle_{Y^*,Y} + \langle N(y) - N(y_n), y - y_n\rangle_{Y^*,Y} \\
&= \langle Ay + N(y), y - y_n\rangle_{Y^*,Y} - \langle Ay_n, y - y_n\rangle_{Y^*,Y} - \langle N(y_n), y - y_n\rangle_{Y^*,Y} \\
&= \langle b, y - y_n\rangle_{Y^*,Y} + \langle (A_n - A)y_n, y - y_n\rangle_{Y^*,Y} - \langle b_n, y - y_n\rangle_{Y^*,Y} \\
&= \langle b - b_n + (A_n - A)y_n, y - y_n\rangle_{Y^*,Y} \\
&\le \big(\|b - b_n\|_{Y^*} + \tfrac{2}{\underline{\kappa}}\|A_n - A\|_{\mathcal{L}(Y,Y^*)}\|b\|_{Y^*}\big)\|y - y_n\|_Y.
\end{aligned}$$

This results in

$$\|y(\xi) - y_n(\xi)\|_Y \le \tfrac{1}{\underline{\kappa}}\|b(\xi) - b_n(\xi)\|_{Y^*} + \tfrac{2}{\underline{\kappa}^2}\|A_n(\xi) - A(\xi)\|_{\mathcal{L}(Y,Y^*)}\|b(\xi)\|_{Y^*}$$

proving the limit $\lim_{n\to\infty} \|y_n(\xi) - y(\xi)\|_Y = 0$ for all $\xi \in \Xi_{\underline{\kappa}} \cap \Xi_A \cap \Xi_b$. Thus, the function $y : \Xi \to Y$ is Bochner measurable. □

**Remark 3.13.** The proof of Theorem 3.12 shows that the solution $y$ of the equation $Ay + N(y) = b$ depends continuously on the operator $A$ and the right hand side $b$.

**Corollary 3.14.** *The function $\xi \mapsto S[\xi](u) \in Y$, where the parametrized control-to-state mapping $S[\xi]$ is implicitly defined by* (3.13)*, is measurable.*

*Proof.* Obviously, all prerequisites from Theorem 3.12 are fulfilled. □

**Corollary 3.15.** *Under Assumption 3.3, the control-to-state mapping from Definition 3.11 is well-defined. It holds that $\boldsymbol{S}(u) \in L^{r_f}(\Xi; Y)$, i. e., $\boldsymbol{S}(u)$ inherits its $\xi$-regularity (integrability or essential boundedness) from $f(\cdot)$.*

*Proof.* By (3.9) we have the estimate

$$\|\boldsymbol{S}(u)(\cdot,\xi)\|_{H_0^1(D)} \le \frac{C_\Omega}{\underline{\kappa}}\|Du + f(\xi) - \varphi(0)\|_{L^2(\Omega)} \quad \text{for a. e. } \xi \in \Xi.$$

Since $f \in L^{r_f}_{\mathbb{P}}(\Xi; L^2(\Omega))$, this shows, together with the Bochner measurability (Corollary 3.14), that $\boldsymbol{S}(u) \in L^{r_f}_{\mathbb{P}}(\Xi; Y) \subset L^p_{\mathbb{P}}(\Xi; Y) = \boldsymbol{Y}$. The inclusion of Bochner spaces holds true by Proposition A.2 if $r_f \ge p$. □

Note that this regularity can be reduced if the function $\xi \mapsto \frac{1}{\underline{\kappa}(\xi)}$ with a parameter-dependent coercivity constant $\underline{\kappa}(\xi) > 0$ for the operator $A(\xi)$ does not belong to $L^\infty_{\mathbb{P}}(\Xi)$. We do not consider this case because we need the uniform coercivity for error estimation later.

**Proposition 3.16.** *Under Assumption 3.3, the function $\boldsymbol{y} = \boldsymbol{S}(u)$ (Definition 3.11) is the unique solution of the equation $\boldsymbol{A}\boldsymbol{y} + \boldsymbol{N}(\boldsymbol{y}) = \boldsymbol{B}u + \boldsymbol{b}$ as defined in* (3.11) *and* (3.12)*.*

*Proof.* The equation can be written as

$$\int_\Xi \langle A(\xi)\boldsymbol{y}(\cdot,\xi) + N(\boldsymbol{y}(\cdot,\xi),\xi) - \tilde{b}(\xi), \boldsymbol{v}(\cdot,\xi)\rangle_{Y^*,Y}\,\mathrm{d}\mathbb{P} = 0 \quad \forall\, \boldsymbol{v} \in \boldsymbol{Y}.$$

The integrand vanishes for almost every $\xi \in \Xi$ if $\boldsymbol{y} = \boldsymbol{S}(u)$ is inserted. Hence, $\boldsymbol{S}(u) \in L^p_{\mathbb{P}}(\Xi; H^1_0(\Omega))$ (Corollary 3.15) is a solution of the equation.

Uniqueness of the solution in $\boldsymbol{Y}$ can be proven as follows: Let $\boldsymbol{y}, \tilde{\boldsymbol{y}} \in \boldsymbol{Y}$ both solve (3.12). Then, subtracting the state equation for $\tilde{\boldsymbol{y}}$ from the one with $\boldsymbol{y}$ and testing with $\boldsymbol{y} - \tilde{\boldsymbol{y}}$ yields

$$0 = \langle \boldsymbol{A}(\boldsymbol{y} - \tilde{\boldsymbol{y}}), \boldsymbol{y} - \tilde{\boldsymbol{y}}\rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}} + \langle \boldsymbol{N}(\boldsymbol{y}) - \boldsymbol{N}(\tilde{\boldsymbol{y}}), \boldsymbol{y} - \tilde{\boldsymbol{y}}\rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}} \geq \underline{\kappa}\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|^2_{L^2_{\mathbb{P}}(\Xi;H^1_0(\Omega))} \geq 0,$$

giving $\boldsymbol{y} = \tilde{\boldsymbol{y}}$ almost everywhere. □

**Example 3.17.** To be more concrete, let us consider a specific form of the coefficient function $\kappa$. We define a mean function $\kappa_0(x)$ and some functions $\eta_i(x)$ $(i \in [m])$ describing the amount of influence of the parameter $\xi_i$ over the domain $\Omega$. We set

$$\kappa(x,\xi) := \kappa_0(x)\Big(1 + \sum_{i=1}^m \xi_i\eta_i(x)\Big).$$

Defining $\kappa_i(x) := \kappa_0(x)\eta_i(x)$ we have $\kappa(x,\xi) = \kappa_0(x) + \sum_{i=1}^n \xi_i\kappa_i(x)$, which is the form for random fields originating from a truncated Karhunen-Loève expansion. If we have $\kappa_0, \eta_i \in L^\infty(\Omega)$ for all $i \in [m]$, $\kappa_0$ uniformly positive ($0 < \underline{\kappa}_0 \leq \kappa_0(x) \leq \overline{\kappa}_0 < \infty$), bounded random variables, i. e., $\exists\, C > 0$ s. t. $\mathbb{P}(\{\xi \in \Xi : |\xi| \leq C\}) = 1$, and $\eta_i$ also suitably bounded (e. g., $|\eta_i(x)| \leq \overline{\eta}_i < \frac{1}{Cm}$ for a. e. $x \in \Omega$ and all $i \in [m]$), then the prerequisites on $\kappa$ from above are fulfilled and we get an $L^2_{\mathbb{P}}(\Xi; H^1_0(\Omega))$-elliptic operator $\boldsymbol{A}$.

With $\kappa$ having the specific form from Example 3.17, the operator $\boldsymbol{A}$ has a useful structure:

**Lemma 3.18.** *The operator $\boldsymbol{A}$ defined in* (3.11) *with $\kappa(x,\xi) = \kappa_0(x)(1 + \sum_{i=1}^m \xi_i\eta_i(x))$ as in Example 3.17 can be written as*

$$\boldsymbol{A} = A_0 \otimes \Big(\bigotimes_{j=1}^m I\Big) + \sum_{i=1}^m A_i \otimes \Big(\bigotimes_{j=1}^m \tilde{S}_{ij}\Big)$$

*with the Kronecker product of operators as in Section 2.2, identifying $\boldsymbol{Y}^* = Y^* \otimes L^{p^*}_{\mathbb{P}_1}(\Xi_1) \otimes \cdots \otimes L^{p^*}_{\mathbb{P}_m}(\Xi_m)$, and with $A_0, A_i \in \mathcal{L}(Y, Y^*)$ and $\tilde{S}_{ij} \in \mathcal{L}(L^p_{\mathbb{P}_j}(\Xi_j), L^{p^*}_{\mathbb{P}_j}(\Xi_j))$ $(i,j \in [m])$ defined by*

$$\langle A_0 y, v\rangle_{Y^*,Y} := (\kappa_0\,\nabla y, \nabla v)_{L^2(\Omega)^n}, \quad \langle A_i y, v\rangle_{Y^*,Y} := (\eta_i\,\kappa_0\,\nabla y, \nabla v)_{L^2(\Omega)^n},$$
$$\tilde{S}_{ij} = I \ \text{ for } i,j \in [m],\ i \neq j, \quad (\tilde{S}_{jj}v_j)(\xi_j) := \xi_j \cdot v_j(\xi_j) \ \text{ for } j \in [m].$$

*Proof.* At first, we split up the parametrized operator $A(\xi)$ using the definitions of $\kappa(\cdot,\xi)$, $A_0$ and $A_i$:

$$\begin{aligned}\langle A(\xi)y, v\rangle_{Y^*,Y} &= (\kappa(\cdot,\xi)\nabla y, \nabla v)_{L^2(\Omega)^n} \\ &= (\kappa_0\nabla y, \nabla v)_{L^2(\Omega)^n} + \sum_{i=1}^m \xi_i(\kappa_0\eta_i\nabla y, \nabla v)_{L^2(\Omega)^n} \\ &= \big\langle \big(A_0 + \sum_{i=1}^m \xi_i A_i\big)y, v\big\rangle_{Y^*,Y}\end{aligned}$$

Next we consider an elementary tensor $\boldsymbol{y} = y \otimes \upsilon_1 \otimes \ldots \otimes \upsilon_m \in \boldsymbol{Y}$, i. e., $\boldsymbol{y}(x,\xi) = y(x) \cdot \prod_{j=1}^m \upsilon_j(\xi_j)$, and compute for fixed $\xi$:

$$\begin{aligned}A(\xi)\boldsymbol{y}(\cdot,\xi) &= \Big(A_0 + \sum_{i=1}^m \xi_i A_i\Big)\Big(y(\cdot)\cdot\prod_{j=1}^m \upsilon_j(\xi_j)\Big) = \\ &= A_0 y(\cdot)\cdot\prod_{j=1}^m \upsilon_j(\xi_j) + \sum_{i=1}^m A_i y(\cdot)\cdot\xi_i\cdot\prod_{j=1}^m \upsilon_j(\xi_j) = \\ &= \Big((A_0 y)\otimes\big(\bigotimes_{j=1}^m \upsilon_j\big)\Big)(\cdot,\xi) + \sum_{i=1}^m \Big((A_i y)\otimes\big(\bigotimes_{j=1}^m \tilde{S}_{ij}\upsilon_j\big)\Big)(\cdot,\xi).\end{aligned}$$

Thus, we can write the operator $\boldsymbol{A}$ as $\boldsymbol{A} = A_0 \otimes \Big(\bigotimes_{j=1}^m I\Big) + \sum_{i=1}^m A_i \otimes \Big(\bigotimes_{j=1}^m \tilde{S}_{ij}\Big)$. Because of linearity and continuity this extends to algebraic and topological tensors $\boldsymbol{y}$ as seen at the end of Section 2.2. The operator $\boldsymbol{A}$ is continuous from $L^p_{\mathbb{P}}(\Xi;Y)$ to $L^{p^*}_{\mathbb{P}}(\Xi;Y^*)$ as already noted. □

The structure of the operator $\boldsymbol{A}$ provided in Lemma 3.18 uses the identification of the state space $\boldsymbol{Y}$ with a tensor product of Banach spaces and is useful because it makes the discrete version of the operator $\boldsymbol{A}$ easily applicable to tensors in low-rank formats that implement $i$-mode matrix multiplication and componentwise sums. Formally, we can write $\boldsymbol{A}\boldsymbol{y} = A_0 \circ_1 \boldsymbol{y} + \sum_{i=1}^m A_i \circ_1 \tilde{S}_{ii} \circ_{i+1} \boldsymbol{y}$ using the application of a linear operator to a certain mode of a tensor, which we have defined only for finite dimensional tensors so that this will make more sense in the discretized setting (Example 6.5). In our work [46] it was already demonstrated that also more difficult operators as resulting from domain parametrizations can be efficiently handled with low-rank tensors. In this thesis we present an adaptive approach and its convergence theory and base our numerical tests only on this example since reliable a posteriori error estimates, which we use for the adaptive solution of the state equation, can be derived in this setting, see Chapter 7.

## 3.3. Existence of a Solution to the Optimal Control Problem

A natural question is whether the optimal control problem (3.3) with the state equation from Section 3.2 admits a solution. Typically, continuity under weak convergence is used in existence proofs. In nonlinear cases, this is often done by means of compact embeddings,

which yield strong convergence (w. r. t. a weaker norm) of weakly converging sequences. In the Bochner space setting considered here, only the continuous embedding $L^p \hookrightarrow L^q$ is available for $p \geq q$ (see Proposition A.2), but this embedding is not compact. Therefore, pointwise considerations for a. e. $\xi \in \Xi$ are used to prove the following existence theorem.

**Theorem 3.19.** *Let Assumption 3.3 hold with $p \in (3, \infty)$ in the case $n = 2$ and $p \in (3, 6)$ in the case $n = 3$. Furthermore, let $\gamma > 0$, $\hat{\boldsymbol{q}} \in L^2_{\mathbb{P}}(\Xi; H)$, $\boldsymbol{Q} \in L^{r_Q}_{\mathbb{P}}(\Xi; \mathcal{L}(Y, H))$ with $\frac{1}{r_Q} + \frac{1}{r_f} = \frac{1}{2}$, and let $Q(\xi) : Y \to H$ be a compact operator for a. e. $\xi$.*
*Then, Problem* (3.3) *with the state equation discussed in Section 3.2 has a solution.*

*Proof.* The statement is proven analogously to [111, Lem. 9.4]. Let $(\boldsymbol{y}_k, u_k)_{k \in \mathbb{N}}$ be a feasible, infimizing sequence, where $\boldsymbol{y}_k = \boldsymbol{S}(u_k)$ is uniquely given due to Proposition 3.16. Because of the regularization term in the objective function with $\gamma > 0$ and the non-negative tracking term, the sequence $(u_k)$ is bounded in $L^2(\Omega_u) = U$. By (3.9), we have for a. e. $\xi \in \Xi$ that the corresponding, pointwise states fulfill

$$\|\boldsymbol{y}_k(\xi)\|_{H^1_0(\Omega)} \leq \tfrac{C_\Omega}{\underline{\kappa}} \big( \|D\|_{\mathcal{L}(U, L^2(\Omega))} \, \|u_k\|_U + \|f(\xi) - \varphi(0)\|_{L^2(\Omega)} \big). \tag{3.14}$$

Note that for a. e. $\xi \in \Xi$ the state equation has a unique solution for *arbitrary* $u$ by Assumption 3.3. By boundedness, and since $U_{\mathrm{ad}}$ is a convex, closed subset of a reflexive Banach space, we can extract subsequences for a. e. $\xi \in \Xi$, again denoted by $(u_k)$ and $(\boldsymbol{y}_k(\xi))$, which converge weakly to some limits $u \in U_{\mathrm{ad}}$ and $\boldsymbol{y}(\xi) \in H^1_0(\Omega)$. The compact embedding $H^1_0(\Omega) \hookrightarrow\hookrightarrow L^p(\Omega)$ implies that the sequences $(\boldsymbol{y}_k(\xi))$ converge strongly in $L^p(\Omega)$ to $\boldsymbol{y}(\xi)$ for a. e. $\xi \in \Xi$. The growth condition and the continuity of $\varphi$ as well as the continuous embedding $L^{p^*}(\Omega) \hookrightarrow H^{-1}(\Omega)$ yield that $\varphi(\boldsymbol{y}_k(\xi)) \to \varphi(\boldsymbol{y}(\xi))$ strongly in $H^{-1}(\Omega)$. It follows that

$$\begin{aligned} &A(\xi)\boldsymbol{y}_k(\xi) + \varphi(\boldsymbol{y}_k(\xi)) \rightharpoonup A(\xi)\boldsymbol{y}(\xi) + \varphi(\boldsymbol{y}(\xi)) \text{ in } H^{-1}(\Omega), \\ &Du_k + f(\xi) \rightharpoonup Du + f(\xi) \text{ in } L^2(\Omega) \subset H^{-1}(\Omega) \end{aligned}$$

so that $\boldsymbol{y}(\xi)$ solves (3.8) for a. e. $\xi$. Thus, we have that $\boldsymbol{y} = \boldsymbol{S}(u) \in L^{r_f}(\Xi; Y)$ by Corollary 3.15 giving that $(\boldsymbol{y}, u)$ is feasible for (3.3). The regularization term of the objective function (3.4) is convex and continuous under strong convergence in $L^2(\Omega_u)$, and thus weakly lower semicontinuous. Furthermore, the tracking term converges: Since $\boldsymbol{y}_k(\xi) \rightharpoonup \boldsymbol{y}(\xi)$ in $Y$ and $Q(\xi)$ maps $Y$ to $H$ compactly, the sequence $(Q(\xi)\boldsymbol{y}_k(\xi))$ converges strongly to $Q(\xi)\boldsymbol{y}(\xi)$ in $H$ for a. e. $\xi$. By continuity, $\|Q(\xi)\boldsymbol{y}_k(\xi) - \hat{q}(\xi)\|^2_H \to \|Q(\xi)\boldsymbol{y}(\xi) - \hat{q}(\xi)\|^2_H$ for a. e. $\xi$ as $k \to \infty$. Because of the boundedness of $\|u_k\|_U$ by some constant $C_u > 0$ and (3.14), we can bound

$$\|Q(\xi)\boldsymbol{y}_k(\xi) - \hat{q}(\xi)\|^2_H \leq \Big( \|Q(\xi)\|_{\mathcal{L}(Y,H)} \tfrac{C_\Omega}{\underline{\kappa}} \big( \|D\|_{\mathcal{L}(U, L^2(\Omega))} \, C_u + \|f(\xi) - \varphi(0)\|_{L^2(\Omega)} \big) + \|\hat{q}(\xi)\|_H \Big)^2$$

for every $k$. The $\xi$-regularities of $Q(\cdot)$, $f(\cdot)$ and $\hat{q}(\cdot)$ are chosen exactly such that the estimate on the right-hand side is integrable w. r. t. $\mathbb{P}$. Therefore, the dominated convergence theorem can be applied to conclude

$$\int_\Xi \|Q(\xi)\boldsymbol{y}_k(\xi) - \hat{q}(\xi)\|^2_H \, \mathrm{d}\mathbb{P} \longrightarrow \int_\Xi \|Q(\xi)\boldsymbol{y}(\xi) - \hat{q}(\xi)\|^2_H \, \mathrm{d}\mathbb{P} \quad \text{as } k \to \infty.$$

It follows from the derived continuity properties of $\boldsymbol{J}$ that $(\boldsymbol{y}, u)$ solves (3.3). □

## 3.4. Derivatives of the Reduced Objective Function

We will apply an adaptive, inexact, gradient-based, nonlinear optimization method to solve problems of type (3.3). This algorithm will be formulated in the function space to be able to apply adaptive solution techniques. Therefore, we need expressions for the derivatives in function space and make the following assumption:

**Assumption 3.20.** The state equation (3.2) has a unique weak solution $\boldsymbol{y} = \boldsymbol{S}(u)$ for every control $u \in U$. The nonlinear operator $\boldsymbol{N} : \boldsymbol{Y} \to \boldsymbol{Z}$ is twice continuously Fréchet-differentiable.

Then we can reduce problem (3.3) to the control and get

$$\min_{u \in U} \ \hat{\boldsymbol{J}}(u) := \boldsymbol{J}(\boldsymbol{S}(u), u) \quad \text{s. t.} \quad u \in U_{\text{ad}}. \tag{3.15}$$

We use the adjoint approach for the computation of the gradient of the reduced objective function. The formal adjoint equation for the adjoint state $\boldsymbol{z} \in \boldsymbol{Y}$ is

$$\boldsymbol{E}_{\boldsymbol{y}}(\boldsymbol{S}(u), u)^* \boldsymbol{z} = -\boldsymbol{J}_{\boldsymbol{y}}(\boldsymbol{S}(u), u)$$

and more concretely

$$\langle \boldsymbol{z}, \boldsymbol{A}\boldsymbol{v} + \boldsymbol{N}'(\boldsymbol{S}(u))\boldsymbol{v} \rangle_{\boldsymbol{Y}, \boldsymbol{Y}^*} = -(\boldsymbol{Q}\boldsymbol{S}(u) - \hat{\boldsymbol{q}}, \boldsymbol{Q}\boldsymbol{v})_{\boldsymbol{H}} \quad \forall\ \boldsymbol{v} \in \boldsymbol{Y}.$$

Typically, one would assume that the partial derivative $\boldsymbol{E}_{\boldsymbol{y}}(\boldsymbol{S}(u), u)$ is boundedly invertible for every $u \in U$. Then, the adjoint equation has a unique solution $\boldsymbol{T}(u)$, the control-to-state mapping $\boldsymbol{S} : U \to \boldsymbol{Y}$ is continuously F-differentiable and the adjoint representation of the first derivative of the reduced objective function is given by

$$\hat{\boldsymbol{J}}'(u) = \boldsymbol{E}_u(\boldsymbol{S}(u), u)^* \boldsymbol{T}(u) + \boldsymbol{J}_u(\boldsymbol{y}(u), u)$$

and in our concrete case

$$\langle \hat{\boldsymbol{J}}'(u), w \rangle_{U^*, U} = \langle \boldsymbol{T}(u), -\boldsymbol{B}w \rangle_{\boldsymbol{Y}, \boldsymbol{Y}^*} + \gamma(u, w)_U \quad \forall\ w \in U$$

or in short notation $\nabla \hat{\boldsymbol{J}}(u) = -\boldsymbol{B}^* \boldsymbol{T}(u) + \gamma u$.

We will see that in our case the operator $\boldsymbol{E}_{\boldsymbol{y}}(\boldsymbol{S}(u), u)$ is not necessarily boundedly invertible. Therefore, we will verify that a pointwisely defined adjoint state is sufficient for the computation of the gradient and discuss this for the example presented beforehand.

### 3.4.1. Discussion of the Example

For the example from Section 3.2, Assumption 3.20 is fulfilled. We have proven that the state equation has a unique weak solution for every control (Proposition 3.16) and that the nonlinear operator is twice continuously differentiable (Proposition 3.6). Next, we show that the formal adjoint equation admits a unique weak solution, which is also constructed pointwise. Again, its regularity w. r. t. $\xi$ is discussed using standard a priori estimates.

First note that the partial derivative $\boldsymbol{E_y}(\boldsymbol{S}(u), u) = \boldsymbol{A} + \boldsymbol{N}'(\boldsymbol{S}(u))$ does not necessarily have a bounded inverse for every $u$. Assume that $\boldsymbol{N}'(\boldsymbol{S}(u)) \equiv 0$, then $\boldsymbol{E_y}(\boldsymbol{S}(u), u) = \boldsymbol{A}$. This operator is indeed boundedly invertible from $L^2_{\mathbb{P}}(\Xi; Y)$ to $L^2_{\mathbb{P}}(\Xi; Y^*)$. For $p > 2$ we have $\boldsymbol{A}(L^p_{\mathbb{P}}(\Xi; Y)) = L^p_{\mathbb{P}}(\Xi; Y^*) \subsetneq L^{p^*}_{\mathbb{P}}(\Xi; Y^*)$. The first equality follows from the fact that $A(\xi) : Y \to Y^*$ is boundedly invertible and $A(\cdot) \in L^\infty_{\mathbb{P}}(\Xi; \mathcal{L}(Y, Y^*))$ as well as $A(\cdot)^{-1} \in L^\infty_{\mathbb{P}}(\Xi; \mathcal{L}(Y^*, Y))$. Hence, $\boldsymbol{A} : \boldsymbol{Y} \to \boldsymbol{Y}^*$ is not surjective and thus not invertible if $p > 2$.

In the deterministic case, the standard adjoint approach [60, Sec. 1.6] can be applied since the operator $A(\xi) + N_y(y, \xi)$ is boundedly invertible for every $y \in Y$ and almost every $\xi \in \Xi$. This follows from Theorem 3.7 using the strong monotonicity of $A(\xi)$ and the monotonicity of $N_y(y, \xi)$ (Lemma A.4). We can therefore conclude that the pointwise adjoint equation

$$A(\xi)z(\xi) + N_y(y(\xi), \xi)z(\xi) = -Q(\xi)^*(Q(\xi)y(\xi) - \hat{q}(\xi)) \text{ for a. e. } \xi \in \Xi \tag{3.16}$$

with $y(\xi) = S[\xi](u)$ has a unique solution $z(\xi) = T[\xi](u)$ for a. e. $\xi \in \Xi$ and that the parametrized control-to-state mapping $u \mapsto S[\xi](u)$ is continuously differentiable by the implicit function theorem. The derivative of the parametric reduced objective function $\hat{J}[\xi](u) := J[\xi](S[\xi](u), u)$, see the definition of $J[\xi]$ in (3.4), is then given by

$$\langle \hat{J}[\xi]'(u), w\rangle_{U^*,U} = \langle T[\xi](u), -B(\xi)w\rangle_{Y^*,Y} + \gamma(u, w)_U \quad \forall\ w \in U. \tag{3.17}$$

For $\boldsymbol{y} \in \boldsymbol{Y}$, we now consider the formal adjoint equation

$$\boldsymbol{Az} + \boldsymbol{N}'(\boldsymbol{y})\boldsymbol{z} = -\boldsymbol{Q}^*(\boldsymbol{Qy} - \hat{\boldsymbol{q}}), \tag{3.18}$$

where we use that the operators $\boldsymbol{A}$ and $\boldsymbol{N}'(\boldsymbol{y})$ are self-adjoint in our example.

**Definition 3.21.** Let $\boldsymbol{T} : U \to L^{p/(p-1)}_{\mathbb{P}}(\Xi; Y)$, $\boldsymbol{T}(u)(x, \xi) := T[\xi](u)(x)$ be defined for almost every $\xi \in \Xi$, where $T[\xi](u)$ is the weak solution of (3.16), i. e.,

$$\langle A(\xi)T[\xi](u) + N_y(S[\xi](u), \xi)T[\xi](u) - \tilde{b}(\xi), v\rangle_{Y^*,Y} = 0 \quad \forall\ v \in Y \tag{3.19}$$

with $\tilde{b}(\xi) := -Q(\xi)^*(Q(\xi)S[\xi](u) - \hat{q}(\xi))$ and $S[\xi]$ as in Definition 3.11.

**Proposition 3.22.** *Assume that Assumption 3.3 holds true and $\boldsymbol{Q} \in L^{2p/(p-2)}_{\mathbb{P}}(\Xi; \mathcal{L}(Y, H))$ as well as $\hat{\boldsymbol{q}} \in L^2_{\mathbb{P}}(\Xi; H)$. Then, the operator $\boldsymbol{T}$ from Definition 3.21 is well-defined.*

*Proof.* Since $z \mapsto N_y(y, \xi)z$ is monotone (Proposition A.4) and $A(\xi)$ is strongly monotone with constant $\underline{\kappa}$, (3.19) has a unique solution $T[\xi](u)$ for a. e. $\xi \in \Xi$ due to Theorem 3.7. The estimate

$$\begin{aligned} \|\boldsymbol{T}(u)(\cdot, \xi)\|_Y &\le \tfrac{1}{\underline{\kappa}}\|Q(\xi)^*(Q(\xi)S[\xi](u) - \hat{q}(\xi))\|_{Y^*} \\ &\le \tfrac{1}{\underline{\kappa}}\|Q(\xi)\|_{\mathcal{L}(Y,H)}\|Q(\xi)S[\xi](u) - \hat{q}(\xi)\|_H \end{aligned} \tag{3.20}$$

can be derived in analogy to (3.9). Therefore, the adjoint state inherits its regularity w. r. t. $\xi$ from the $\xi$-regularity of $Q^*(\cdot)(Q(\cdot)S[\cdot](u) - \hat{q}(\cdot))$. Since $S[\cdot](u) \in L^p_{\mathbb{P}}(\Xi; Y)$ by Corollary 3.15, Hölder's inequality implies that $Q(\cdot)S[\cdot](u) \in L^2_{\mathbb{P}}(\Xi; H)$, hence $Q(\cdot)^*(Q(\cdot)S[\cdot](u) - \hat{q}(\cdot)) \in L^{p/(p-1)}_{\mathbb{P}}(\Xi; Y^*)$.

The function $\xi \mapsto T[\xi](u)$ is measurable due to Theorem 3.12 as long as the mapping $\xi \mapsto N_y(S[\xi](u), \xi) \in \mathcal{L}(Y, Y^*)$ is measurable. This is true because $N_y$ is deterministic and continuous and $\xi \mapsto S[\xi](u)$ is measurable by Corollary 3.14. Measurability and the regularity estimate (3.20) yield $\boldsymbol{T}(u) \in L_{\mathbb{P}}^{p/(p-1)}(\Xi; Y)$. □

**Lemma 3.23.** *Let Assumption 3.3 hold and assume $\boldsymbol{y} \in L_{\mathbb{P}}^{r_f}(\Xi; Y)$ for some $r_f \in [p, \infty]$, $\boldsymbol{Q} \in L_{\mathbb{P}}^{r_Q}(\Xi; \mathcal{L}(Y, H))$ for some $r_Q \in [\frac{2p}{p-2}, \infty]$, and $\hat{\boldsymbol{q}} \in L_{\mathbb{P}}^{r_{\hat{q}}}(\Xi; H)$ for some $r_{\hat{q}} \in [2, \infty]$.*
*Then, $\boldsymbol{T}(u) \in L_{\mathbb{P}}^{r_z}(\Xi; Y)$ (Definition 3.21) holds, where $r_z \in [\frac{p}{p-1}, \infty]$ depends on $r_f$, $r_Q$, and $r_{\hat{q}}$ and is defined as in the column "Exponents" of Table 3.1.*

*Proof.* In Proposition 3.22, $\boldsymbol{T}(u) \in L_{\mathbb{P}}^{p/(p-1)}(\Xi; Y)$ is shown. The regularity exponent is deduced from (3.20). In the same way, the regularity exponents for $\boldsymbol{Qy}$, $\boldsymbol{Qy} - \hat{\boldsymbol{q}}$, and $\boldsymbol{T}(u)$ depicted in Table 3.1 follow from Hölder's inequality. □

**Proposition 3.24.** *Let Assumption 3.3 hold and assume $\boldsymbol{y} \in L_{\mathbb{P}}^{r_f}(\Xi; Y)$ for some $r_f \in [p, \infty]$, $\boldsymbol{Q} \in L_{\mathbb{P}}^{r_Q}(\Xi; \mathcal{L}(Y, H))$ and $\hat{\boldsymbol{q}} \in L_{\mathbb{P}}^{r_{\hat{q}}}(\Xi; H)$, where $r_Q \in [2p, \infty]$ and $r_{\hat{q}} \in [p, \infty]$ fulfill the properties in columns "Estimation 1" and "Estimation 2" of Table 3.1, respectively.*
*Then, the formal adjoint equation* (3.18) *has the unique solution $\boldsymbol{T}(u) \in \boldsymbol{Y}$ (Definition 3.21).*

*Proof.* This can be proven analogously to Proposition 3.16.

Equation (3.18) can be written as

$$\int_\Xi \langle A(\xi)\boldsymbol{z}(\cdot, \xi) + N_y(\boldsymbol{y}(\cdot, \xi), \xi)\boldsymbol{z}(\cdot, \xi) - \tilde{b}(\xi), \boldsymbol{v}(\cdot, \xi)\rangle_{Y^*, Y} \, \mathrm{d}\mathbb{P} = 0 \quad \forall\, \boldsymbol{v} \in \boldsymbol{Y}.$$

It is well-posed in $\boldsymbol{Y}$ since Hölder's inequality and the growth of $\varphi'$ (Lemma A.10) yield $\boldsymbol{Az} \in L_{\mathbb{P}}^{p}(\Xi; Y^*)$ with $p \geq p^*$, $\varphi'(\boldsymbol{y}) \in L_{\mathbb{P}}^{r_f/(p-2)}(\Xi; L^{p/(p-2)}(\Omega))$, $\boldsymbol{N}'(\boldsymbol{y})\boldsymbol{z} \in L_{\mathbb{P}}^{r_{N'z}}(\Xi; Y^*)$ with $r_{N'z} = \frac{pr_f}{r_f + p(p-2)} \geq \frac{p}{p-1} = p^*$ if $r_f < \infty$ and $r_{N'z} = p$ if $r_f = \infty$, and $\tilde{b}(\cdot) = -Q(\cdot)^*(Q(\cdot)S[\cdot](u) - \hat{q}(\cdot)) \in L_{\mathbb{P}}^{p}(\Xi; Y^*)$ for $\boldsymbol{z} \in \boldsymbol{Y}$ and the assumed $\xi$-regularities of $\boldsymbol{y}$, $\boldsymbol{Q}$, and $\hat{\boldsymbol{q}}$. The integrand vanishes for almost every $\xi \in \Xi$ if $\boldsymbol{z} = \boldsymbol{T}(u)$ is inserted. Hence, $\boldsymbol{T}(u) \in L_{\mathbb{P}}^{p}(\Xi; H_0^1(\Omega)) = \boldsymbol{Y}$ is a solution of the equation. Lemma 3.23 provides the additional $\xi$-regularity.

Since $\boldsymbol{N}'(\boldsymbol{y})$ is monotone by Proposition A.4, the operator $\boldsymbol{A} + \boldsymbol{N}'(\boldsymbol{y})$ is strictly monotone on $\boldsymbol{Y}$. This proves uniqueness of a solution as in Proposition 3.16. □

Table 3.1 shows how the $\xi$-regularity of $\boldsymbol{y}$, $\boldsymbol{Q}$, and $\hat{\boldsymbol{q}}$ affects the $\xi$-regularity of the adjoint state $\boldsymbol{z}$ as deduced from (3.20). The column "Exponents" depicts the minimum requirement on the exponents for every function in (3.3) to be well-defined. Then, $L_{\mathbb{P}}^{p/(p-1)}$-regularity of the adjoint state can be shown (Proposition 3.22). In the columns "Estimation 1" and "Estimation 2" it is listed which values of the exponent $r_Q$ and $r_{\hat{q}}$ are required for the adjoint state to belong to $\boldsymbol{Y} = L_{\mathbb{P}}^{p}(\Xi, Y)$ if $\boldsymbol{y} \in L_{\mathbb{P}}^{r_f}(\Xi; Y)$ holds (e. g., by Corollary 3.15). In the case $r_Q = \infty$ and $r_{\hat{q}} \geq r_f$ (column "Example"), we get $\boldsymbol{z} \in L_{\mathbb{P}}^{r_f}(\Xi; Y)$, i. e., the adjoint state has the same $\xi$-regularity as the state itself. In the following, we will restrict the discussion to this case, i. e., $\boldsymbol{Q} \in L_{\mathbb{P}}^{\infty}(\Xi; \mathcal{L}(Y, H))$ and $\hat{\boldsymbol{q}} \in L_{\mathbb{P}}^{r_f}(\Xi; L^2(\Omega))$, since then the state, the desired state, and the adjoint state all enjoy $L_{\mathbb{P}}^{r_f}$-regularity.

| Function and space | Exponents | Estimation 1 | Estimation 2 | Example |
|---|---|---|---|---|
| $\boldsymbol{y} \in L^{r_f}_{\mathbb{P}}(\Xi; Y)$ | $r_f \geq p \in (3, \infty)$ | $r_f \in [p, \infty)$ | $r_f = \infty$ | $r_f \geq p$ |
| $\boldsymbol{Q} \in L^{r_Q}_{\mathbb{P}}(\Xi; \mathcal{L}(Y, H))$ | $r_Q \geq \frac{2p}{p-2}$ | $r_Q \geq \frac{2pr_f}{r_f - p}$ | $r_Q \geq 2p$ | $r_Q = \infty$ |
| $\hat{\boldsymbol{q}} \in L^{r_{\hat{q}}}_{\mathbb{P}}(\Xi; H)$ | $r_{\hat{q}} \geq 2$ | $r_{\hat{q}} \geq \frac{2pr_f}{r_f + p}$ | $r_{\hat{q}} \geq r_Q$ | $r_{\hat{q}} \geq r_f$ |
| $\boldsymbol{Qy} \in L^{r_{Qy}}_{\mathbb{P}}(\Xi; H)$ | $\frac{1}{r_{Qy}} = \frac{1}{r_Q} + \frac{1}{r_f}, r_{Qy} \geq 2$ | $r_{Qy} \geq \frac{2pr_f}{r_f + p}$ | $r_{Qy} = r_Q$ | $r_{Qy} = r_f$ |
| $\boldsymbol{Qy} - \hat{\boldsymbol{q}} \in L^{\hat{r}}_{\mathbb{P}}(\Xi; H)$ | $\hat{r} = \min\{r_{Qy}, r_{\hat{q}}\} \geq 2$ | $\hat{r} \geq \frac{2pr_f}{r_f + p}$ | $\hat{r} = r_Q$ | $\hat{r} = r_f$ |
| $\boldsymbol{z} \in L^{r_z}_{\mathbb{P}}(\Xi; Y)$ | $\frac{1}{r_z} = \frac{1}{\hat{r}} + \frac{1}{r_Q}, r_z \geq \frac{p}{p-1}$ | $r_z \geq p$ | $r_z = \frac{r_Q}{2} \geq p$ | $r_z = r_f$ |

Table 3.1.: Overview of the integrability inheritance

**Remark 3.25.** The adjoint equation can be well-defined in a larger space than $\boldsymbol{Y}$. If, e. g., $\boldsymbol{y} \in L^{\infty}_{\mathbb{P}}(\Xi; H^1_0(\Omega) \cap L^{\infty}(\Omega))$ ($r_f = \infty$), we have $\varphi'(\boldsymbol{y}) \in L^{\infty}_{\lambda \otimes \mathbb{P}}(\Omega \times \Xi)$ by the continuity of $\varphi'$. Note that $L^{\infty}_{\mathbb{P}}(\Xi; L^{\infty}(\Omega)) \subset L^{\infty}_{\lambda \otimes \mathbb{P}}(\Omega \times \Xi)$. It follows that

$$\varphi'(\boldsymbol{y})\boldsymbol{v} \in L^2_{\lambda \otimes \mathbb{P}}(\Omega \times \Xi) \cong L^2_{\mathbb{P}}(\Xi; L^2(\Omega))$$

and thus by $Y^* \subset L^2(\Omega) \subset Y$ that $\boldsymbol{N}'(\boldsymbol{y}) \in \mathcal{L}(L^2_{\mathbb{P}}(\Xi; Y), L^2_{\mathbb{P}}(\Xi; Y^*))$. The adjoint equation is then even well-posed in $L^2_{\mathbb{P}}(\Xi; Y) \supset \boldsymbol{Y}$.
From $\boldsymbol{y} \in L^{r_f}_{\mathbb{P}}(\Xi; H^1_0(\Omega) \cap L^{r_f}(\Omega))$ with $r_f \in [p, \infty)$ we can conclude that

$$\varphi'(\boldsymbol{y}) \in L^{r_f/(p-2)}_{\mathbb{P}}(\Xi; L^{r_f/(p-2)}(\Omega))$$

and that $\boldsymbol{N}'(\boldsymbol{y}) \in \mathcal{L}(L^{\tilde{r}}_{\mathbb{P}}(\Xi; Y), L^{\tilde{r}^*}_{\mathbb{P}}(\Xi; Y^*))$ holds with $\tilde{r} = \frac{2r_f}{r_f - p + 2} \leq p$ by Proposition 3.1. The adjoint equation is then even well-defined in $L^{\tilde{r}}_{\mathbb{P}}(\Xi; Y) \supset \boldsymbol{Y}$.

Now we use that the derivative of the parametric, reduced objective function is given by

$$\hat{J}[\xi]'(u) = -B(\xi)^* T[\xi](u) + \gamma(u, \cdot)_U. \tag{3.21}$$

It is measurable as a composition of measurable functions and has the same $\xi$-regularity as the adjoint state because $B(\cdot) \in L^{\infty}_{\mathbb{P}}(\Xi; \mathcal{L}(U, Y^*))$, especially $\|\hat{J}[\cdot]'(u)\|_{U^*} \in L^{p/(p-1)}_{\mathbb{P}}(\Xi) \subset L^1_{\mathbb{P}}(\Xi)$ by Propositions 3.22 and A.2. Furthermore, the function $\xi \mapsto \hat{J}[\xi](u)$ belongs to $L^{p/2}_{\mathbb{P}}(\Xi) \subset L^1_{\mathbb{P}}(\Xi)$ if $\boldsymbol{Q} \in L^{\infty}_{\mathbb{P}}(\Xi; \mathcal{L}(Y, H))$. Therefore, we can apply the chain rule and get

$$\frac{\mathrm{d}}{\mathrm{d}u}\hat{\boldsymbol{J}}(u) = \frac{\mathrm{d}}{\mathrm{d}u}\int_{\Xi} \hat{J}[\xi](u)\,\mathrm{d}\mathbb{P} = \int_{\Xi} \hat{J}[\xi]'(u)\,\mathrm{d}\mathbb{P} = -\boldsymbol{B}^*\boldsymbol{T}(u) + \gamma(u, \cdot)_U$$

since $\mathbb{E} : L^1_{\mathbb{P}}(\Xi) \to \mathbb{R}$ is linear and bounded. The pointwise adjoint state can thus be used for the computation of the gradient of the reduced objective function $\hat{\boldsymbol{J}}$.

### 3.4.2. Differentiability of the Control-to-State Mapping

By the argumentation above, we see that $\hat{\boldsymbol{J}}$ is continuously differentiable, but have not used differentiability properties of the control-to-state mapping $\boldsymbol{S}$ (Definition 3.11) in a concrete way. As discussed in Corollary 3.15, the control-to-state mapping $\boldsymbol{S}$ is well-defined. It is even continuously differentiable:

**Theorem 3.26.** *Under Assumption 3.3 and using the Definitions* (3.10)*, the control-to-state mapping $\boldsymbol{S}$ (Definition 3.11) is continuously differentiable with the derivative $\boldsymbol{S}' : U \to \mathcal{L}(U, \boldsymbol{Y}), [\boldsymbol{S}'(u)w](\cdot, \xi) := S[\xi]'(u)w$.*

*Proof.* The pointwise derivative applied to $w \in U$ is the solution of

$$[A(\xi) + N_y(S[\xi](u), \xi)](S[\xi]'(u)w) = -B(\xi)w.$$

In analogy to the discussion of the adjoint state in the proof of Proposition 3.22 we conclude that $S[\xi]'(u)w$ admits the same $\xi$-regularity as $B(\cdot)$ by

$$\|S[\xi]'(u)w\|_Y \leq \tfrac{1}{\underline{\kappa}}\|B(\xi)w\|_{Y^*} \leq \tfrac{1}{\underline{\kappa}}\|B(\xi)\|_{\mathcal{L}(U,Y^*)}\|w\|_U \tag{3.22}$$

giving $\|S[\xi]'(u)\|_{\mathcal{L}(U,Y)} \leq \frac{1}{\underline{\kappa}}\|B(\xi)\|_{\mathcal{L}(U,Y^*)}$. In our case $B$ is constant and thus it holds that $B \in L^\infty_{\mathbb{P}}(\Xi; \mathcal{L}(U, Y^*))$ and therefore $S[\xi]'(u) \in L^\infty_{\mathbb{P}}(\Xi; \mathcal{L}(U, Y)) \subset \boldsymbol{Y}$. Linearity w. r. t. $w$ is obvious; the given operator $\boldsymbol{S}'$ is well-defined. Since $S[\xi]$ is F-differentiable for a. e. $\xi$ by the implicit function theorem, we have that

$$\lim_{\|w\|_U \to 0} \frac{\|S[\xi](u+w) - S[\xi](u) - S[\xi]'(u)w\|_Y}{\|w\|_U} = 0$$

for a. e. $\xi \in \Xi$.

Next, the following holds true, cf. the proof of [111, Prop. A.11]:

$$\begin{aligned}
&\lim_{\|w\|_U \to 0} \frac{\|\boldsymbol{S}(u+w) - \boldsymbol{S}(u) - \boldsymbol{S}'(u)w\|_{\boldsymbol{Y}}}{\|w\|_U} \\
&= \lim_{\|w\|_U \to 0} \frac{\left(\int_\Xi \|S[\xi](u+w) - S[\xi](u) - S[\xi]'(u)w\|_Y^p \,\mathrm{d}\mathbb{P}\right)^{1/p}}{\|w\|_U} \\
&= \Big(\lim_{\|w\|_U \to 0} \int_\Xi \frac{\|S[\xi](u+w) - S[\xi](u) - S[\xi]'(u)w\|_Y^p}{\|w\|_U^p} \,\mathrm{d}\mathbb{P}\Big)^{1/p}. \\
&= \Big(\lim_{\|w\|_U \to 0} \int_\Xi \frac{\|\int_0^1 [S[\xi]'(u+\tau w) - S[\xi]'(u)]w \,\mathrm{d}\tau\|_Y^p}{\|w\|_U^p} \,\mathrm{d}\mathbb{P}\Big)^{1/p} = 0.
\end{aligned}$$

This limit is zero by the dominated convergence theorem because the integrand can be bounded by

$$\sup_{\tau \in [0,1]} \|S[\xi]'(u+\tau w) - S[\xi]'(u)\|^p_{\mathcal{L}(U,Y)} \leq \tfrac{2^p}{\underline{\kappa}^p}\|B(\xi)\|^p_{\mathcal{L}(U,Y^*)},$$

where the upper bound is an $L^1_{\mathbb{P}}(\Xi)$-function w. r. t. $\xi$ for all $w \in U$. We obtain that $\boldsymbol{S}$ is F-differentiable with the given derivative.

Continuity of $\boldsymbol{S}'$ can be shown as follows: We know that $S[\xi]'$ is continuous for a. e. $\xi \in \Xi$. Moreover, we have $\|S[\xi]'(u)\|_{\mathcal{L}(U,Y)} \leq \frac{1}{\underline{\kappa}} \|B(\xi)\|_{\mathcal{L}(U,Y^*)}$ independently of $u \in U$ by (3.22). For $u, \tilde{u}, w \in U$ it holds that

$$\begin{aligned} \|(\boldsymbol{S}'(u) - \boldsymbol{S}'(\tilde{u}))w\|_{\boldsymbol{Y}}^p &= \int_\Xi \|(S[\xi]'(u) - S[\xi]'(\tilde{u}))w\|_Y^p \,\mathrm{d}\mathbb{P} \\ &\leq \int_\Xi \|S[\xi]'(u) - S[\xi]'(\tilde{u})\|_{\mathcal{L}(U,Y)}^p \,\mathrm{d}\mathbb{P} \cdot \|w\|_U^p, \end{aligned}$$

giving

$$0 \leq \|(\boldsymbol{S}'(u) - \boldsymbol{S}'(\tilde{u}))\|_{\mathcal{L}(U,\boldsymbol{Y})}^p \leq \int_\Xi \|S[\xi]'(u) - S[\xi]'(\tilde{u})\|_{\mathcal{L}(U,Y)}^p \,\mathrm{d}\mathbb{P} \longrightarrow 0 \quad \text{as} \quad \tilde{u} \xrightarrow{U} u.$$

This limit is obtained because the integrand converges to zero almost everywhere and can be bounded by $\frac{2^p}{\underline{\kappa}^p} \|B(\cdot)\|_{\mathcal{L}(U,Y^*)}^p$, which is an $L_{\mathbb{P}}^1(\Xi)$-function if at least $B \in L_{\mathbb{P}}^p(\Xi; \mathcal{L}(U, Y^*))$. □

Overall, we have shown that the control-to-state mapping $\boldsymbol{S}$ is well-defined and continuously differentiable. Therefore, using the chain rule and the boundedness and linearity of the expectation operator, we obtain that the reduced objective function $\hat{\boldsymbol{J}}$ is well-defined and continuously differentiable with the given derivative. Other risk measures can also be included as long as they are at least once continuously differentiable, cf. Section 9.2.

### 3.4.3. Second Derivatives

We can follow a similar strategy as in Subsection 3.4.1 to derive an expression for the second derivative of the reduced objective function applied to a given direction $s \in U$. Following [60, Sec. 1.6.5], we see that

$$\nabla^2 \hat{J}[\xi](u)s = \iota\, B(\xi)^* h(\xi) + \gamma s \in U, \tag{3.23}$$

where $h(\xi) \in Y$ solves

$$[A(\xi) + N_y(S[\xi](u), \xi)]^* h(\xi) = Q(\xi)^* Q(\xi)\delta(\xi) + \langle T[\xi](u), [N_{yy}(S[\xi](u), \xi)\delta(\xi)]\,\cdot\rangle_{Y,Y^*} \tag{3.24}$$

with $\delta(\xi) = S[\xi]'(u)s \in Y$ solving

$$[A(\xi) + N_y(S[\xi](u), \xi)]\delta(\xi) = B(\xi)s. \tag{3.25}$$

The mapping $\iota : U^* \to U$ is the Riesz representation operator. It is used that $N$ is twice continuously differentiable w. r. t. $y$. Both $\delta(\xi)$ and $h(\xi)$ are unique as discussed in Subsection 3.4.1. Since $A(\xi)$ and $N_y(S[\xi](u), \xi)$ are self-adjoint, the results for the adjoint equation are also applicable to the linearized state equation. As in Proposition 3.22 we have

$$\|\delta(\xi)\|_Y \leq \tfrac{1}{\underline{\kappa}} \|B(\xi)\|_{\mathcal{L}(U,Y^*)} \|s\|_U \leq \|D\|_{\mathcal{L}(U,L^2(\Omega))} \|s\|_U$$

and therefore $\delta \in L_{\mathbb{P}}^\infty(\Xi; Y)$. In analogy to (3.20), it follows that

$$\|h(\xi)\|_Y \leq \tfrac{1}{\underline{\kappa}} \big( \|Q(\xi)\|_{\mathcal{L}(Y,H)} \|Q(\xi)\delta(\xi)\|_H + \|T[\xi](u)\|_Y \|N_{yy}(S[\xi](u), \xi)\delta(\xi)\|_{\mathcal{L}(Y,Y^*)} \big) \tag{3.26}$$

and thus $h \in L^{r_h}_{\mathbb{P}}(\Xi; Y)$ with $r_h = \min\{\frac{r_Q}{2}, (\frac{1}{r_z} + \frac{p-3}{r_f})^{-1}\} \geq \frac{p}{p-2} > 1$ with the prerequisites and notation from Proposition 3.24, especially $r_z \geq p$. We obtain $\nabla^2 \hat{J}[\cdot](u)s \in L^{r_h}_{\mathbb{P}}(\Xi; U) \subset L^1_{\mathbb{P}}(\Xi; U)$.

Again, the parametrized linearized state equation and adjoint equation can be written in weak form also w. r. t. the parameters. As in the case of the first derivative, the regularity estimates and measurability yield

$$\nabla^2 \hat{\boldsymbol{J}}(u)s = \int_\Xi \nabla^2 \hat{J}[\xi](u)s \, \mathrm{d}\mathbb{P}.$$

# 4. An Inexact Trust-Region Algorithm for the Solution of Optimal Control Problems with Control Constraints

We present and discuss an inexact and projection-based trust-region algorithm, which can be used for the solution of optimal control problems of the form

$$\min_{(y,u)\in Y\times U} J(y,u) \quad \text{s. t.} \quad E(y,u)=0, \quad u\in U_{\text{ad}}. \tag{4.1}$$

We want to use this algorithm later to solve problem (3.3) adaptively and to control all arising errors caused by discretization, inexact solution of linear systems and tensor truncation during this procedure.

One possible solution algorithm would be an an inexact, trust-region-based SQP algorithm such as [118], which extends [119, 57] to the case of control constraints handled by projections onto $U_{\text{ad}}$. Here it is necessary to minimize the residual of the linearized state equation approximately. If the space $Z$ of the residual is not a Hilbert space as in the example from Section 3.2, this is possibly a nonsmooth problem so that this approach is not preferable in our setting. To circumvent this difficulty, we can work with the reduced problem (4.3) and use an extension of the algorithm presented in [70], which is based on [69]. The algorithm in [70] can handle inexact gradients as well as inexact objective function evaluations. Earlier versions of inexact trust-region algorithms formulated in Hilbert space [26, 27] have the disadvantage that the relative errors in the gradient and the objective function evaluation have to be bounded by fixed, prescribed constants of a certain magnitude, whereas [70] allows for fixed, but possibly unknown and arbitrarily large constants. This fits very well to PDE applications, where error estimates often contain unknown multiplicative constants. We extend [70, Algorithm 4.1] to the case of control constraints using an inexact projection (inspired by [118]) and present our version in the following.

**Assumption 4.1.** We make the following assumptions on problem (4.1):

- The space $U$ is a Hilbert space; $Y$, $Z$ are reflexive Banach spaces.
- The feasible control set $U_{\text{ad}} \subset U$ is nonempty, convex and closed.
- For each control $u \in \tilde{U} \subset U$, where $\tilde{U} \supset U_{\text{ad}}$ is an open set, there exists a unique state $y = S(u)$ fulfilling $E(S(u), u) = 0$. $S : \tilde{U} \to Y$ is the control-to-state mapping.
- The control-to-state mapping $S : \tilde{U} \to Y$ and the objective function $J : Y \times \tilde{U} \to \mathbb{R}$ are such that the reduced objective function $\hat{J} : \tilde{U} \to \mathbb{R}$ in (4.3) is continuously differentiable.

If Assumption 4.1 is fulfilled, Problem (4.1) can be reduced to the control:

$$\min_{u\in U} \hat{J}(u) := J(S(u), u) \quad \text{s. t.} \quad u \in U_{\text{ad}}. \tag{4.2}$$

This motivates that we describe an inexact trust-region algorithm for the solution of general problems of the form

$$\min_{u\in U} \hat{J}(u) \quad \text{s. t.} \quad u \in U_{\text{ad}}. \tag{4.3}$$

in this chapter and prove its convergence. Later, in Chapter 5, we return to the original problem (4.2), where the objective function $\hat{J}$ is of reduced form.

Using parts of Assumption 4.1, we make the following assumption for problem (4.3):

**Assumption 4.2.**

- $U$ is a Hilbert space.
- The feasible set $U_{\text{ad}} \subset U$ is nonempty, closed and convex.
- The objective function $\hat{J} : \tilde{U} \to \mathbb{R}$ is continuously differentiable on an open set $\tilde{U}$, $U_{\text{ad}} \subset \tilde{U} \subset U$, and bounded from below on $U_{\text{ad}}$. The Fréchet approximation condition holds uniformly on every level set, i. e.,

$$\sup_{u\in U_{\text{ad}}:\hat{J}(u)\leq\hat{J}(\tilde{u})} |\hat{J}(u+s) - \hat{J}(u) - (\nabla\hat{J}(u), s)_U| = o(\|s\|_U) \quad (s \to 0), \tag{4.4}$$

  for every $\tilde{u} \in U_{\text{ad}}$.

## 4.1. Formulation of the Algorithm

For a comprehensive introduction to trust-region algorithms we refer to [31]. In the algorithm presented here, we use a typically, but not necessarily quadratic model $m_k(s)$ of $\hat{J}(u^k + s) - \hat{J}(u^k)$ with the current control $u^k$ for the computation of the current step $s^k \in U$. The step computation approximately solves

$$\min_{s\in U} m_k(s) \quad \text{s. t.} \quad u^k + s \in U_{\text{ad}}, \quad \|s\|_U \leq \Delta_k \tag{4.5}$$

with the current trust region radius $\Delta_k > 0$. For the acceptance of the step, we allow for inexact evaluations of $\hat{J}$ by using an approximation $\hat{J}_k$ instead of $\hat{J}$. We define the actual, computed (as introduced in [26]), and predicted reduction, respectively, as

$$\text{ared}_k := \hat{J}(u^k) - \hat{J}(u^k + s^k), \tag{4.6a}$$

$$\text{cred}_k := \hat{J}_k(u^k) - \hat{J}_k(u^k + s^k), \tag{4.6b}$$

$$\text{pred}_k := m_k(0) - m_k(s^k). \tag{4.6c}$$

Furthermore, we define a criticality measure for the original problem (4.3), namely

$$\chi : \tilde{U} \to \mathbb{R}_{\geq 0}, \quad \chi(u) := \|u - P_{U_{\text{ad}}}(u - \tau\nabla\hat{J}(u))\|_U \tag{4.7}$$

with a fixed parameter $\tau > 0$ and the projection $P_{U_{\text{ad}}}$ onto the feasible set. The function $\chi$ is continuous, and $\chi(\bar{u}) = 0$ holds if and only if $\bar{u}$ is a first order critical point for (4.3). In addition, a criticality measure for problem (4.5) without the trust-region constraint is defined:

$$\tilde{\chi}_k : U \to \mathbb{R}_{\geq 0}, \quad \tilde{\chi}_k(s) := \|u^k + s - P_{U_{\text{ad}}}(u^k + s - \tau \nabla m_k(s))\|_U.$$

The condition $\tilde{\chi}_k(\bar{s}) = 0$ holds if and only if $\bar{s}$ is first order critical for the problem

$$\min_{s \in U} m_k(s) \quad \text{s. t.} \quad u^k + s \in U_{\text{ad}}. \tag{4.8}$$

Typically, the projection is also not computed exactly. Therefore, we introduce the *approximate criticality measure* for problem (4.8),

$$\chi_k : U \to \mathbb{R}_{\geq 0}, \quad \chi_k(s) := \|u^k + s - \hat{\mathrm{P}}_{U_{\text{ad}}}(u^k + s - \tau \nabla m_k(s))\|_U, \tag{4.9}$$

now with an *approximate* projection $\hat{\mathrm{P}}_{U_{\text{ad}}} : U \to U_{\text{ad}}$ onto the feasible set. This can be *any* mapping approximating the exact projection. Especially, it does not have to fulfill the variational inequality defining the projection on some discrete subspace $\mathrm{U} \subset U$. In contrast to that, the properties of the discrete projection are used in [118, Lem. 5.3, Lem. 5.5] when proving the Cauchy decrease condition although an approximate version is used in the final implementation there [118, Sec. 5.2].

**Remark 4.3.** Consider the case that $U = L^2(\Omega)$. The gradient of the reduced objective function at $\mathrm{u} \in \mathrm{U}$ (discrete subspace of $U$) is, e. g., of the form $\nabla \hat{J}(\mathrm{u}) = -B^* z + \gamma \mathrm{u}$ with $B^* \equiv \iota : H_0^1(\Omega) \hookrightarrow L^2(\Omega)$ (canonical embedding). An approximation of it is obtained using an approximate adjoint state $\mathrm{z} \in \mathrm{Y}$, computed, e. g., by a finite element method with piecewise linear, continuous ansatz functions. The function $B^*\mathrm{z} \in U$ then belongs to in the finite element space Y. To avoid further approximations, it makes sense that the current control u belongs to the space U of linear FE functions (not necessarily with zero boundary data). Then, the approximate gradient and any control obtained by a linesearch in its direction are also in U. This means that we do not explicitly discretize the control space, cf. the variational discretization concept of optimal control problems [59], but its discretization is implied by the one of the adjoint state.
In general, one could equip the discrete space U of linear FE functions with the inner product induced by the lumped mass matrix. The obtained discrete projection onto a box-constrained feasible set $U_{\text{ad}} \cap \mathrm{U}$ can then be computed node-wise. As described, this is not possible in our case since the algorithm uses the exact inner product of $U$. The resulting discrete projection would have to be computed by solving a high-dimensional quadratic program with box-constraints. In order to use the node-wise projection instead as in [118, Sec. 5.2], we have to allow for an inexact projection and cannot use any projection properties of it.
Alternatively, one could use piecewise constant functions for the control. Then, the exact $L^2$-projection onto a box (with constant bounds) can be computed simply, but an additional error in the approximate gradient occurs since the operator $B^*$ cannot be applied exactly then.

To ensure global convergence of the algorithm we have to control the following quantities:

- The inexactness of the model gradient:

$$\|\nabla m_k(0) - \nabla \hat{J}(u^k)\|_U \leq \varrho_g(\Delta_k), \tag{4.10}$$

  where $\varrho_g : \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ is a function satisfying $\lim_{t\to 0^+} \varrho_g(t) = 0$, e. g., $\varrho_g(t) = \mathfrak{c}_g t$, $\mathfrak{c}_g > 0$.

- The inexactness of the approximate criticality measure:

$$|\chi_k(0) - \chi(u^k)| \leq \varrho_c(\chi_k(0)), \tag{4.11}$$

  where $\varrho_c : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a function satisfying $\lim_{t\to 0^+} \varrho_c(t) = 0$ and $\varrho_c(0) = 0$, e. g., $\varrho_c(t) = \mathfrak{c}_c t$, $\mathfrak{c}_c > 0$. Note that we do not have to use the exact criticality measure for the model problem here since it is not required in the convergence proof. It will only be used for the control of the inexactness.

- The quality of the computed reduction:

$$|\mathrm{ared}_k - \mathrm{cred}_k| \leq \varrho_r(\eta_3 \min\{\mathrm{pred}_k, \mathfrak{r}_k\}), \tag{4.12}$$

  with $\eta_3 < \min\{\eta_1, 1 - \eta_2\}$, where $0 < \eta_1 < \eta_2 < 1$ are a priori chosen parameters for assessing the quality of the model, and with a forcing sequence $(\mathfrak{r}_k)_{k\in\mathbb{N}_0} \subset \mathbb{R}_{>0}$ fulfilling $\lim_{k\to\infty} \mathfrak{r}_k = 0$ and a function $\varrho_r : \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ fulfilling $\varrho_r(t) \leq t$ for all $t \in (0, \bar{t}]$ with some fixed $\bar{t} > 0$, e. g., $\varrho_r(t) = \mathfrak{c}_r t^{\mathfrak{e}_r}$, $\mathfrak{c}_r > 0$, $\mathfrak{e}_r > 1$. Note that $\varrho_r(t) = t$ would also be possible, but then it is not sufficient to know the error in (4.12) up to an unknown, multiplicative constant.

A trial step $s^k \in U_{\mathrm{ad}} - u^k$, $\|s^k\|_U \leq \Delta_k$, has to fulfill the decrease condition

$$\mathrm{pred}_k = m_k(0) - m_k(s^k) \geq \varrho_{t1}(\chi_k(0)) \cdot \min\{\varrho_{t2}(\chi_k(0)), \Delta_k\} \tag{4.13}$$

with monotonically increasing functions $\varrho_{t1}, \varrho_{t2} : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$. These functions must be chosen such that (4.13) is satisfiable by, e. g., the generalized Cauchy point, see Section 4.3. A possible example is $\varrho_{t1}(t) = \mathfrak{c}_{t1} t$, $\varrho_{t1}(t) = \mathfrak{c}_{t2} t$ with $\mathfrak{c}_{t1}, \mathfrak{c}_{t2} > 0$.

The complete method is listed in Algorithm 1. Note that all iterates $u^k$ belong to $U_{\mathrm{ad}}$ since $s^k \in U_{\mathrm{ad}} - u^k$ is required for all trial steps. Therefore, it is sufficient to assume differentiability of $\hat{J}$ only in an open neighborhood $\tilde{U}$ of $U_{\mathrm{ad}}$, see Assumption 4.2.

**Remark 4.4.** In case of an unsuccessful step, the formulation of Algorithm 1 allows to choose $\Delta_{k+1} \in (0, \nu_1 \|s^k\|_U] \subset (0, \nu_1 \Delta_k]$ if $\|s^k\|_U > 0$, which is a suitable strategy to avoid that $s^k$ is feasible for the trust-region subproblem in the next iteration.

## 4.2. Convergence Proof

Provided all conditions in Algorithm 1 can be satisfied, we prove its convergence. This means that we assume for now that an adequate model, approximate projection, trial step, and inexact objective function exist in each iteration. We prove this in Section 4.3.

**Algorithm 1:** Inexact Trust-Region Method for Solving Problem (4.3)

**Input**: Initial iterate $u^0 \in U_{\mathrm{ad}}$

**Parameters** : $\tau > 0$, error control functions $\varrho_{\mathrm{c}}$, $\varrho_{\mathrm{g}}$, $\varrho_{\mathrm{t}1}$, $\varrho_{\mathrm{t}2}$, $\varrho_{\mathrm{r}}$,
forcing sequence $(\mathfrak{r}_k)_{k\in\mathbb{N}_0} \subset \mathbb{R}_{>0}$, $\lim_{k\to\infty} \mathfrak{r}_k = 0$.
$\Delta_{\max} \in (0,\infty]$, $\Delta_0 \in \mathbb{R}_{>0}$ s. t. $\Delta_0 \le \Delta_{\max}$,
$0 < \eta_1 < \eta_2 < 1$ and $0 < \eta_3 \le \min\{\eta_1, 1-\eta_2\}$,
$0 < \nu_1 < 1 \le \nu_2 < \nu_3$.

**Output**: Sequences $(u^k)_{k\in\mathbb{N}_0} \subset U_{\mathrm{ad}}$, $(\Delta_k)_{k\in\mathbb{N}_0} \subset \mathbb{R}_{>0}$, $(\chi_k(0))_{k\in\mathbb{N}_0} \subset \mathbb{R}_{\ge 0}$

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

Choose a model $m_k : U \to \mathbb{R}$ and an approximate projection $\hat{\mathrm{P}}_{U_{\mathrm{ad}}} : U \to U_{\mathrm{ad}}$ such that (4.10) and (4.11) hold. Compute $\chi_k(0)$ using $m_k$ and $\hat{\mathrm{P}}_{U_{\mathrm{ad}}}$.

**if** $\chi_k(0) = 0$, **then**
Set $u^\ell := u^k$, $\Delta_\ell := \Delta_k$, and $\chi_\ell(0) = 0$ for all $\ell \ge k+1$ and STOP.
**end**

Compute a trial step $s^k \in U_{\mathrm{ad}} - u^k$, $\|s^k\|_U \le \Delta_k$, fulfilling (4.13) with the computed $\chi_k(0)$, see Section 4.3.

Compute $\mathrm{pred}_k$ by (4.6c) and $\mathrm{cred}_k$ by (4.6b) with $\hat{J}_k$ such that (4.12) holds.

**if** $\frac{\mathrm{cred}_k}{\mathrm{pred}_k} < \eta_1$ *(unsuccessful step),* **then**
$u^{k+1} := u^k$, choose $\Delta_{k+1} \in (0, \nu_1\Delta_k]$.
**else if** $\frac{\mathrm{cred}_k}{\mathrm{pred}_k} \in [\eta_1, \eta_2)$ *(successful step),* **then**
$u^{k+1} := u^k + s^k$, choose $\Delta_{k+1} \in [\nu_1\Delta_k, \min\{\nu_2\Delta_k, \Delta_{\max}\}]$.
**else if** $\frac{\mathrm{cred}_k}{\mathrm{pred}_k} \ge \eta_2$ *(very successful step),* **then**
$u^{k+1} := u^k + s^k$, choose $\Delta_{k+1} \in [\min\{\nu_2\Delta_k, \Delta_{\max}\}, \min\{\nu_3\Delta_k, \Delta_{\max}\}]$.
**end**

**end**

We require the following assumption in addition to Assumption 4.2 to prove the convergence result given in Theorem 4.6.

**Assumption 4.5.** Each model $m_k : U \to \mathbb{R}$ is continuously differentiable. The Fréchet approximation condition holds uniformly over all models:

$$\sup_{k\in\mathbb{N}_0} |m_k(s) - m_k(0) - (\nabla m_k(0), s)_U| = o(\|s\|) \quad (s \to 0). \tag{4.14}$$

The model gradients are Lipschitz continuous on the sets of feasible search directions, i. e.,

$$\|\nabla m_k(s) - \nabla m_k(\hat{s})\|_U \le \mathfrak{c}_{m_k} \|s - \hat{s}\|_U$$

holds for all $s, \hat{s} \in U_{\mathrm{ad}} - u^k$, $\|s\|_U \le \Delta_k$, $\|\hat{s}\|_U \le \Delta_k$ with some $\mathfrak{c}_{m_k} > 0$. The Lipschitz constants shall be bounded uniformly: $\mathfrak{c}_{m_k} \le \mathfrak{c}_m$ for some $\mathfrak{c}_m > 0$ and all $k \in \mathbb{N}_0$.

**Theorem 4.6.** *Let Assumptions 4.2 and 4.5 hold and let the sequence $(u^k)_{k\in\mathbb{N}_0} \subset U$ be generated by Algorithm 1. Then,*

$$\liminf_{k\to\infty} \chi(u^k) = 0$$

*holds with the criticality measure $\chi$ defined in* (4.7).

We apply the following two lemmas to prove Theorem 4.6:

**Lemma 4.7.** *Let Assumptions 4.2 and 4.5 hold and let the sequence of inexact criticality measures (as defined in* (4.9)*) generated by Algorithm 1 satisfy*

$$\chi_k(0) \geq \varepsilon > 0 \quad \text{for all } k \geq K_1 \in \mathbb{N}_0 \tag{4.15}$$

*for some fixed* $\varepsilon > 0$.
*Then,* $\lim_{k\to\infty} \Delta_k = 0$ *holds for the sequence of corresponding trust-region radii.*

*Proof.* First observe that the termination criterion "$\chi_k(0) = 0$" in the algorithm is not met for *any* $k \in \mathbb{N}_0$ by assumption, because $\chi_k(0) = 0$ for some $k \in \mathbb{N}_0$ would yield $\chi_\ell(0) = 0$ for all $\ell \geq k$, which contradicts (4.15). Moreover, $\mathrm{pred}_k > 0$ holds for all $k \in \mathbb{N}_0$ due to (4.13), the positivity property of $\varrho_{\mathrm{t}1}$ and $\varrho_{\mathrm{t}2}$, and $\chi_k(0) > 0$, $\Delta_k > 0$.

Due to $\lim_{k\to\infty} \mathfrak{r}_k = 0$ and $\varrho_{\mathrm{r}}(t) \leq t$ for small enough $t$, it holds that

$$\varrho_{\mathrm{r}}(\eta_3 \min\{\mathrm{pred}_k, \mathfrak{r}_k\}) \leq \eta_3 \mathrm{pred}_k \quad \text{for all } k \geq K_2 \in \mathbb{N}_0$$

with some $K_2 \geq K_1$. By (4.12) we thus get

$$|\mathrm{ared}_k - \mathrm{cred}_k| \leq \eta_3 \mathrm{pred}_k \quad \text{for all } k \geq K_2.$$

This implies

$$\mathrm{ared}_k = \mathrm{cred}_k + \mathrm{ared}_k - \mathrm{cred}_k \geq \mathrm{cred}_k - \eta_3 \mathrm{pred}_k = \big(\tfrac{\mathrm{cred}_k}{\mathrm{pred}_k} - \eta_3\big)\mathrm{pred}_k \tag{4.16}$$

for all $k \geq K_2$. This is well-defined due to $\mathrm{pred}_k > 0$.

Now we show that $\sum_{k=0}^{\infty} \Delta_k < \infty$ follows from (4.15).

For all unsuccessful steps $k \in \mathcal{I}_{\mathrm{u}} \subset \mathbb{N}_0$ we have $\Delta_{k+1} \leq \nu_1 \Delta_k$ with the parameter $\nu_1 \in (0, 1)$. Thus, if there are only finitely many (very) successful steps, the sequence $(\Delta_k)_{k\in\mathbb{N}_0}$ is summable since then $\Delta_k \leq {\nu_1}^{k-K_3} \Delta_{K_3}$ for all $k \geq K_3$ for some $K_3 \in \mathbb{N}_0$. In the following, we consider the case of infinitely many (very) successful steps.

For a (very) successful step $k \in \mathcal{I}_{\mathrm{s}} = \mathbb{N}_0 \setminus \mathcal{I}_{\mathrm{u}}$, i. e.,

$$\tfrac{\mathrm{cred}_k}{\mathrm{pred}_k} \geq \eta_1 \text{ and } \Delta_{k+1} \leq \min\{\nu_3 \Delta_k, \Delta_{\max}\} \leq \nu_3 \Delta_k,$$

we deduce from (4.16) and (4.13):

$$\begin{aligned} \mathrm{ared}_k &\geq \big(\tfrac{\mathrm{cred}_k}{\mathrm{pred}_k} - \eta_3\big)\mathrm{pred}_k \geq (\eta_1 - \eta_3)\mathrm{pred}_k \\ &\geq (\eta_1 - \eta_3) \cdot \varrho_{\mathrm{t}1}(\chi_k(0)) \cdot \min\{\varrho_{\mathrm{t}2}(\chi_k(0)), \Delta_k\} \\ &\geq (\eta_1 - \eta_3) \cdot \varrho_{\mathrm{t}1}(\varepsilon) \cdot \min\{\varrho_{\mathrm{t}2}(\varepsilon), \Delta_k\} > 0, \end{aligned} \tag{4.17}$$

for all $k \in \mathcal{I}_{\mathrm{s}}$, $k \geq K_2$, where we used in the last estimate that $\varrho_{\mathrm{t}1}$ and $\varrho_{\mathrm{t}2}$ are increasing. Since the sequence $(\hat{J}(u^k))_{k\in\mathcal{I}_{\mathrm{s}}}$ is bounded from below by Assumption 4.2, we get

$$0 \leq \sum_{k\in\mathcal{I}_{\mathrm{s}}, k\geq K_2} \mathrm{ared}_k = \sum_{k\in\mathcal{I}_{\mathrm{s}}, k\geq K_2} \big(\hat{J}(u^k) - \hat{J}(u^{k+1})\big) = \sum_{k=K_2}^{\infty} \big(\hat{J}(u^k) - \hat{J}(u^{k+1})\big) < \infty \tag{4.18}$$

using that $u^{k+1} = u^k + s^k$ in the case of a (very) successful step and $u^{k+1} = u^k$ in the case of an unsuccessful step. Due to $\eta_1 > \eta_3$, $\varrho_{\mathrm{t}1}(\varepsilon) > 0$, $\varrho_{\mathrm{t}2}(\varepsilon) > 0$ and $\lim_{\mathcal{I}_\mathrm{s} \ni k \to \infty} \mathrm{ared}_k = 0$ (by (4.18)), it follows from (4.17) that $\Delta_k \leq \frac{\mathrm{ared}_k}{(\eta_1 - \eta_3)\cdot \varrho_{\mathrm{t}1}(\varepsilon)}$ for all $k \in \mathcal{I}_\mathrm{s}$, $k \geq K_4$, with some sufficiently large $K_4 \in \mathbb{N}_0$, $K_4 \geq K_2$, and thus

$$0 \leq \sum_{k \in \mathcal{I}_\mathrm{s}} \Delta_k < \infty \tag{4.19}$$

by (4.18).

Now we consider two (very) successful steps $\tilde{k}, \hat{k} \in \mathcal{I}_\mathrm{s}$, $\hat{k} \geq \tilde{k}+2$ with only unsuccessful steps $k \in \{\tilde{k}+1, \ldots, \hat{k}-1\}$ in between. Hence, we have $\Delta_k \leq \nu_3 {\nu_1}^{k-\tilde{k}-1} \Delta_{\tilde{k}}$ for $\tilde{k}+1 \leq k \leq \hat{k}-1$ and thus (using the geometric series with $\nu_1 \in (0,1)$)

$$\Sigma(\tilde{k}) := \sum_{k=\tilde{k}}^{\hat{k}-1} \Delta_k \leq \Delta_{\tilde{k}} \Big(1 + \nu_3 \sum_{\ell=0}^{\hat{k}-\tilde{k}-2} {\nu_1}^\ell\Big) \leq \Delta_{\tilde{k}} \Big(1 + \tfrac{\nu_3}{1-\nu_1}\Big).$$

Additionally, for $\tilde{k} \in \mathcal{I}_\mathrm{s}$ such that $\tilde{k}+1 \in \mathcal{I}_\mathrm{s}$ we set $\Sigma(\tilde{k}) = \Delta_{\tilde{k}}$ and in the case $0 \notin \mathcal{I}_\mathrm{s}$ we set and estimate

$$\Sigma(0) := \sum_{k=0}^{\hat{k}-1} \Delta_k \leq \Delta_0 \Big(\sum_{\ell=0}^{\hat{k}-1} {\nu_1}^\ell\Big) \leq \Delta_0 \cdot \tfrac{1}{1-\nu_1},$$

where $\hat{k} = \min \mathcal{I}_\mathrm{s}$. Therefore,

$$0 \leq \sum_{k=0}^{\infty} \Delta_k = \sum_{\tilde{k} \in \mathcal{I}_\mathrm{s} \cup \{0\}} \Sigma(\tilde{k}) \leq \Big(1 + \tfrac{\nu_3}{1-\nu_1}\Big) \cdot \sum_{\tilde{k} \in \mathcal{I}_\mathrm{s} \cup \{0\}} \Delta_{\tilde{k}} < \infty$$

follows with $\Sigma(\tilde{k}) \leq \Delta_{\tilde{k}} \big(1 + \frac{\nu_3}{1-\nu_1}\big)$ for all $\tilde{k} \in \mathcal{I}_\mathrm{s} \cup \{0\}$ (using $\nu_3 \geq 1$) and (4.19). We see that $(\Delta_k)_{k \in \mathbb{N}_0}$ is summable and thus $\lim_{k \to \infty} \Delta_k = 0$. □

**Lemma 4.8.** *Under Assumptions 4.2 and 4.5*

$$\liminf_{k \to \infty} \chi_k(0) = 0 \tag{4.20}$$

*holds true for every sequence generated by Algorithm 1, where the inexact criticality measure $\chi_k$ is defined as in* (4.9).

*Proof.* For a proof by contradiction, assume that (4.20) is false, giving that (4.15) is true for some fixed $\varepsilon > 0$. By Lemma 4.7 we have that $\lim_{k \to \infty} \Delta_k = 0$ and thus $\lim_{k \to \infty} \|s^k\|_U = 0$.

In analogy to (4.16), we can estimate

$$\mathrm{cred}_k \geq \mathrm{ared}_k - |\mathrm{ared}_k - \mathrm{cred}_k| \geq \mathrm{ared}_k - \eta_3 \mathrm{pred}_k \tag{4.21}$$

for all $k \geq K_2 \geq K_1$.

As in (4.17) we infer

$$\mathrm{pred}_k \geq \varrho_{\mathrm{t}1}(\varepsilon) \cdot \Delta_k$$

for all $k \geq K_5$ with some sufficiently large $K_5 \in \mathbb{N}_0$, $K_5 \geq K_2$, due to (4.13) and $\lim_{k\to\infty} \Delta_k = 0$, i. e., $\Delta_k \leq \varrho_{\mathrm{t}2}(\varepsilon)$ for all $k \geq K_5$ because $\varrho_{\mathrm{t}2}(\varepsilon) > 0$. Thus, we can bound

$$|o(\Delta_k)| \leq (1 - \eta_3 - \eta_2)\mathrm{pred}_k \tag{4.22}$$

for all $k \geq K_6$ with $K_6 \geq K_5 \geq K_2$ sufficiently large, since $(1 - \eta_3 - \eta_2) > 0$. Using this and the bounds indicated below, we estimate for $k \geq K_6$ and using $s_k \to 0$:

$$\begin{aligned}
\mathrm{cred}_k &\overset{(4.21)}{\geq} \mathrm{ared}_k - \eta_3\mathrm{pred}_k \overset{(4.4)}{\geq} -(\nabla\hat{J}(u^k), s^k)_U - \eta_3\mathrm{pred}_k - |o(\|s^k\|)| \\
&\overset{(4.14)}{\geq} \mathrm{pred}_k + (\nabla m_k(0), s^k)_U - (\nabla J(u^k), s^k)_U - \eta_3\mathrm{pred}_k - |o(\|s^k\|)| \\
&\geq (1 - \eta_3)\mathrm{pred}_k - \|\nabla m_k(0) - \nabla\hat{J}(u^k)\|_U\|s^k\|_U - |o(\|s^k\|)| \\
&\overset{\substack{(4.10),\\ \|s^k\|_U \leq \Delta_k}}{\geq} (1 - \eta_3)\mathrm{pred}_k - \varrho_{\mathrm{g}}(\Delta_k)\Delta_k - |o(\Delta_k)| \overset{\substack{\varrho_{\mathrm{g}}(t)\to 0\\ (t\to 0^+)}}{\geq} (1 - \eta_3)\mathrm{pred}_k - |o(\Delta_k)| \\
&\overset{(4.22)}{\geq} (1 - \eta_3)\mathrm{pred}_k - (1 - \eta_3 - \eta_2)\mathrm{pred}_k = \eta_2\mathrm{pred}_k
\end{aligned}$$

Note that $u^k \in \{u \in U_{\mathrm{ad}} : \hat{J}(u) \leq \hat{J}(u^{K_2})\}$ holds for all $k \geq K_2$ due to (4.17) so that (4.4) is applicable. In fact, the objective function values $(\hat{J}(u^k))_{k\geq K_2}$ are non-increasing since (4.17) holds for (very) successful steps and the function values do not change for unsuccessful steps. Using $\mathrm{pred}_k > 0$ as in the proof of Lemma 4.7, it follows that $\frac{\mathrm{cred}_k}{\mathrm{pred}_k} \geq \eta_2$ for all $k \geq K_6$, i. e., all steps $k \geq K_6$ are successful giving $\Delta_{k+1} \geq \min\{\nu_2\Delta_k, \Delta_{\max}\} \geq \Delta_k > 0$ due to $\nu_2 \geq 1$. This contradicts $\lim_{k\to\infty} \Delta_k = 0$, proving (4.20). □

Using Lemma 4.8, the proof of Theorem 4.6 is very short:

*Proof of Theorem 4.6.* Due to (4.11) we have

$$\chi(u^k) \leq \chi_k(0) + |\chi_k(0) - \chi(u^k)| \leq \chi_k(0) + \varrho_{\mathrm{c}}(\chi_k(0)).$$

This is also true if the algorithm is stopped due to $\chi_k(0) = 0$, because then $\chi(u^k) = 0$ follows from (4.11). Therefore, $\chi(u^\ell) = \chi(u^k) = 0 = \chi_\ell(0)$ for all $\ell \geq k$. The bound on $\chi(u^k)$ and $\lim_{t\to 0^+} \varrho_{\mathrm{c}}(t) = 0$ as well as $\varrho_{\mathrm{c}}(0) = 0$ show

$$0 \leq \liminf_{k\to\infty} \chi(u^k) \leq \liminf_{k\to\infty} \chi_k(0) + \varrho_{\mathrm{c}}(\chi_k(0)) = 0.$$

□

## 4.3. Satisfying the Conditions Required by the Algorithm

We show that Algorithm 1 is realizable, i. e., that all conditions can be satisfied under certain assumptions. This includes the computation of a generalized Cauchy point satisfying (4.13). In addition to Assumptions 4.2 and 4.5 we require:

**Assumption 4.9.**

- For every iterate $u^k \in U_{\mathrm{ad}}$, one can compute $\nabla m_k(0) \in U$ such that

$$\|\nabla m_k(0) - \nabla \hat{J}(u^k)\|_U \leq c_{\mathrm{g}} \varepsilon_{\mathrm{g}} \tag{4.23}$$

  holds with some fixed but possibly unknown constant $c_{\mathrm{g}} > 0$ and a still to be specified $\varepsilon_{\mathrm{g}} > 0$.

- For every $w^k(t) := u^k - t\nabla m_k(0) \in U$ with the values of $t > 0$ specified later, one can compute $\hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t)) \in U_{\mathrm{ad}}$ such that

$$\|P_{U_{\mathrm{ad}}}(w^k(t)) - \hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t))\|_U \leq c_{\mathrm{p}} \varepsilon_{\mathrm{p}} \tag{4.24}$$

  holds with some fixed but possibly unknown constant $c_{\mathrm{p}} > 0$ and a still to be specified $\varepsilon_{\mathrm{p}} > 0$.

- For every iterate $u^k \in U_{\mathrm{ad}}$ and every trial step $s^k \in U$, one can compute $\hat{J}_k(u) \in \mathbb{R}$ such that

$$|\hat{J}_k(u) - \hat{J}(u)| \leq c_{\mathrm{o}} \varepsilon_{\mathrm{o}}$$

  holds with some fixed but possibly unknown constant $c_{\mathrm{o}} > 0$ and a still to be specified $\varepsilon_{\mathrm{o}} > 0$, where $u \in \{u^k, u^k + s^k\}$.

**Remark 4.10.** It is necessary to require (4.24) not only for $t = \tau$ because an Armijo type linesearch, which tests different values of $t$, is employed for computing the generalized Cauchy point later.

### Model and Approximate Projection

To ensure (4.10) and (4.11), the following result is helpful:

**Proposition 4.11.** *Let $\mathfrak{c}_{\mathrm{s}} \in [0, 1]$ be given and let the following be fulfilled:*

$$\|\nabla m_k(0) - \nabla \hat{J}(u^k)\|_U \leq \min\{\tfrac{1-\mathfrak{c}_{\mathrm{s}}}{\tau} \varrho_{\mathrm{c}}(\chi_k(0)), \varrho_{\mathrm{g}}(\Delta_k)\}, \tag{4.25}$$

$$\|[P_{U_{\mathrm{ad}}} - \hat{\mathrm{P}}_{U_{\mathrm{ad}}}](u^k - \tau \nabla m_k(0))\|_U \leq \mathfrak{c}_{\mathrm{s}} \varrho_{\mathrm{c}}(\chi_k(0)). \tag{4.26}$$

*Then,* (4.10) *and* (4.11) *hold, where $\chi$ and $\chi_k$ are defined as in* (4.7) *and* (4.9)*, respectively.*

*Proof.* The estimate (4.10) follows directly from (4.25). Using the definitions (4.7) and (4.9), we estimate

$$\begin{aligned}
|\chi_k(0) - \chi(u^k)| &\leq |\chi_k(0) - \tilde{\chi}_k(0)| + |\tilde{\chi}_k(0) - \chi(u^k)| \\
&\leq \|P_{U_{\mathrm{ad}}}(u^k - \tau \nabla m_k(0)) - \hat{\mathrm{P}}_{U_{\mathrm{ad}}}(u^k - \tau \nabla m_k(0))\|_U \\
&\quad + \|P_{U_{\mathrm{ad}}}(u^k - \tau \nabla \hat{J}(u^k)) - P_{U_{\mathrm{ad}}}(u^k - \tau \nabla m_k(0))\|_U \\
&\leq \|[P_{U_{\mathrm{ad}}} - \hat{\mathrm{P}}_{U_{\mathrm{ad}}}](u^k - \tau \nabla m_k(0))\|_U + \tau \|\nabla m_k(0) - \nabla \hat{J}(u^k)\|_U \\
&\leq \mathfrak{c}_{\mathrm{s}} \varrho_{\mathrm{c}}(\chi_k(0)) + (1 - \mathfrak{c}_{\mathrm{s}}) \varrho_{\mathrm{c}}(\chi_k(0)) = \varrho_{\mathrm{c}}(\chi_k(0)),
\end{aligned}$$

where the second inequality is established using $\big|\|u\|_U - \|w\|_U\big| \leq \|u - w\|_U$ for any $u, w \in U$ and the last one follows from (4.26) and (4.25). This proves that condition (4.11) is satisfied. □

The two conditions (4.25) and (4.26) can be satisfied as follows:

**Lemma 4.12.** *Let* (4.23) *and* (4.24) *with* $t = \tau$ *be fulfilled with* $\varepsilon_\mathrm{g}$ *and* $\varepsilon_\mathrm{p}$ *such that*

$$(1 + (1 - \mathfrak{c}_\mathrm{s})\mathfrak{c}_\mathrm{c})c_\mathrm{g}\varepsilon_\mathrm{g} + \tfrac{1-\mathfrak{c}_\mathrm{s}}{\tau}\mathfrak{c}_\mathrm{c}c_\mathrm{p}\varepsilon_\mathrm{p} \leq \tfrac{1-\mathfrak{c}_\mathrm{s}}{\tau}\mathfrak{c}_\mathrm{c} \cdot \chi(u^k), \tag{4.27a}$$

$$c_\mathrm{g}\varepsilon_\mathrm{g} \leq \mathfrak{c}_\mathrm{g} \cdot \Delta_k, \tag{4.27b}$$

$$\mathfrak{c}_\mathrm{s}\mathfrak{c}_\mathrm{c}\tau c_\mathrm{g}\varepsilon_\mathrm{g} + (1 + \mathfrak{c}_\mathrm{s}\mathfrak{c}_\mathrm{c})c_\mathrm{p}\varepsilon_\mathrm{p} \leq \mathfrak{c}_\mathrm{s}\mathfrak{c}_\mathrm{c} \cdot \chi(u^k) \tag{4.27c}$$

*hold with some constants* $\mathfrak{c}_\mathrm{c}, \mathfrak{c}_\mathrm{g} > 0$.
*Then, the estimates* (4.25) *and* (4.26) *hold true with the choices* $\varrho_\mathrm{c}(t) = \mathfrak{c}_\mathrm{c}t$ *and* $\varrho_\mathrm{g}(t) = \mathfrak{c}_\mathrm{g}t$.

*Proof.* From

$$|\chi_k(0) - \chi(u^k)| \leq \|[P_{U_\mathrm{ad}} - \hat{\mathrm{P}}_{U_\mathrm{ad}}](u^k - \tau\nabla m_k(0))\|_U + \tau\|\nabla m_k(0) - \nabla\hat{J}(u^k)\|_U$$

(see the proof of Proposition 4.11), (4.23), and (4.24) it follows that

$$\chi_k(0) \geq \chi(u^k) - c_\mathrm{p}\varepsilon_\mathrm{p} - \tau c_\mathrm{g}\varepsilon_\mathrm{g}$$

holds. Computing $\nabla m_k(0)$, $\hat{\mathrm{P}}_{U_\mathrm{ad}}(u^k - \tau\nabla m_k(0))$, and $\chi_k(0)$ (in this order) according to Assumption 4.9, with $\varepsilon_\mathrm{g}$ and $\varepsilon_\mathrm{p}$ as in (4.27), we get

$$\begin{aligned}\|\nabla m_k(0) - \nabla\hat{J}(u^k)\|_U &\leq c_\mathrm{g}\varepsilon_\mathrm{g} \\ &\overset{(4.27a)}{\leq} \tfrac{1-\mathfrak{c}_\mathrm{s}}{\tau}\mathfrak{c}_\mathrm{c}\big(\chi(u^k) - c_\mathrm{p}\varepsilon_\mathrm{p} - \tau c_\mathrm{g}\varepsilon_\mathrm{g}\big) \leq \tfrac{1-\mathfrak{c}_\mathrm{s}}{\tau}\mathfrak{c}_\mathrm{c} \cdot \chi_k(0),\end{aligned}$$

$$\|\nabla m_k(0) - \nabla\hat{J}(u^k)\|_U \leq c_\mathrm{g}\varepsilon_\mathrm{g} \overset{(4.27b)}{\leq} \mathfrak{c}_\mathrm{g} \cdot \Delta_k,$$

and

$$\begin{aligned}\|[P_{U_\mathrm{ad}} - \hat{\mathrm{P}}_{U_\mathrm{ad}}](u^k - \tau\nabla m_k(0))\|_U &\leq c_\mathrm{p}\varepsilon_\mathrm{p} \\ &\overset{(4.27c)}{\leq} \mathfrak{c}_\mathrm{s}\mathfrak{c}_\mathrm{c}\big(\chi(u^k) - c_\mathrm{p}\varepsilon_\mathrm{p} - \tau c_\mathrm{g}\varepsilon_\mathrm{g}\big) \leq \mathfrak{c}_\mathrm{s}\mathfrak{c}_\mathrm{c} \cdot \chi_k(0).\end{aligned}$$

The choices $\varrho_\mathrm{c}(t) = \mathfrak{c}_\mathrm{c}t$ and $\varrho_\mathrm{g}(t) = \mathfrak{c}_\mathrm{g}t$ yield (4.25) and (4.26). □

We see that (4.27) can always be fulfilled by choosing $\varepsilon_\mathrm{g}$ and $\varepsilon_\mathrm{p}$ small enough: Given some parameters $\mathfrak{a}_1, \mathfrak{a}_2 \in [0, 1]$, we require

$$\varepsilon_\mathrm{g} \leq \min\big\{\tfrac{\mathfrak{a}_1(1-\mathfrak{c}_\mathrm{s})\mathfrak{c}_\mathrm{c}}{\tau c_\mathrm{g}(1+(1-\mathfrak{c}_\mathrm{s})\mathfrak{c}_\mathrm{c})}\chi(u^k), \tfrac{\mathfrak{c}_\mathrm{g}}{c_\mathrm{g}}\Delta_k, \tfrac{\mathfrak{a}_2}{\tau c_\mathrm{g}}\chi(u^k)\big\} \tag{4.28}$$

as well as

$$\varepsilon_\mathrm{p} \leq \min\big\{\tfrac{1-\mathfrak{a}_1}{c_\mathrm{p}}\chi(u^k), \tfrac{(1-\mathfrak{a}_2)\mathfrak{c}_\mathrm{s}\mathfrak{c}_\mathrm{c}}{(1+\mathfrak{c}_\mathrm{s}\mathfrak{c}_\mathrm{c})c_\mathrm{p}}\chi(u^k)\big\} \tag{4.29}$$

to fulfill (4.27).

If we choose $\mathfrak{a}_1 := \frac{1+(1-\mathfrak{c}_\mathrm{s})\mathfrak{c}_\mathrm{c}}{1+\mathfrak{c}_\mathrm{c}} \in (0,1]$ and $\mathfrak{a}_2 := \frac{(1-\mathfrak{c}_\mathrm{s})\mathfrak{c}_\mathrm{c}}{1+\mathfrak{c}_\mathrm{c}} \in [0,1)$, the bound (4.28) becomes

$$\varepsilon_\mathrm{g} \leq \min\bigl\{(1-\mathfrak{c}_\mathrm{s})\tfrac{\mathfrak{c}_\mathrm{c}}{1+\mathfrak{c}_\mathrm{c}} \cdot \tfrac{\chi(u^k)}{\tau c_\mathrm{g}}, \tfrac{\mathfrak{c}_\mathrm{g}}{c_\mathrm{g}}\Delta_k\bigr\}$$

and the bound (4.29) becomes

$$\varepsilon_\mathrm{p} \leq \mathfrak{c}_\mathrm{s}\tfrac{\mathfrak{c}_\mathrm{c}}{1+\mathfrak{c}_\mathrm{c}} \cdot \tfrac{\chi(u^k)}{c_\mathrm{p}},$$

i. e., we balance the bounds involving $\chi(u^k)$ by this choice.

As long as $\chi(u^k) > 0$ and $\mathfrak{c}_\mathrm{s} \in (0,1)$, the respective bounds on $\varepsilon_\mathrm{g}$ and $\varepsilon_\mathrm{p}$ are positive. From $\varepsilon_\mathrm{g} > 0$ and $\varepsilon_\mathrm{p} > 0$ it then follows that $\chi_k(0) > 0$ must hold at the end of the refinement procedure by (4.25) and (4.26).

In the case $\chi(u^k) = 0$ it could happen that an adaptive algorithm for computing $\nabla m_k(0)$ and the projection keeps increasing the accuracy, i. e., decreasing $\varepsilon_\mathrm{g}$ and $\varepsilon_\mathrm{p}$ towards 0, without being able to fulfill (4.25) and (4.26). For theoretical considerations, we assume then, e. g., that the exact gradient and projection are used in this case. In a practical implementation, the refinement procedure should be stopped if $\chi_k(0) \leq c_\chi\varepsilon_\mathrm{tol}$, $\varepsilon_\mathrm{g} \leq \frac{\varepsilon_\mathrm{tol}}{\tau}$, and $\varepsilon_\mathrm{p} \leq \varepsilon_\mathrm{tol}$ holds for some tolerance $\varepsilon_\mathrm{tol} > 0$ and a constant $c_\chi > 0$, because then

$$\begin{aligned}\chi(u^k) &\leq \chi_k(0) + |\chi(u^k) - \chi_k(0)| \leq c_\chi\varepsilon_\mathrm{tol} + c_\mathrm{p}\varepsilon_\mathrm{p} + \tau c_\mathrm{g}\varepsilon_\mathrm{g}\\ &\leq (c_\chi + c_\mathrm{p} + c_\mathrm{g})\varepsilon_\mathrm{tol}.\end{aligned}$$

*Exact computation of the projection:* When we consider a box-constrained problem in $L^2(\Omega_u)$ and a discretization of the control space by linear finite elements for example, we have to evaluate the exact projection to compute the error in $\hat{\mathrm{P}}_{U_\mathrm{ad}}$. Therefore, it makes sense to compute the inexact criticality measure directly with the exact projection, i. e., $\chi_k(0) = \tilde{\chi}_k(0)$, but not to refine the $U$-grid in the following to save computational costs. The bound (4.26) is then always satisfied and it remains to fulfill (4.25) with $\chi_k \equiv \tilde{\chi}_k$ and $\mathfrak{c}_\mathrm{s} = 0$. By setting $\varepsilon_\mathrm{p} = 0$, we arrive at the bound

$$\varepsilon_\mathrm{g} \leq \min\bigl\{\tfrac{\mathfrak{c}_\mathrm{c}}{\mathfrak{c}_\mathrm{c}+1} \cdot \tfrac{\chi(u^k)}{\tau c_\mathrm{g}}, \tfrac{\mathfrak{c}_\mathrm{g}}{c_\mathrm{g}}\Delta_k\bigr\}$$

for $\varepsilon_\mathrm{g}$, which yields that (4.25) holds with $\mathfrak{c}_\mathrm{s} = 0$. Again, the bound $\frac{\mathfrak{c}_\mathrm{c}}{\mathfrak{c}_\mathrm{c}+1} \cdot \frac{\chi(u^k)}{\tau c_\mathrm{g}}$ on $\varepsilon_\mathrm{g}$ cannot be computed explicitly, but is positive as long as $u^k$ is not stationary. If the model gradient is not computed exactly, $\chi_k(0) > 0$ holds at the end of the refinement procedure.

Often, we do not have access to the constant $c_\mathrm{g}$ in (4.23), but only can compute the error estimate $\varepsilon_\mathrm{g}$. Therefore, we set $\mathfrak{c}_\mathrm{c} := c_\mathrm{g}\tau\tilde{\mathfrak{c}}_\mathrm{c}$ and $\mathfrak{c}_\mathrm{g} := c_\mathrm{g}\tilde{\mathfrak{c}}_\mathrm{g}$ for two constants $\tilde{\mathfrak{c}}_\mathrm{c}, \tilde{\mathfrak{c}}_\mathrm{g} > 0$ and require

$$\varepsilon_\mathrm{g} \leq \min\bigl\{\tilde{\mathfrak{c}}_\mathrm{c}\chi_k(0), \tilde{\mathfrak{c}}_\mathrm{g}\Delta_k\bigr\}$$

to ensure (4.25) with $\mathfrak{c}_\mathrm{s} = 0$, $\varrho_\mathrm{c}(t) = \mathfrak{c}_\mathrm{c}t$, and $\varrho_\mathrm{g}(t) = \mathfrak{c}_\mathrm{g}t$.

### Generalized Cauchy Point

Condition (4.13) will be satisfied by a generalized Cauchy point $s_{\mathrm{C}}^k \in U_{\mathrm{ad}} - u^k$, $\|s_{\mathrm{C}}^k\|_U \leq \Delta_k$, which is computed by some type of linesearch with an inexact, and possibly refined projection $\hat{\mathrm{P}}_{U_{\mathrm{ad}}}$. It is very important to permit an inexact projection in this procedure because then $U$-grid refinement may not be necessary in every iteration. In this way, computational cost can be saved and the quality of the FE grid can be preserved by a suitable refinement method instead of adapting the refinement exactly to the projection which has to be computed. Based on the generalized Cauchy point, the decrease condition (4.13) can be evaluated for improved trial steps. One can use an Armijo type linesearch on a line segment or a projected linesearch for example. We describe the latter here, following [118].

**Lemma 4.13.** *Given $u^k \in U_{\mathrm{ad}}$, $\nabla m_k(0) \in U$ and $t > 0$, the direction*

$$p^k = p^k(t) := \hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t)) - u^k$$

*with $w^k = w^k(t) := u^k - t\nabla m_k(0)$ is a descent direction for $m_k$ in 0 in the sense that $(\nabla m_k(0), p^k)_U \leq -\frac{\mathfrak{c}_{\mathrm{i}}}{t}\|p^k\|_U^2$, provided the inexact projection $\hat{\mathrm{P}}_{U_{\mathrm{ad}}}$ satisfies*

$$\big(\hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t)) - w^k(t), \hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t)) - u^k\big)_U \leq (1 - \mathfrak{c}_{\mathrm{i}})\|p^k(t)\|_U^2 \tag{4.30}$$

*for some arbitrary, but fixed constant $\mathfrak{c}_{\mathrm{i}} \in (0, 1]$ and $\|p^k\|_U > 0$.*

*Proof.* We estimate

$$\begin{aligned} t\big(\nabla m_k(0), p^k\big)_U &= \big(u^k - w^k, \hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k) - u^k\big)_U \\ &= -\|p^k\|_U^2 + \big(\hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k) - w^k, \hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k) - u^k\big)_U \leq -\mathfrak{c}_{\mathrm{i}}\|p^k\|_U^2. \end{aligned}$$

Due to $t > 0$, $\mathfrak{c}_{\mathrm{i}} > 0$ and $\|p^k\| > 0$, $p^k$ is a descent direction for $m_k$ in 0. □

**Remark 4.14.** This lemma is essentially different from [118, Lem. 5.3] in the sense that we do not use any projection property of $\hat{\mathrm{P}}_{U_{\mathrm{ad}}}$. If the discrete projection is used, (4.30) is trivially fulfilled.

**Lemma 4.15.** *Let $\hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k) \in U_{\mathrm{ad}}$ be computed such that (4.24) holds with $w^k(t) = w^k$ and with $\varepsilon_{\mathrm{p}}$ fulfilling*

$$\big(\mathfrak{c}_{\mathrm{i}} c_{\mathrm{p}}\varepsilon_{\mathrm{p}} + \|2\mathfrak{c}_{\mathrm{i}} P_{U_{\mathrm{ad}}}(w^k) - w^k + (1 - 2\mathfrak{c}_{\mathrm{i}})u^k\|_U\big)c_{\mathrm{p}}\varepsilon_{\mathrm{p}} \leq (1 - \mathfrak{c}_{\mathrm{i}})\|P_{U_{\mathrm{ad}}}(w^k) - u^k\|_U^2. \tag{4.31}$$

*Then, (4.30) holds.*

*Proof.* With the choice of $\varepsilon_{\mathrm{p}}$ and writing $P = P_{U_{\mathrm{ad}}}(w^k)$, $\hat{\mathrm{P}} = \hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k)$, $w = w^k$, $u = u^k$ we estimate

$$\begin{aligned} &\big(\hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k) - w^k, \hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k) - u^k\big)_U = \big(\hat{\mathrm{P}} - w, \hat{\mathrm{P}} - u\big)_U \\ &= \big(\hat{\mathrm{P}} - P + P - w, \hat{\mathrm{P}} - P + P - u\big)_U \\ &= \big\|\hat{\mathrm{P}} - P\big\|_U^2 + \big(\hat{\mathrm{P}} - P, 2P - w - u\big)_U + \big(P - w, P - u\big)_U \end{aligned}$$

$$
\begin{aligned}
&= (1-\mathfrak{c}_\mathrm{i})\left[\big\|\hat{\mathrm{P}}-P\big\|_U^2+\big(\hat{\mathrm{P}}-P,2P-2u\big)_U+\big\|P-u\big\|_U^2\right]+\mathfrak{c}_\mathrm{i}\big\|\hat{\mathrm{P}}-P\big\|_U^2\\
&\quad+\big(\hat{\mathrm{P}}-P,2\mathfrak{c}_\mathrm{i}P-w+(1-2\mathfrak{c}_\mathrm{i})u\big)_U+\big(\mathfrak{c}_\mathrm{i}P-w+(1-\mathfrak{c}_\mathrm{i})u,P-u\big)_U\\
&= (1-\mathfrak{c}_\mathrm{i})\big(\hat{\mathrm{P}}-P+P-u,\hat{\mathrm{P}}-P+P-u\big)_U+\mathfrak{c}_\mathrm{i}\big\|\hat{\mathrm{P}}-P\big\|_U^2\\
&\quad+\big(\hat{\mathrm{P}}-P,2\mathfrak{c}_\mathrm{i}P-w+(1-2\mathfrak{c}_\mathrm{i})u\big)_U\\
&\quad+\big(P-w,P-u\big)_U-(1-\mathfrak{c}_\mathrm{i})\big(P-u,P-u\big)_U\\
&\le (1-\mathfrak{c}_\mathrm{i})\big\|p^k\big\|_U^2+\mathfrak{c}_\mathrm{i}\big\|\hat{\mathrm{P}}-P\big\|_U^2+\big\|\hat{\mathrm{P}}-P\big\|_U\cdot\|2\mathfrak{c}_\mathrm{i}P-w+(1-2\mathfrak{c}_\mathrm{i})u\|_U\\
&\qquad-(1-\mathfrak{c}_\mathrm{i})\big\|P-u\big\|_U^2\\
&\overset{(4.24)}{\le} (1-\mathfrak{c}_\mathrm{i})\big\|p^k\big\|_U^2+\mathfrak{c}_\mathrm{i}c_\mathrm{p}^2\varepsilon_\mathrm{p}^2+c_\mathrm{p}\varepsilon_\mathrm{p}\cdot\|2\mathfrak{c}_\mathrm{i}P-w+(1-2\mathfrak{c}_\mathrm{i})u\big\|_U\\
&\qquad-(1-\mathfrak{c}_\mathrm{i})\big\|P-u\big\|_U^2\\
&\overset{(4.31)}{\le} (1-\mathfrak{c}_\mathrm{i})\big\|p^k\big\|_U^2.
\end{aligned}
$$

For the first estimate we have used the variational inequality property of the exact projection and the Cauchy-Schwarz inequality. □

**Remark 4.16.** If $\tilde{\chi}_k(0)>0$, then $\|P_{U_\mathrm{ad}}(w^k(t))-u^k\|_U>0$ holds for every $t>0$: If there existed $t>0$ such that $\|P_{U_\mathrm{ad}}(w^k(t))-u^k\|_U=0$, then $0\in U$ would be first-order stationary for problem (4.8), which would yield $\tilde{\chi}_k(0)=0$. Therefore, the bound on the right hand side of (4.31) can be satisfied by taking $\varepsilon_\mathrm{p}$ small enough. If we use the exact projection to compute $\chi_k(0)=\tilde{\chi}_k(0)$, $\tilde{\chi}_k(0)>0$ is ensured when performing the linesearch by the stopping criterion in Algorithm 1.

**Definition 4.17** (Armijo condition)**.** Choose the largest $t_k\in\{\mathfrak{c}_{\mathrm{a},k}\cdot\mathfrak{c}_\mathrm{f}^j, j\in\mathbb{N}_0\}$ such that

$$
m_k(p^k(t_k))\le m_k(0)-\tfrac{\mathfrak{c}_\mathrm{i}\mathfrak{c}_\mathrm{e}}{t_k}\|p^k(t_k)\|_U^2, \tag{4.32a}
$$

$$
\|p^k(t_k)\|_U\le\Delta_k \tag{4.32b}
$$

with two fixed parameters $\mathfrak{c}_\mathrm{f},\mathfrak{c}_\mathrm{e}\in(0,1)$ are satisfied. Here, $\mathfrak{c}_{\mathrm{a},k}\in\mathbb{R}$ are constants satisfying $\mathfrak{c}_\mathrm{a}\le\mathfrak{c}_{\mathrm{a},k}\le\tau$ for all $k\in\mathbb{N}_0$ with some $\mathfrak{c}_\mathrm{a}\in(0,\tau]$.

**Remark 4.18.** Note that

$$
-\tfrac{\mathfrak{c}_\mathrm{i}\mathfrak{c}_\mathrm{e}}{t_k}\|p^k(t_k)\|_U^2\ge\mathfrak{c}_\mathrm{e}(\nabla m_k(0),p^k(t_k))_U
$$

holds by Lemma 4.13 if $\hat{\mathrm{P}}_{U_\mathrm{ad}}$ satisfies (4.30). Then, condition (4.32a) is implied by

$$
m_k(p^k(t_k))\le m_k(0)+\mathfrak{c}_\mathrm{e}(\nabla m_k(0),p^k(t_k))_U.
$$

**Lemma 4.19.** *If Assumption 4.5 holds and if $p^k(t_k)$ is computed according to Lemma 4.13, condition* (4.32a) *is satisfied for all $t_k\in\big(0,\frac{2(1-\mathfrak{c}_\mathrm{e})\mathfrak{c}_\mathrm{i}}{\mathfrak{c}_{m_k}}\big]$, where $\mathfrak{c}_{m_k}>0$ is the Lipschitz constant of the model gradient $\nabla m_k$.*

*Proof.* We estimate using the fundamental theorem of calculus:

$$\begin{aligned} m_k(p^k(t_k)) - m_k(0) &= \int_0^1 \big(\nabla m_k(\sigma p^k(t_k)), p^k(t_k)\big)_U \,\mathrm{d}\sigma \\ &\leq \big(\nabla m_k(0), p^k(t_k)\big)_U + \int_0^1 \big\|\nabla m_k(\sigma p^k(t_k)) - \nabla m_k(0)\big\|_U \cdot \big\|p^k(t_k)\big\|_U \,\mathrm{d}\sigma \\ &\leq -\tfrac{\mathfrak{c}_\mathrm{i}}{t_k}\|p^k(t_k)\|_U^2 + \tfrac{\mathfrak{c}_{m_k}}{2}\|p^k(t_k)\|_U^2 = \big(-\tfrac{\mathfrak{c}_\mathrm{i}}{t_k} + \tfrac{\mathfrak{c}_{m_k}}{2}\big)\|p^k(t_k)\|_U^2. \end{aligned}$$

In the last estimate we have used Lemma 4.13 and the Lipschitz continuity of $\nabla m_k$. With the given choice of $t_k$, (4.32a) follows. □

**Lemma 4.20.** *Let Assumption 4.5 hold and let $t_k$ be computed according to the Armijo condition (Definition 4.17), where the inexact projection $\hat{\mathrm{P}}_{U_\mathrm{ad}}$ satisfies* (4.30) *and additionally*

$$\|P_{U_\mathrm{ad}}(w^k(t)) - \hat{\mathrm{P}}_{U_\mathrm{ad}}(w^k(t))\|_U \leq \tfrac{1-\mathfrak{c}_{\mathrm{l}1}}{\mathfrak{c}_{\mathrm{l}1}}\|\hat{\mathrm{P}}_{U_\mathrm{ad}}(w^k(t)) - u^k\|_U \tag{4.33}$$

*as well as*

$$\|P_{U_\mathrm{ad}}(w^k(t)) - \hat{\mathrm{P}}_{U_\mathrm{ad}}(w^k(t))\|_U \leq \tfrac{1-\mathfrak{c}_{\mathrm{l}2}}{\mathfrak{c}_{\mathrm{l}2}}\|P_{U_\mathrm{ad}}(w^k(t)) - u^k\|_U \tag{4.34}$$

*for every $t = t_k$ tested during the Armijo linesearch with some constants $\mathfrak{c}_{\mathrm{l}1}, \mathfrak{c}_{\mathrm{l}2} \in (0,1]$, cf. [118, Eq. (3.24)].*
*Then, the trial step $s_\mathrm{C}^k := p^k(t_k)$ fulfills condition* (4.13) *with*

$$\varrho_{\mathrm{t}1}(t) := \tfrac{\mathfrak{c}_\mathrm{i}\mathfrak{c}_{\mathrm{l}1}^2\mathfrak{c}_{\mathrm{l}2}\mathfrak{c}_\mathrm{e}\mathfrak{c}_\mathrm{f}}{\tau}t, \qquad \varrho_{\mathrm{t}2}(t) := \tfrac{1}{\mathfrak{c}_{\mathrm{l}2}\tau}\cdot\min\big\{\tfrac{2(1-\mathfrak{c}_\mathrm{e})\mathfrak{c}_\mathrm{i}}{\mathfrak{c}_m}, \tfrac{\mathfrak{c}_\mathrm{a}}{\mathfrak{c}_\mathrm{f}}\big\}t,$$

*and $\chi_k(0) = \tilde{\chi}_k(0)$.*

*Proof.* From (4.33) we get that

$$\begin{aligned} &\mathfrak{c}_{\mathrm{l}1}\|P_{U_\mathrm{ad}}(w^k(t)) - u^k\|_U \\ &\leq \mathfrak{c}_{\mathrm{l}1}\|P_{U_\mathrm{ad}}(w^k(t)) - \hat{\mathrm{P}}_{U_\mathrm{ad}}(w^k(t))\|_U + \mathfrak{c}_{\mathrm{l}1}\|\hat{\mathrm{P}}_{U_\mathrm{ad}}(w^k(t)) - u^k\|_U \\ &\leq (1-\mathfrak{c}_{\mathrm{l}1})\|\hat{\mathrm{P}}_{U_\mathrm{ad}}(w^k(t)) - u^k\|_U + \mathfrak{c}_{\mathrm{l}1}\|\hat{\mathrm{P}}_{U_\mathrm{ad}}(w^k(t)) - u^k\|_U \\ &= \|\hat{\mathrm{P}}_{U_\mathrm{ad}}(w^k(t)) - u^k\|_U = \|p^k(t)\|_U \end{aligned} \tag{4.35}$$

for every tested $t = t_k$. Using

$$\tfrac{1}{t}\|P_{U_\mathrm{ad}}(w^k(t)) - u^k\|_U \geq \tfrac{1}{\tau}\tilde{\chi}_k(0) \tag{4.36}$$

for $t \leq \tau$ by [60, Lem. 1.10 (e)], which states that the function $\phi(t) := \frac{1}{t}\|P_{U_\mathrm{ad}}(u^k - t\nabla m_k(0)) - u^k\|_U$ $(t > 0)$ is non-increasing, and (4.32a) as well as (4.35), we conclude that

$$\begin{aligned} \mathrm{pred}_k = m_k(0) - m_k(p^k(t_k)) &\geq \tfrac{\mathfrak{c}_\mathrm{i}\mathfrak{c}_\mathrm{e}}{t_k}\|p^k(t_k)\|_U^2 \\ &\geq \mathfrak{c}_\mathrm{i}\mathfrak{c}_\mathrm{e}\mathfrak{c}_{\mathrm{l}1}\cdot\tfrac{1}{t_k}\|P_{U_\mathrm{ad}}(w^k(t_k)) - u^k\|_U \cdot \|p^k(t_k)\|_U \\ &\geq \mathfrak{c}_\mathrm{i}\mathfrak{c}_\mathrm{e}\mathfrak{c}_{\mathrm{l}1}\cdot\tfrac{1}{\tau}\tilde{\chi}_k(0)\cdot\|p^k(t_k)\|_U \end{aligned} \tag{4.37}$$

holds by $t_k \leq \mathfrak{c}_{\mathrm{a},k} \leq \tau$.

Now consider the case that $t_k$ found by the standard projected Armijo linesearch for (4.32a) already satisfies (4.32b). Thus, Lemma 4.19 can be applied and $t_k \geq \min\{\frac{2(1-\mathfrak{c}_\mathrm{e})\mathfrak{c}_\mathrm{i}\mathfrak{c}_\mathrm{f}}{\mathfrak{c}_{m_k}}, \mathfrak{c}_{\mathrm{a},k}\}$ holds. From (4.37), (4.35), and (4.36) it now follows that

$$\begin{aligned}
\mathrm{pred}_k &\geq \tfrac{\mathfrak{c}_\mathrm{i}\mathfrak{c}_\mathrm{e}\mathfrak{c}_{\mathrm{l}1}}{\tau}\tilde{\chi}_k(0) \cdot \tfrac{\mathfrak{c}_{\mathrm{l}1}t_k}{\tau}\tilde{\chi}_k(0) \\
&\geq \tfrac{\mathfrak{c}_\mathrm{i}\mathfrak{c}_\mathrm{e}\mathfrak{c}_{\mathrm{l}1}}{\tau}\tilde{\chi}_k(0) \cdot \min\{\tfrac{2(1-\mathfrak{c}_\mathrm{e})\mathfrak{c}_\mathrm{i}\mathfrak{c}_\mathrm{f}}{\mathfrak{c}_{m_k}}, \mathfrak{c}_{\mathrm{a},k}\} \cdot \tfrac{\mathfrak{c}_{\mathrm{l}1}}{\tau}\tilde{\chi}_k(0) \\
&= \tfrac{\mathfrak{c}_\mathrm{i}\mathfrak{c}_{\mathrm{l}1}^2\mathfrak{c}_{\mathrm{l}2}\mathfrak{c}_\mathrm{e}\mathfrak{c}_\mathrm{f}}{\tau}\tilde{\chi}_k(0) \cdot \tfrac{1}{\mathfrak{c}_{\mathrm{l}2}\tau} \cdot \min\{\tfrac{2(1-\mathfrak{c}_\mathrm{e})\mathfrak{c}_\mathrm{i}}{\mathfrak{c}_m}, \tfrac{\mathfrak{c}_\mathrm{a}}{\mathfrak{c}_\mathrm{f}}\}\tilde{\chi}_k(0) \\
&= \varrho_{\mathrm{t}1}(\tilde{\chi}_k(0)) \cdot \varrho_{\mathrm{t}2}(\tilde{\chi}_k(0)) \geq \varrho_{\mathrm{t}1}(\tilde{\chi}_k(0)) \cdot \min\{\varrho_{\mathrm{t}2}(\tilde{\chi}_k(0)), \Delta_k\}.
\end{aligned}$$

In the case that the standard search for (4.32a) does not yield $t_k$ satisfying (4.32b), $t_k$ has to be decreased further. It follows that $\|p^k(\frac{t_k}{\mathfrak{c}_\mathrm{f}})\|_U > \Delta_k$.
In analogy to (4.35) we can conclude from (4.34) that

$$\mathfrak{c}_{\mathrm{l}2}\|p^k(t)\|_U = \mathfrak{c}_{\mathrm{l}2}\|\hat{\mathrm{P}}_{U_\mathrm{ad}}(w^k(t)) - u^k\|_U \leq \|P_{U_\mathrm{ad}}(w^k(t)) - u^k\|_U \tag{4.38}$$

for every tested $t = t_k$. With (4.35), [60, Lem. 1.10 (e)], and (4.38) we obtain

$$\begin{aligned}
\|p^k(t_k)\|_U &\geq \mathfrak{c}_{\mathrm{l}1}\|P_{U_\mathrm{ad}}(w^k(t_k)) - u^k\|_U \\
&\geq \mathfrak{c}_{\mathrm{l}1}\mathfrak{c}_\mathrm{f}\|P_{U_\mathrm{ad}}(w^k(\tfrac{t_k}{\mathfrak{c}_\mathrm{f}})) - u^k\|_U \\
&\geq \mathfrak{c}_{\mathrm{l}1}\mathfrak{c}_{\mathrm{l}2}\mathfrak{c}_\mathrm{f}\|p^k(\tfrac{t_k}{\mathfrak{c}_\mathrm{f}})\|_U > \mathfrak{c}_{\mathrm{l}1}\mathfrak{c}_{\mathrm{l}2}\mathfrak{c}_\mathrm{f}\Delta_k.
\end{aligned} \tag{4.39}$$

Hence, by (4.37), we get

$$\begin{aligned}
\mathrm{pred}_k &\geq \tfrac{\mathfrak{c}_\mathrm{i}\mathfrak{c}_\mathrm{e}\mathfrak{c}_{\mathrm{l}1}}{\tau}\tilde{\chi}_k(0) \cdot \|p^k(t_k)\|_U \\
&> \tfrac{\mathfrak{c}_\mathrm{i}\mathfrak{c}_\mathrm{e}\mathfrak{c}_{\mathrm{l}1}}{\tau}\tilde{\chi}_k(0) \cdot \mathfrak{c}_{\mathrm{l}1}\mathfrak{c}_{\mathrm{l}2}\mathfrak{c}_\mathrm{f}\Delta_k \\
&= \varrho_{\mathrm{t}1}(\tilde{\chi}_k(0)) \cdot \Delta_k \geq \varrho_{\mathrm{t}1}(\tilde{\chi}_k(0)) \cdot \min\{\varrho_{\mathrm{t}2}(\tilde{\chi}_k(0)), \Delta_k\}.
\end{aligned}$$

Therefore, in both cases, (4.13) is satisfied with $\chi_k(0) = \tilde{\chi}_k(0)$.

□

**Remark 4.21.**

- If (4.33) holds with $\mathfrak{c}_{\mathrm{l}1} \in (\frac{1}{2}, 1]$, (4.34) follows if we choose $\mathfrak{c}_{\mathrm{l}2} = \frac{2\mathfrak{c}_{\mathrm{l}1}-1}{\mathfrak{c}_{\mathrm{l}1}} \in (0, 1]$: From (4.33), we obtain

$$\begin{aligned}
&\|P_{U_\mathrm{ad}}(w^k(t)) - \hat{\mathrm{P}}_{U_\mathrm{ad}}(w^k(t))\|_U \\
\leq\ &\tfrac{1-\mathfrak{c}_{\mathrm{l}1}}{\mathfrak{c}_{\mathrm{l}1}}\|\hat{\mathrm{P}}_{U_\mathrm{ad}}(w^k(t)) - u^k\|_U \\
\leq\ &\tfrac{1-\mathfrak{c}_{\mathrm{l}1}}{\mathfrak{c}_{\mathrm{l}1}}\left(\|\hat{\mathrm{P}}_{U_\mathrm{ad}}(w^k(t)) - P_{U_\mathrm{ad}}(w^k(t))\|_U + \|P_{U_\mathrm{ad}}(w^k(t)) - u^k\|_U\right).
\end{aligned}$$

  This is equivalent to

$$\tfrac{2\mathfrak{c}_{\mathrm{l}1}-1}{\mathfrak{c}_{\mathrm{l}1}}\|P_{U_\mathrm{ad}}(w^k(t)) - \hat{\mathrm{P}}_{U_\mathrm{ad}}(w^k(t))\|_U \leq \tfrac{1-\mathfrak{c}_{\mathrm{l}1}}{\mathfrak{c}_{\mathrm{l}1}}\|P_{U_\mathrm{ad}}(w^k(t)) - u^k\|_U,$$

  which yields (4.34) with the given value of $\mathfrak{c}_{\mathrm{l}2}$.

- The prerequisite (4.34) in Lemma 4.20 can be dropped if the discrete projection $\hat{\mathrm{P}}_{U_{\mathrm{ad}}}$ onto $U_{\mathrm{ad}} \cap \mathrm{U}$ is used (cf. [118]) since then

$$\|p^k(t_k)\|_U \geq \mathfrak{c}_{\mathrm{f}} \|p^k(\tfrac{t_k}{\mathfrak{c}_{\mathrm{f}}})\|_U$$

  holds by the projection property of $\hat{\mathrm{P}}_{U_{\mathrm{ad}}}$, which we do not assume in general, and then replaces (4.39). The constants in $\varrho_{\mathrm{t}1}$ and $\varrho_{\mathrm{t}2}$ change accordingly.

- Instead of (4.33) and (4.34) one can directly assume (4.35) and (4.38) in Lemma 4.20, but these conditions can be more difficult to evaluate than (4.33) and (4.34). In particular, the verification of condition (4.35) only requires the estimation of the projection error and the evaluation of the inexact projection.

- Given $\mathfrak{c}_{\mathrm{a}} \in (0, \tau]$, it makes sense to choose

$$\mathfrak{c}_{\mathrm{a},k} := \max\{\mathfrak{c}_{\mathrm{a}}, \min\{\tau, \tfrac{\mathfrak{c}_{\mathrm{l}2}\Delta_k}{\|\nabla m_k(0)\|_U}\}\} \tag{4.40}$$

  in the Armijo condition because with $t_k \leq \frac{\mathfrak{c}_{\mathrm{l}2}\Delta_k}{\|\nabla m_k(0)\|_U}$, which is not always ensured by (4.40), and (4.38) we obtain

$$\begin{aligned}\|p^k(t_k)\|_U &\leq \tfrac{1}{\mathfrak{c}_{\mathrm{l}2}} \|P_{U_{\mathrm{ad}}}(u^k - t_k \nabla m_k(0)) - u^k\|_U \\ &= \tfrac{1}{\mathfrak{c}_{\mathrm{l}2}} \|P_{U_{\mathrm{ad}}}(u^k - t_k \nabla m_k(0)) - P_{U_{\mathrm{ad}}}(u^k)\|_U \\ &\leq \tfrac{t_k}{\mathfrak{c}_{\mathrm{l}2}} \|\nabla m_k(0)\|_U \leq \Delta_k,\end{aligned}$$

  which is exactly (4.32b). Note that this choice of $t_k$ is not necessary (only sufficient) for (4.32b). This observation yields together with Lemma 4.19 that the projected Armijo linesearch terminates after finitely many times decreasing $t_k$.

**Lemma 4.22.** *Let $t \leq \tau$ and let $\hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t)) \in U_{\mathrm{ad}}$ be computed such that* (4.24) *holds with $\varepsilon_{\mathrm{p}}$ fulfilling*

$$\varepsilon_{\mathrm{p}} \leq \tfrac{(1-\mathfrak{c}_{\mathrm{l}1})t}{c_{\mathrm{p}}\tau} \tilde{\chi}_k(0). \tag{4.41}$$

*Then,* (4.33) *holds.*

*Proof.* We write $\hat{\mathrm{P}} = \hat{\mathrm{P}}_{U_{\mathrm{ad}}}(w^k(t))$, $P = P_{U_{\mathrm{ad}}}(w^k(t))$, and $u = u^k$ and get

$$\begin{aligned}\tfrac{1-\mathfrak{c}_{\mathrm{l}1}}{\mathfrak{c}_{\mathrm{l}1}} \|\hat{\mathrm{P}} - u\|_U &\geq \tfrac{1-\mathfrak{c}_{\mathrm{l}1}}{\mathfrak{c}_{\mathrm{l}1}} \big(\|P - u\|_U - \|P - \hat{\mathrm{P}}\|_U\big) \\ &\geq \tfrac{1-\mathfrak{c}_{\mathrm{l}1}}{\mathfrak{c}_{\mathrm{l}1}} \big(\tfrac{t}{\tau}\tilde{\chi}_k(0) - c_{\mathrm{p}}\varepsilon_{\mathrm{p}}\big) \overset{(4.41)}{\geq} \tfrac{c_{\mathrm{p}}}{\mathfrak{c}_{\mathrm{l}1}}\varepsilon_{\mathrm{p}} - \tfrac{1-\mathfrak{c}_{\mathrm{l}1}}{\mathfrak{c}_{\mathrm{l}1}} c_{\mathrm{p}}\varepsilon_{\mathrm{p}} \\ &= c_{\mathrm{p}}\varepsilon_{\mathrm{p}} \overset{(4.24)}{\geq} \|\hat{\mathrm{P}} - P\|_U,\end{aligned}$$

which proves (4.33). In the second estimate we have used (4.24) and (4.36). □

**Remark 4.23.** In the same way, we obtain that (4.34) is satisfied if

$$\varepsilon_{\mathrm{p}} \leq \tfrac{(1-\mathfrak{c}_{\mathrm{l}2})t}{\mathfrak{c}_{\mathrm{l}2}c_{\mathrm{p}}\tau} \tilde{\chi}_k(0).$$

Having computed the generalized Cauchy point $s_{\mathrm{C}}^k$ yields a simple criterion for (4.13):

**Lemma 4.24** (Fraction of generalized Cauchy decrease)**.** *Let $s_{\mathrm{C}}^k$ be computed according to Lemma 4.20 and let $s^k \in U_{\mathrm{ad}} - u^k$ satisfy*

$$m_k(0) - m_k(s^k) \geq \mathfrak{c}_{\mathrm{d}}(m_k(0) - m_k(s_{\mathrm{C}}^k))$$

*with some $\mathfrak{c}_{\mathrm{d}} \in (0, 1]$.*
*Then, $s^k$ satisfies* (4.13) *with $\chi_k(0) = \tilde{\chi}_k(0)$ and*

$$\varrho_{\mathrm{t}1}(t) := \tfrac{\mathfrak{c}_{\mathrm{d}}\mathfrak{c}_{\mathrm{i}}\mathfrak{c}_{\mathrm{l}1}^2\mathfrak{c}_{\mathrm{l}2}\mathfrak{c}_{\mathrm{e}}\mathfrak{c}_{\mathrm{f}}}{\tau}t, \qquad \varrho_{\mathrm{t}2}(t) := \tfrac{1}{\mathfrak{c}_{\mathrm{l}2}\tau} \cdot \min\{\tfrac{2(1-\mathfrak{c}_{\mathrm{e}})\mathfrak{c}_{\mathrm{i}}}{\mathfrak{c}_m}, \tfrac{\mathfrak{c}_{\mathrm{a}}}{\mathfrak{c}_{\mathrm{f}}}\}t.$$

*Proof.* From Lemma 4.20 we know that $s_{\mathrm{C}}^k$ satisfies (4.13) with

$$\varrho_{\mathrm{t}1}(t) := \tfrac{\mathfrak{c}_{\mathrm{i}}\mathfrak{c}_{\mathrm{l}1}^2\mathfrak{c}_{\mathrm{l}2}\mathfrak{c}_{\mathrm{e}}\mathfrak{c}_{\mathrm{f}}}{\tau}t, \qquad \varrho_{\mathrm{t}2}(t) := \tfrac{1}{\mathfrak{c}_{\mathrm{l}2}\tau} \cdot \min\{\tfrac{2(1-\mathfrak{c}_{\mathrm{e}})\mathfrak{c}_{\mathrm{i}}}{\mathfrak{c}_m}, \tfrac{\mathfrak{c}_{\mathrm{a}}}{\mathfrak{c}_{\mathrm{f}}}\}t,$$

and $\chi_k(0) = \tilde{\chi}_k(0)$. The stated result follows immediately. □

### Computed Reduction

The computed reduction is only evaluated as long as $\chi_k(0)$ is positive, which is ensured by the stopping criterion of Algorithm 1. Then the predicted reduction $\mathrm{pred}_k$ is positive by (4.13). The inequality (4.12) can be reduced to a bound on the inexact objective function evaluation:

**Lemma 4.25.** *Let*

$$|\hat{J}(u) - \hat{J}_k(u)| \leq \tfrac{1}{2}\varrho_{\mathrm{r}}(\eta_3 \min\{\mathrm{pred}_k, \mathfrak{r}_k\})$$

*hold for all $u \in \{u^k, u^k + s^k\} \subset U$. Then, condition* (4.12) *is fulfilled.*

*Proof.* Using the definitions (4.6a) and (4.6b), we estimate

$$|\mathrm{ared}_k - \mathrm{cred}_k| \leq |\hat{J}(u^k) - \hat{J}_k(u^k)| + |\hat{J}_k(u^k + s^k) - \hat{J}(u^k + s^k)| \leq \varrho_{\mathrm{r}}(\eta_3 \min\{\mathrm{pred}_k, \mathfrak{r}_k\}),$$

which is (4.12). □

Choosing $\varepsilon_{\mathrm{o}} \leq \tilde{\mathfrak{c}}_{\mathrm{o}} \cdot (\eta_3 \min\{\mathrm{pred}_k, \mathfrak{r}_k\})^{\mathfrak{e}_{\mathrm{o}}} > 0$ with some constant $\tilde{\mathfrak{c}}_{\mathrm{o}} > 0$ and $\varrho_{\mathrm{r}}(t) = 2c_{\mathrm{o}}\tilde{\mathfrak{c}}_{\mathrm{o}}t^{\mathfrak{e}_{\mathrm{o}}}$, the condition given in Lemma 4.25 can be fulfilled by Assumption 4.9.

### Update of the Trust-Region Radius

Due to $\mathrm{pred}_k > 0$, the quantity $\frac{\mathrm{cred}_k}{\mathrm{pred}_k}$ is well-defined. Since we have $0 < \Delta_k \leq \Delta_{\max}$ and $0 < \nu_1 < 1 \leq \nu_2 < \nu_3$, the intervals from which the new radius $\Delta_{k+1}$ is chosen are always nonempty.

## 4.4. Solution of Subproblems by a Semismooth Newton Method

To obtain fast convergence of the algorithm, it is necessary to apply a more sophisticated solver to the trust-region subproblem (4.5). Since a semismooth Newton method [92, 110, 58, 111] gave very good results previously [46] and also in different contexts involvings PDEs [60, 111], we want to apply it here also. As it is difficult to handle, e. g., box constraints and an additional trust-region constraint at the same time, we replace the latter in (4.5) by quadratic regularization and obtain

$$\min_{s\in U} m_k(s) + \tfrac{\mathfrak{c}_{\mathrm{n},k}}{2\Delta_k}\|s\|_U^2 \quad \text{s. t.} \quad u^k + s \in U_{\mathrm{ad}} \tag{4.42}$$

with some constant $\mathfrak{c}_{\mathrm{n},k} > 0$. A solution of this problem is not necessarily feasible for (4.5). Still, there is a connection between the two problems:

**Lemma 4.26.** *Let $\bar{s} \in U$ be a global solution of (4.42) with $\|\bar{s}\|_U = \Delta_k$. Then, $\bar{s}$ solves (4.5).*

*Proof.* It holds that $m_k(\bar{s}) + \frac{\mathfrak{c}_{\mathrm{n},k}}{2\Delta_k}\|\bar{s}\|_U^2 \le m_k(s) + \frac{\mathfrak{c}_{\mathrm{n},k}}{2\Delta_k}\|s\|_U^2$ for all $s \in U$ such that $u^k + s \in U_{\mathrm{ad}}$. Thus, $m_k(\bar{s}) \le m_k(s) + \frac{\mathfrak{c}_{\mathrm{n},k}}{2\Delta_k}\big(\|s\|_U^2 - \|\bar{s}\|_U^2\big) \le m_k(s)$ for all $s \in U$ such that $u^k + s \in U_{\mathrm{ad}}$ and $\|s\|_U^2 \le \Delta_k^2$. □

For the application of semismooth Newton we follow [60, Chap. 2] and [111]. A necessary optimality condition for (4.42) is given by

$$R(\bar{s}) := \bar{s} - P_{U_{\mathrm{ad}}-u^k}\big(\bar{s} - \tau_{\mathrm{n},k}(\nabla m_k(\bar{s}) + \tfrac{\mathfrak{c}_{\mathrm{n},k}}{\Delta_k}\bar{s})\big) = 0 \tag{4.43}$$

with an arbitrary $\tau_{\mathrm{n},k} > 0$. We consider the case of a quadratic model

$$m_k(s) = m_k(0) + (\nabla m_k(0), s)_U + \tfrac{1}{2}(\nabla^2 m_k(0)s, s)_U$$

with gradient $\nabla m_k(s) = \nabla m_k(0) + \nabla^2 m_k(0)s$. In our application we have $\nabla m_k(0) = -B^*\tilde{T}(u^k) + \gamma u^k =: \hat{G}(u^k) + \gamma u^k$, where $\tilde{T}(u^k)$ is the inexact adjoint state. The Hessian of the model can be a positive multiple of the identity or a better approximation of the true Hessian. The Hessian is often of the form

$$\nabla^2 m_k(0)s = \hat{H}(u^k)s + \gamma s, \tag{4.44}$$

cf. (3.23), with $\hat{H}(u^k) \in \mathcal{L}(U,U)$. The operator $\hat{H}(u^k)$ need not necessarily correspond to the exact Hessian as derived in (3.23), but is only required to fulfill certain properties to establish semismoothness.

Choosing $\tau_{\mathrm{n},k} := (\gamma + \frac{\mathfrak{c}_{\mathrm{n},k}}{\Delta_k})^{-1}$, (4.43) becomes

$$\begin{aligned} R(\bar{s}) &= \bar{s} - P_{U_{\mathrm{ad}}-u^k}\big(\bar{s} - \tau_{\mathrm{n},k}(\nabla m_k(\bar{s}) + \tfrac{\mathfrak{c}_{\mathrm{n},k}}{\Delta_k}\bar{s})\big) \\ &= \bar{s} - P_{U_{\mathrm{ad}}-u^k}\big(\bar{s} - \tau_{\mathrm{n},k}(\nabla m_k(0) + \nabla^2 m_k(0)\bar{s} + \tfrac{\mathfrak{c}_{\mathrm{n},k}}{\Delta_k}\bar{s})\big) = \\ &= \bar{s} - P_{U_{\mathrm{ad}}-u^k}\big(-\tau_{\mathrm{n},k}(\nabla m_k(0) + \hat{H}(u^k)\bar{s})\big) \\ &= \bar{s} + u^k - P_{U_{\mathrm{ad}}}\big(u^k - \tau_{\mathrm{n},k}\nabla m_k(0) - \tau_{\mathrm{n},k}\hat{H}(u^k)\bar{s}\big) = 0. \end{aligned} \tag{4.45}$$

Now, assume that $u^k \in L^q(\Omega_u)$ and $\nabla m_k(0) \in L^q(\Omega_u)$ holds with some $q \geq 2$,[8] which can be ensured by $u^k \in U_{\mathrm{ad}} \subset L^q(\Omega_u)$ and $\hat{G}(u^k) \in L^q(\Omega_u)$ for example, as well as $\hat{H}(u^k) \in \mathcal{L}(L^2(\Omega_u), L^q(\Omega_u))$. Then,

$$f_1 : L^2(\Omega_u) \to L^q(\Omega_u), s \mapsto f_1(s) := u^k - \tau_{\mathrm{n},k} \nabla m_k(0) - \tau_{\mathrm{n},k} \hat{H}(u^k) s$$

is affine, bounded and thus $\{-\tau_{\mathrm{n},k}\hat{H}(u^k)\}$-semismooth by [111, Prop. 3.4]. If additionally $P_{U_{\mathrm{ad}}} : L^q(\Omega_u) \to L^2(\Omega_u)$ is $\partial P_{U_{\mathrm{ad}}}$-semismooth and bounded near $f_1(s)$, we can apply the chain rule [111, Prop. 3.8] to show that $R$ is $\partial R$-semismooth at $s$ with $\partial R(s) = \{I + \tau_{\mathrm{n},k} M_P \hat{H}(u^k) : M_P \in \partial P_{U_{\mathrm{ad}}}(f_1(s))\}$.

Given a current iterate $s^{k,\ell}$ and $M_P \in \partial P_{U_{\mathrm{ad}}}(f_1(s^{k,\ell}))$, the semismooth Newton equation for an update $\tilde{d}^{k,\ell}$ reads

$$\tilde{d}^{k,\ell} + \tau_{\mathrm{n},k} M_P \hat{H}(u^k) \tilde{d}^{k,\ell} = -R(s^{k,\ell}).$$

Having computed a solution of it (approximately), we set $s^{k,\ell+1} := s^{k,\ell} + \tilde{d}^{k,\ell}$.

**Remark 4.27.** To prove superlinear convergence of the semismooth Newton method, a regularity condition is required. Such a condition can be derived depending on the concrete choice of $U_{\mathrm{ad}}$ and the solution of (4.42), see [111, Sec. 9.1].

In contrast to the projected linesearch developed in Section 4.3, the (approximate) solution found by applying semismooth Newton to (4.42) may not be feasible for (4.5). Thus, we set $s^k_{\mathrm{SSN}} := P_{\{s \in U : \|s\|_U \leq \Delta_k\}}\big(P_{U_{\mathrm{ad}} - u^k}(\hat{s}^k)\big)$, where $\hat{s}^k$ is an (approximate) solution of (4.42).

**Lemma 4.28.** *Given $\hat{s} \in U$, $u \in U$ and $\Delta_k > 0$, $s := P_{\{s \in U : \|s\|_U \leq \Delta_k\}}\big(P_{U_{\mathrm{ad}} - u}(\hat{s})\big)$ is feasible for (4.5), where*

$$P_{\{s \in U : \|s\|_U \leq \Delta_k\}}(s) = \tfrac{\Delta_k}{\max\{\|s\|_U, \Delta_k\}} s.$$

*Proof.* It is obvious that $\|\hat{s}\|_U \leq \Delta_k$ holds. Moreover, by $\frac{\Delta_k}{\max\{\|s\|_U, \Delta_k\}} \leq 1$ for all $s \in U$ and the convexity of $U_{\mathrm{ad}}$, we have

$$\begin{aligned} u + s &= u + \tfrac{\Delta_k}{\max\{\|P_{U_{\mathrm{ad}} - u}(\hat{s})\|_U, \Delta_k\}} P_{U_{\mathrm{ad}} - u}(\hat{s}) \\ &= \tfrac{\Delta_k}{\max\{\|P_{U_{\mathrm{ad}} - u}(\hat{s})\|_U, \Delta_k\}} \big(\underbrace{u + P_{U_{\mathrm{ad}} - u}(\hat{s})}_{\in U_{\mathrm{ad}}}\big) + \big(1 - \tfrac{\Delta_k}{\max\{\|P_{U_{\mathrm{ad}} - u}(\hat{s})\|_U, \Delta_k\}}\big) u \in U_{\mathrm{ad}}. \end{aligned}$$

□

[8] Often, depending on the concrete choice of $U_{\mathrm{ad}}$, we need $q > 2$ to establish semismoothness.

# 5. Realization of the Required Error Estimates for the Model Problem

We want to apply Algorithm 1 to the example from Section 3.2. To ensure global convergence we have to control the following quantities, see Chapter 4:

- The inexactness of the model gradient:

$$\|\nabla m_k(0) - \nabla \hat{J}(u^k)\|_U \leq \varrho_{\mathrm{g}}(\Delta_k), \tag{5.1}$$

  where $\varrho_{\mathrm{g}} : \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ is a function satisfying $\lim_{t\to 0^+} \varrho_{\mathrm{g}}(t) = 0$, e. g., $\varrho_{\mathrm{g}}(t) = \mathfrak{c}_{\mathrm{g}} t$, $\mathfrak{c}_{\mathrm{g}} > 0$.

- The inexactness of the approximate criticality measure:

$$|\chi_k(0) - \chi(u^k)| \leq \varrho_{\mathrm{c}}(\chi_k(0)), \tag{5.2}$$

  where $\varrho_{\mathrm{c}} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a function satisfying $\lim_{t\to 0^+} \varrho_{\mathrm{c}}(t) = 0$ and $\varrho_{\mathrm{c}}(0) = 0$, e. g., $\varrho_{\mathrm{c}}(t) = \mathfrak{c}_{\mathrm{c}} t$, $\mathfrak{c}_{\mathrm{c}} > 0$.

- The quality of the computed reduction:

$$|\mathrm{ared}_k - \mathrm{cred}_k| \leq \varrho_{\mathrm{r}}(\eta_3 \min\{\mathrm{pred}_k, \mathfrak{r}_k\}), \tag{5.3}$$

  with $\eta_3 < \min\{\eta_1, 1-\eta_2\}$, where $0 < \eta_1 < \eta_2 < 1$ are chosen a priori, and with a forcing sequence $(\mathfrak{r}_k)_{k\in\mathbb{N}_0} \subset \mathbb{R}_{>0}$ fulfilling $\lim_{k\to\infty} \mathfrak{r}_k = 0$ and a function $\varrho_{\mathrm{r}} : \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ fulfilling $\varrho_{\mathrm{r}}(t) \leq t$ for all $t \in (0, \bar{t}]$ with some fixed $\bar{t} > 0$, e. g., $\varrho_{\mathrm{r}}(t) = \mathfrak{c}_{\mathrm{r}} t^{\mathfrak{e}_{\mathrm{r}}}$, $\mathfrak{c}_{\mathrm{r}} > 0$, $\mathfrak{e}_{\mathrm{r}} > 1$.

In this chapter, we discuss how the required error estimates (5.1), (5.2), and (5.3) can be ensured for the example from Section 3.2, but several estimates will hold in a more general setting.

## 5.1. Realization of the Error Estimates in the Deterministic Case

First, we focus on the deterministic case with fixed $\xi \in \Xi$, where the objective function

$$J[\xi](y, u) = \tfrac{1}{2}\|Q(\xi)y - \hat{q}(\xi)\|_H^2 + \tfrac{\gamma}{2}\|u\|_U^2 \tag{5.4}$$

is of tracking type with a real Hilbert space $H$, a desired state $\hat{q}(\xi) \in H$, $Q(\xi) \in \mathcal{L}(Y, H)$, and $\gamma > 0$ as in (3.4) and the state equation is

$$E[\xi](y, u) = A(\xi)y + N(y) - B(\xi)u - b(\xi) = 0 \tag{5.5}$$

with a strongly monotone (with constant $\underline{\kappa}$), linear, bounded operator $A(\xi) : Y \to Y^*$, a monotone and continuously differentiable operator $N : Y \to Y^*$, $B(\xi) \in \mathcal{L}(U, Y^*)$, and $b(\xi) \in Y^*$, cf. (3.10) and Assumption 3.3. In the following, we will skip the dependence on $\xi$ because it is fixed. Still, all derived results are valid for almost every $\xi \in \Xi$. Furthermore, we skip the index $k$ denoting the iteration number in the algorithm for readability purposes in this section as far as possible.

In the following, we describe the procedure which is employed to realize the error control in the deterministic case.

### Model Gradient Error

The model gradient is computed by the (formal) adjoint approach with inexact solutions $\tilde{y}$ and $\tilde{z}$ of the state and (formal) adjoint equation, respectively. Let $y = S(u) \in Y$ be the exact state solving $E(S(u), u) = 0$ and let $\tilde{y} \in Y$ be an inexact solution. The perturbed adjoint equation reads

$$E_y(\tilde{y}, u)^* \hat{z} = -J_y(\tilde{y}, u). \tag{5.6}$$

Its exact solution is denoted by $\hat{z}$ and its inexact solution by $\tilde{z}$, whereas $z$ is the exact adjoint state solving

$$E_y(y, u)^* z = -J_y(y, u). \tag{5.7}$$

**Theorem 5.1.** *Let $J : Y \times U \to \mathbb{R}$ be defined as in* (5.4)*, let $E : Y \times U \to Y^*$ be as in* (5.5)*, and let $u \in U$ and $\tilde{y}, \tilde{z} \in Y$ be given. Moreover, let $y \in Y$ be the exact solution of $E(y, u) = 0$ and let $z \in Y$ and $\hat{z} \in Y$ be the exact solutions of* (5.7) *and* (5.6)*, respectively.*
*Choosing $m$ such that $m'(0) = E_u(\tilde{y}, u)^* \tilde{z} + J_u(\tilde{y}, u)$, it then holds that*

$$\begin{aligned} \|m'(0) - \hat{J}'(u)\|_{U^*} \leq \|B^*(\hat{z} - \tilde{z})\|_{U^*} + \tfrac{1}{\underline{\kappa}} \|B\|_{\mathcal{L}(U,Y^*)} \|Q\|_{\mathcal{L}(Y,H)} \big( \|Q(y - \tilde{y})\|_H \\ + \tfrac{1}{\underline{\kappa}} \|N'(\tilde{y}) - N'(y)\|_{\mathcal{L}(Y,Y^*)} (\|Q\tilde{y} - \hat{q}\|_H + \|Q(y - \tilde{y})\|_H) \big). \end{aligned} \tag{5.8}$$

*Proof.* From the choice of $m'(0)$ and (5.7) we obtain using that $E_u \equiv -B$ and that $J_u$ is independent of $y$:

$$\begin{aligned} \|m'(0) - \hat{J}'(u)\|_{U^*} &= \|E_u(\tilde{y}, u)^* \tilde{z} - E_u(y, u)^* z + J_u(\tilde{y}, u) - J_u(y, u)\|_{U^*} \\ &= \|B^*(z - \tilde{z})\|_{U^*} \leq \|B^*(z - \hat{z})\|_{U^*} + \|B^*(\hat{z} - \tilde{z})\|_{U^*}. \end{aligned} \tag{5.9}$$

Since both equations (5.6) and (5.7) are uniquely solvable because of the strong monotonicity of $A^* + N'(y)^*$ for every $y \in Y$ by Proposition A.4, we can compute

$$\begin{aligned} \hat{z} - z = & - E_y(\tilde{y}, u)^{-*} J_y(\tilde{y}, u) + E_y(y, u)^{-*} J_y(y, u) \\ = & E_y(\tilde{y}, u)^{-*} (J_y(y, u) - J_y(\tilde{y}, u)) + (E_y(y, u)^{-*} - E_y(\tilde{y}, u)^{-*}) J_y(\tilde{y}, u) \\ & + (E_y(y, u)^{-*} - E_y(\tilde{y}, u)^{-*})(J_y(y, u) - J_y(\tilde{y}, u)). \end{aligned} \tag{5.10}$$

With (5.4) and (5.5), the perturbed adjoint equation (5.6) reads

$$E_y(\tilde{y},u)^* z = A^* z + N'(\tilde{y})^* z = -Q^*(Q\tilde{y} - \hat{q}) = -J_y(\tilde{y},u).$$

Compared to the exact adjoint equation (5.7), the error in the right-hand side is

$$J_y(y,u) - J_y(\tilde{y},u) = Q^*Q(y-\tilde{y}). \tag{5.11}$$

For the estimation of the error caused by the approximate left-hand side operator, we introduce for $\tilde{b} \in Y^*$ the unique solutions $v, \tilde{v} \in Y$ of the equations

$$\begin{aligned} A^* v + N'(y)^* v &= \tilde{b}, \\ A^* \tilde{v} + N'(\tilde{y})^* \tilde{v} &= \tilde{b}, \end{aligned}$$

respectively. We have $v - \tilde{v} = (E_y(y,u)^{-*} - E_y(\tilde{y},u)^{-*})\tilde{b}$. Using the monotonicity of $N'(y)$ (Proposition A.4) and the strong monotonicity of $A$ with constant $\underline{\kappa}$, we estimate:

$$\begin{aligned} \underline{\kappa}\|v-\tilde{v}\|_Y^2 &\le \langle v-\tilde{v}, A(v-\tilde{v})\rangle_{Y,Y^*} \le \langle v-\tilde{v}, (A+N'(y))(v-\tilde{v})\rangle_{Y,Y^*} \\ &= \langle A^*v - A^*\tilde{v} + N'(y)^*v - N'(y)^*\tilde{v}, v-\tilde{v}\rangle_{Y^*,Y} \\ &= \langle \tilde{b} - A^*\tilde{v} - N'(\tilde{y})^*\tilde{v} + N'(\tilde{y})^*\tilde{v} - N'(y)^*\tilde{v}, v-\tilde{v}\rangle_{Y^*,Y} \\ &= \langle (N'(\tilde{y}) - N'(y))^*\tilde{v}, v-\tilde{v}\rangle_{Y^*,Y} \le \|(N'(\tilde{y}) - N'(y))^*\tilde{v}\|_{Y^*}\|v-\tilde{v}\|_Y, \end{aligned}$$

cf. the proof of Theorem 3.12. This results in

$$\begin{aligned} \|(E_y(y,u)^{-*} - E_y(\tilde{y},u)^{-*})\tilde{b}\|_Y &\le \tfrac{1}{\underline{\kappa}}\|(N'(\tilde{y}) - N'(y))^*\tilde{v}\|_{Y^*} \\ &\le \tfrac{1}{\underline{\kappa}^2}\|N'(\tilde{y}) - N'(y)\|_{\mathcal{L}(Y,Y^*)}\|\tilde{b}\|_{Y^*}, \end{aligned} \tag{5.12}$$

where the last estimate is due to the strong monotonicity of $A^* + N'(\tilde{y})^*$. Inserting (5.11) into (5.10) and using (5.12), we obtain (again using strong monotonicity):

$$\begin{aligned} \|\hat{z} - z\|_Y \le& \tfrac{1}{\underline{\kappa}}\|Q^*Q(y-\tilde{y})\|_{Y^*} \\ &+ \tfrac{1}{\underline{\kappa}^2}\|N'(\tilde{y}) - N'(y)\|_{\mathcal{L}(Y,Y^*)}\big(\|Q^*(Q\tilde{y} - \hat{q})\|_{Y^*} + \|Q^*Q(y-\tilde{y})\|_{Y^*}\big). \end{aligned} \tag{5.13}$$

Combining this and (5.9) results in (5.8). □

To bound the gradient error, we therefore have to control the error $\|B^*(\hat{z}-\tilde{z})\|_{U^*}$ caused by the inexact solution of the perturbed adjoint equation (5.6). If, e. g., $B \equiv \iota : L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$, it is sufficient to control the adjoint state error in the $L^2(\Omega)$-norm. This is no longer true if a boundary control problem is considered. We will therefore estimate $\|B^*(\hat{z}-\tilde{z})\|_{U^*} \le \|B\|_{\mathcal{L}(U,Y^*)}\|\hat{z}-\tilde{z}\|_Y$ and control the error in the $Y$-norm. Another reason for this is that a posteriori error estimation techniques to estimate the $L^2(\Omega)$-error require the PDE solution to have $H^2(\Omega)$-regularity, see, e. g., [2, Sec. 2.4]. This cannot be guaranteed if the coefficient function $\kappa(\cdot,\xi)$ in the definition (3.10) of the operator $A(\xi)$ is only $L^\infty(\Omega)$-regular, i. e., it can contain jumps along edges for example, or if the domain $\Omega$ is non-convex.

Moreover, we have to control the errors $\|Q(y-\tilde{y})\|_H$ or even $\|Q^*Q(y-\tilde{y})\|_{Y^*}$, see (5.13), and $\|N'(\tilde{y})-N'(y)\|_{\mathcal{L}(Y,Y^*)}$ introduced by the inexact solution of the state equation. Again, if, e. g., $Q \equiv \iota : H_0^1(\Omega) \hookrightarrow L^2(\Omega)$, it is sufficient to control the state error in the $L^2(\Omega)$-norm, which is no longer true if we have a problem with, e. g., boundary observation. Furthermore, we will see that an error estimate possibly in a stronger norm than the $L^2(\Omega)$-norm is required to bound $\|N'(\tilde{y})-N'(y)\|_{\mathcal{L}(Y,Y^*)}$.

The error in the computed state and adjoint state can be controlled by standard a posteriori techniques for elliptic PDEs. If $N'$ is locally Lipschitz continuous w. r. t. $y$, bounding the error $\|N'(\tilde{y})-N'(y)\|_{\mathcal{L}(Y,Y^*)}$ reduces to the estimation of the local Lipschitz constant $c_{N'}$ due to $\|N'(y,u)-N'(\tilde{y},u)\|_{\mathcal{L}(Y,Y^*)} \leq c_{N'}\|\tilde{y}-y\|_Y$. But then, the state error in the possibly stronger $Y$-norm has to be estimated. For the example from Section 3.2, this local Lipschitz constant can be bounded as follows:

**Lemma 5.2.** *Let $N : Y \to Y^*$ be defined as in* (3.10) *(skipping the dependence on $\xi$) and let $\varphi : \mathbb{R} \to \mathbb{R}$ fulfill the respective conditions in Assumption 3.3.*
*Then it holds that*

$$\|N'(\tilde{y})-N'(y)\|_{\mathcal{L}(Y,Y^*)} \leq c_p^3\Big(a''_{\varphi''}\,\lambda(\Omega)^{(p-3)/p} + c''_{\varphi''}c_p^{p-3}\left(\|\tilde{y}\|_Y + \|y-\tilde{y}\|_Y\right)^{p-3}\Big)\|y-\tilde{y}\|_Y, \tag{5.14}$$

*where $\lambda$ is the Lebesgue measure on $\Omega$ and $c_p > 0$ is the Sobolev constant such that $\|y\|_{L^p(\Omega)} \leq c_p\|y\|_{H_0^1(\Omega)}$ holds for every $y \in Y$.*

*Proof.* We have that $\langle N'(y)v,\tilde{v}\rangle_{Y^*,Y} = \int_\Omega \varphi'(y)v\tilde{v}\,\mathrm{d}x$ for $y,v,\tilde{v} \in Y$ and that $\varphi : \mathbb{R} \to \mathbb{R}$ is a twice continuously differentiable, increasing function, which fulfills the growth condition (3.7), i. e., $|\varphi''(t)| \leq a''_{\varphi''} + c''_{\varphi''}|t|^{p-3}$ with $a''_{\varphi''}, c''_{\varphi''} \geq 0$ and $p \in (3,\infty)$ for $n=2$ and $p \in (3,6]$ for $n=3$. Thus, we can estimate with $r_i \in [1,\infty]$ $(i \in \{1,2,3,4,5\})$, $\frac{1}{r_1}+\frac{1}{r_2}+\frac{1}{r_3}=1$, $\frac{1}{r_4}+\frac{1}{r_5}=\frac{1}{r_1}$, and $r_4(p-3) \geq 1$ (to be specified later):

$$\begin{aligned}
&|\langle (N'(y)-N'(\tilde{y}))v,\tilde{v}\rangle_{Y^*,Y}| \\
&\leq \|\varphi'(y)-\varphi'(\tilde{y})\|_{L^{r_1}(\Omega)}\|v\|_{L^{r_2}(\Omega)}\|\tilde{v}\|_{L^{r_3}(\Omega)} \\
&\leq \int_0^1 \|\varphi''(\tilde{y}+\tau(y-\tilde{y}))(y-\tilde{y})\|_{L^{r_1}(\Omega)}\,\mathrm{d}\tau \cdot \|v\|_{L^{r_2}(\Omega)}\|\tilde{v}\|_{L^{r_3}(\Omega)} \\
&\leq \Big(a''_{\varphi''}\cdot\lambda(\Omega)^{1/r_4} + c''_{\varphi''}\cdot\sup_{\tau\in[0,1]}\|\tilde{y}+\tau(y-\tilde{y})\|^{p-3}_{L^{r_4(p-3)}(\Omega)}\Big) \\
&\qquad\cdot\|y-\tilde{y}\|_{L^{r_5}(\Omega)}\|v\|_{L^{r_2}(\Omega)}\|\tilde{v}\|_{L^{r_3}(\Omega)} \\
&\leq c_{r_2}c_{r_3}\Big(a''_{\varphi''}\cdot\lambda(\Omega)^{1/r_4} + c''_{\varphi''}\cdot\big(\max\{\|y\|_{L^{r_4(p-3)}(\Omega)},\|\tilde{y}\|_{L^{r_4(p-3)}(\Omega)}\}\big)^{p-3}\Big) \\
&\qquad\cdot\|y-\tilde{y}\|_{L^{r_5}(\Omega)}\|v\|_{H_0^1(\Omega)}\|\tilde{v}\|_{H_0^1(\Omega)},
\end{aligned} \tag{5.15}$$

where $c_{\hat{r}}$ is the constant from the Sobolev embedding $Y = H_0^1(\Omega) \hookrightarrow L^{\hat{r}}(\Omega)$ with adequately chosen $\hat{r} \in [1,\infty)$ or $\hat{r} \in [1,6]$ dependent on $n$. Choosing $r_2 = r_3 = \tilde{r}$ with $\tilde{r} \in (2,\infty)$ for $n=2$ and $\tilde{r} \in (2,6]$ for $n=3$, $r_5 = \frac{\tilde{r}(p-2)}{\tilde{r}-2} = p-2+\frac{2(p-2)}{\tilde{r}-2} > p-2 > 1$, and $r_4 = \frac{r_5}{p-3} > 1+\frac{1}{p-3}$,

this gives

$$\begin{aligned}&\|N'(y) - N'(\tilde{y})\|_{\mathcal{L}(Y,Y^*)}\\ &\leq c_{\tilde{r}}^2\Big(a''_{\varphi''}\cdot\lambda(\Omega)^{(p-3)/r_5} + c''_{\varphi''}\left(\max\{\|y\|_{L^{r_5}(\Omega)}, \|\tilde{y}\|_{L^{r_5}(\Omega)}\}\right)^{p-3}\Big)\|y-\tilde{y}\|_{L^{r_5}(\Omega)}.\end{aligned} \tag{5.16}$$

The concrete choice $\tilde{r} = p$ in (5.16) (giving $r_5 = p$, $r_4 = \frac{p}{p-3}$) and the Sobolev embedding $Y \hookrightarrow L^p(\Omega)$ yield (5.14). □

**Remark 5.3.** From (5.16) we see that it is enough to control the $L^{r_5}(\Omega)$-error in the computed state. For $n = 2$ we can choose $\tilde{r}$ arbitrarily large, but not $\tilde{r} = \infty$, giving that $r_5$ can be arbitrarily close to $p - 2$.

### Criticality Measure Error

By Proposition 4.11, ensuring

$$\|\nabla m_k(0) - \nabla\hat{J}(u^k)\|_U \leq \min\{\tfrac{1-\mathfrak{c}_\mathrm{s}}{\tau}\varrho_\mathrm{c}(\chi_k(0)), \varrho_\mathrm{g}(\Delta_k)\}$$

as well as

$$\|[P_{U_\mathrm{ad}} - \hat{\mathrm{P}}_{U_\mathrm{ad}}](u^k - \tau\nabla m_k(0))\|_U \leq \mathfrak{c}_\mathrm{s}\varrho_\mathrm{c}(\chi_k(0))$$

for some constant $\mathfrak{c}_\mathrm{s} \in [0, 1]$ yields (4.10) and (4.11).

Thus, if we can control the error in the model gradient as discussed in Theorem 5.1, it remains to control the inexactness in the projection.

**Remark 5.4.** In certain cases, the difference between the exact and the approximate projection can be computed exactly:

- If $U = \mathbb{R}^{n_u}$ is finite dimensional and $U_\mathrm{ad}$ is a convex set for which the projection can be computed simply, e. g., a box or an ellipsoid, there is no need for introducing an approximate projection, i. e., $P_{U_\mathrm{ad}} - \hat{\mathrm{P}}_{U_\mathrm{ad}} \equiv 0$.
- If $U = L^2(\Omega_u)$ with $\Omega_u$ being a measurable subset of $\Omega$ or $\partial\Omega$ and the discretization allows for the exact computation of norms, the projection onto a ball $U_\mathrm{ad} := \{u \in U : \|u\|_U \leq \varrho\}$ with $\varrho > 0$ can also be computed exactly.
- The same holds true for $U = L^2(\Omega_u)$ and $U_\mathrm{ad} := \{u \in U : u_\mathrm{l} \leq u \leq u_\mathrm{u} \text{ a. e.}\}$ with $u_\mathrm{l} < u_\mathrm{u} \in \mathbb{R}$ and a discretization by piecewise constant functions.
- Assume that $U = L^2(\Omega_u)$ with $\Omega_u$ being a subset of the domain $\Omega$ or its boundary $\partial\Omega$, and that continuous, linear finite elements with nodal bases are used for its discretization. Let $U_\mathrm{ad}$ be described by pointwise bound constraints with continuous, piecewise linear functions as bounds, which can be represented exactly in the current discretization. Then, the approximate projection is typically computed by pointwisely projecting the nodal function values onto the box. In [118] a method for computing the error exactly is presented. It is used that the $L^2$-projection on the continuous level can be written in a pointwise fashion. The error is computed on each element of the FE

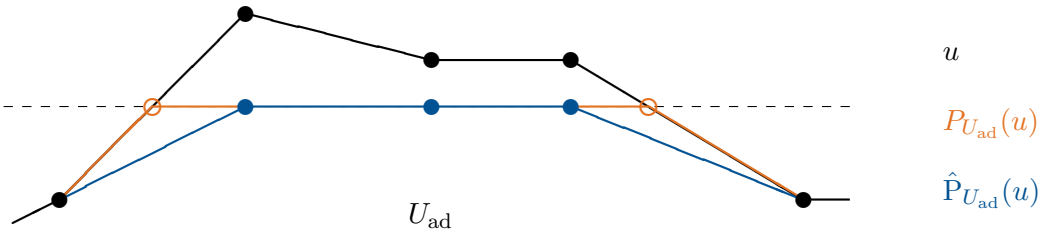



Figure 5.1.: Exact and inexact $L^2$-projection onto a box in 1D.

discretization separately and can be reduced by refining the elements with the largest error contribution. In Figure 5.1, the exact and the inexact projection of a linear FE function $u$ onto a box with constant upper bound in 1D is depicted. It can be recognized that the error in the node-wise projection occurs exactly on two elements, namely the ones where the function $u$ crosses the upper bound. This error can be computed exactly. The elements with the largest error contribution can be refined uniformly if a higher accuracy of the node-wise projection is required.

- If the projection is computed by solving

$$\bar{p} := P_{U_{\text{ad}}}(u) = \arg\min_{p \in U_{\text{ad}}} \tfrac{1}{2}\|u - p\|_U^2$$

approximately, the distance of an $\varepsilon$-solution $\hat{p} \in U_{\text{ad}}$ to $\bar{p} \in U_{\text{ad}}$ can be estimated as $\|\hat{p} - \bar{p}\|_U \leq \sqrt{2\varepsilon}$ due to

$$\begin{aligned} 2\varepsilon &\geq \|u - \hat{p}\|_U^2 - \|u - \bar{p}\|_U^2 = \|u - \bar{p} + \bar{p} - \hat{p}\|_U^2 - \|u - \bar{p}\|_U^2 \\ &= 2\underbrace{(u - \bar{p}, \bar{p} - \hat{p})_U}_{\geq 0} + \|\hat{p} - \bar{p}\|_U^2. \end{aligned}$$

Note that in all cases where the projection can be computed exactly, we can set $c_P = 0$ in Proposition 4.11.

### Objective Function Evaluation Error

We assume that the inexact reduced objective function $\hat{J}_k$ is evaluated using an inexact solution $\tilde{y} \in Y$ of the state equation, i. e., $\hat{J}(u) = J(S(u), u)$ and $\hat{J}_k(u) = J(\tilde{y}, u)$ for some $\tilde{y} \in Y$. By Lemma 4.25, (5.3) holds if

$$|\hat{J}(u) - \hat{J}_k(u)| \leq \frac{1}{2}\varrho_{\text{r}}(\eta_3 \min\{\text{pred}_k, \mathfrak{r}_k\}) \text{ for all } u \in \{u^k, u^k + s^k\} \subset U.$$

Thus, it is enough to control the error $|\hat{J}(u) - \hat{J}_k(u)| = |J(y, u) - J(\tilde{y}, u)|$ for $u \in U$, where $y = S(u)$ is the exact solution of the state equation. If $J$ is locally Lipschitz continuous w. r. t. $y$, it holds that $|J(y, u) - J(\tilde{y}, u)| \leq c_J\|y - \tilde{y}\|_Y$, and for error estimation we have to

estimate the error in the computed state and possibly the local Lipschitz constant $c_J$. For a tracking-type objective function, we have a more explicit estimate:

**Proposition 5.5.** *Let $J : Y \times U \to \mathbb{R}$ be of tracking type form (5.4) and let $y, \tilde{y} \in Y$, $u \in U$ be given. Then the following estimate holds true:*

$$|J(y,u) - J(\tilde{y},u)| \leq \tfrac{1}{2}\|Q(y-\tilde{y})\|_H^2 + \|Q\tilde{y} - \hat{q}\|_H \|Q(y-\tilde{y})\|_H. \tag{5.17}$$

*Proof.* We compute

$$\begin{aligned} J(y,u) - J(\tilde{y},u) &= \tfrac{1}{2}\|Qy - Q\tilde{y} + Q\tilde{y} - \hat{q}\|_H^2 - \tfrac{1}{2}\|Q\tilde{y} - \hat{q}\|_H^2 \\ &= \tfrac{1}{2}\|Q(y-\tilde{y})\|_H^2 + (Q(y-\tilde{y}), Q\tilde{y} - \hat{q})_H. \end{aligned}$$

This shows (5.17). □

Combining Lemma 4.25 and Proposition 5.5, we see that we have to estimate the error $\|Q(y-\tilde{y})\|_H \leq \|Q\|_{\mathcal{L}(Y,H)}\|y-\tilde{y}\|_Y$, which again reduces to the question of error estimation for the computed state. An alternative to the estimate given in Proposition 5.5 would be the dual-weighted-residual method [18], which is well-suited for the estimation of the error in the objective function of an optimal control problem. We do not employ it here having our stochastic application in mind. We want to rely on already established error estimation techniques for PDEs with uncertain inputs, which can be implemented with low-rank tensors.

In conclusion, we only have to control a few errors to ensure (5.1), (5.2), and (5.3):

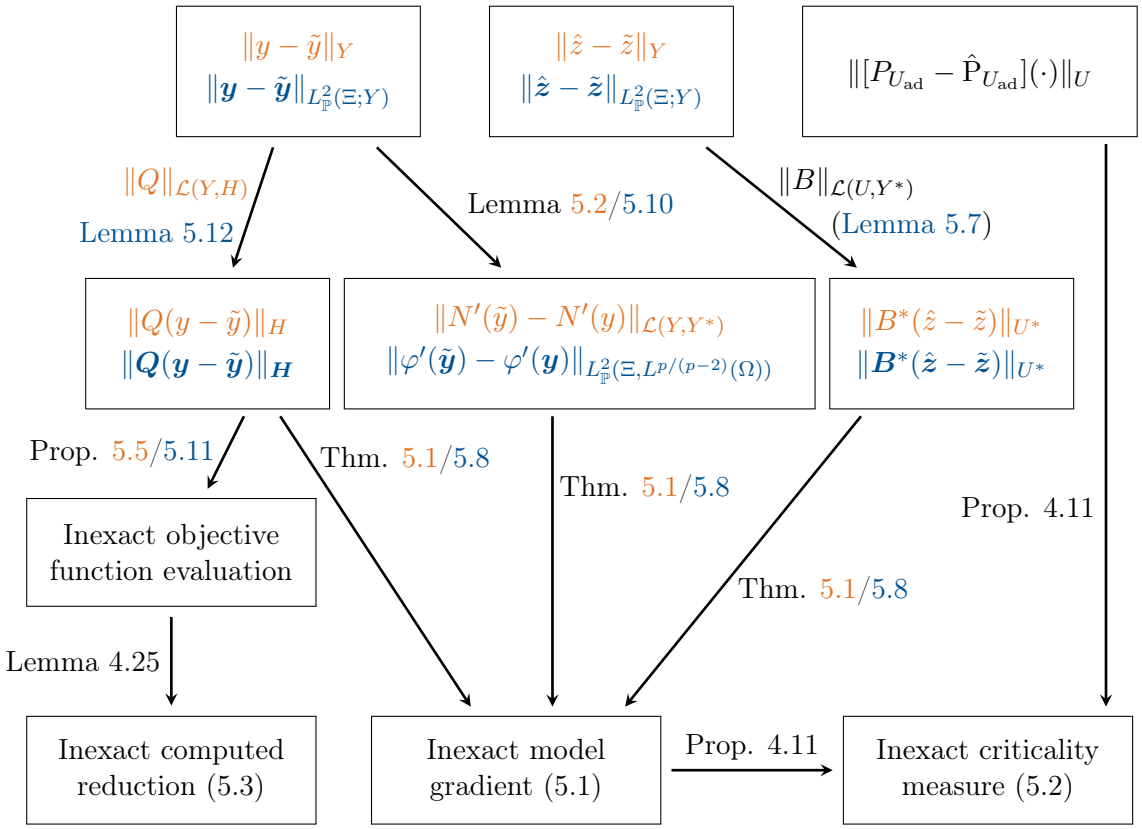
- The error $\|Q(y-\tilde{y})\|_H$ of the computed state observation. This error controls the accuracy of the objective function evaluation and the right-hand side in the perturbed adjoint equation. For the adjoint equation it would even be enough to control $\|Q^*Q(y-\tilde{y})\|_{Y^*}$.

- The error $\|B^*(\hat{z}-\tilde{z})\|_{U^*}$ introduced by the computed adjoint state. This influences the error in the computed gradient (together with the perturbation error of the adjoint equation). The accuracy of the computed gradient is also relevant for the accuracy of the computed criticality measure.

- The error $\|N'(y,u) - N'(\tilde{y},u)\|_{\mathcal{L}(Y,Y^*)}$, which is relevant for the perturbed adjoint equation. A concrete estimator reducing to estimating $\|y-\tilde{y}\|_Y$ is given in Lemma 5.2.

- The error made by the discrete projection in the $U$-norm. In certain cases this error can be computed exactly and the $U$-grid can be refined appropriately.

Overall, everything can be reduced to estimating the errors in the inexact state and adjoint state as well as the error caused by the inexact projection. An overview of the error estimation procedure is given in Figure 5.2, where everything highlighted in orange corresponds to the deterministic case.




Additionally, the following quantities have to be computed or estimated:
$\|Q\tilde{y} - \hat{q}\|_H$ (Theorem 5.1), $\|\tilde{y}\|_Y$ (Lemma 5.2),
$\|\boldsymbol{Q}\tilde{\boldsymbol{y}} - \hat{\boldsymbol{q}}\|_{L^2_{\mathbb{P}}(\Xi,H)}$ (Theorem 5.8), $\|\boldsymbol{y}\|_{L^\infty_{\mathbb{P}}(\Xi;Y)}$ and $\|\tilde{\boldsymbol{y}}\|_{L^\infty_{\mathbb{P}}(\Xi;Y)}$ (Lemma 5.10).

Figure 5.2.: Overview of the error estimation procedure for the deterministic problem and for the stochastic problem in the case $r_f = \infty$. Everything highlighted in orange corresponds to the deterministic case, whereas blue stands for the stochastic case.

## 5.2. Realization of the Error Estimates in the Stochastic Case

We now extend the considerations from Section 5.1 to the stochastic case, for the example from Section 3.2 with

$$\boldsymbol{E} : \boldsymbol{Y} \times U \to \boldsymbol{Y}^*, \ \boldsymbol{E}(\boldsymbol{y}, u) = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{N}(\boldsymbol{y}) - \boldsymbol{B}u - \boldsymbol{b}$$

defined in (3.11) and

$$\boldsymbol{J} : \boldsymbol{Y} \times U \to \mathbb{R}, \ \boldsymbol{J}(\boldsymbol{y}, u) = \int_\Xi J[\xi](\boldsymbol{y}(\xi), u) \, \mathrm{d}\mathbb{P} \tag{5.18}$$

from (3.4) and $J[\xi]$ from (5.4). In Section 3.4 we have already discussed that the adjoint state can be used to compute the gradient $\nabla\hat{\boldsymbol{J}}(u)$ of the reduced objective function, see (3.15).

**Assumption 5.6.** In this section we require the following $\xi$-regularities:

- $f \in L^{r_f}_{\mathbb{P}}(\Xi; L^2(\Omega))$, $\hat{\boldsymbol{q}} \in L^{r_f}_{\mathbb{P}}(\Xi; H)$ for some $r_f \in [p, \infty]$ with $p$ from Assumption 3.3,
- $Q \in L^{\infty}_{\mathbb{P}}(\Xi; \mathcal{L}(Y, H))$, and
- $\boldsymbol{B}$ is constant and defined as in (3.11).

Then we can ensure $\boldsymbol{y}, \boldsymbol{z} \in L^{r_f}_{\mathbb{P}}(\Xi; Y)$ by Corollary 3.15 and Lemma 3.23 and have the a priori bounds (3.9) and (3.20) on the state and the adjoint state. These estimates yield bounds for the respective $L^{r_f}_{\mathbb{P}}(\Xi; Y)$-norm, which is at least as strong as the $\boldsymbol{Y}$-norm by Proposition A.2 because $r_f \geq p$.

### Model Gradient Error

The model gradient is computed as described in Section 3.4, but with inexact solutions of the respective state and adjoint equations.

Quantities of the form $\|\boldsymbol{B}^*(\boldsymbol{z} - \tilde{\boldsymbol{z}})\|_{U^*}$ can be estimated as follows:

**Lemma 5.7.** *Let $\boldsymbol{B} : U \to \boldsymbol{Y}^*$ be defined by*

$$\langle \boldsymbol{B}u, \boldsymbol{v} \rangle_{\boldsymbol{Y}^*, \boldsymbol{Y}} = \int_{\Xi} \langle Bu, \boldsymbol{v}(\cdot, \xi) \rangle_{Y^*, Y} \, \mathrm{d}\mathbb{P}$$

*for some operator $B \in \mathcal{L}(U, Y^*)$ (cf. (3.11)) and let $\boldsymbol{z}, \tilde{\boldsymbol{z}} \in \boldsymbol{Y}$ be given.*
*Then it holds that*

$$\|\boldsymbol{B}^*(\boldsymbol{z} - \tilde{\boldsymbol{z}})\|_{U^*} \leq \|B\|_{\mathcal{L}(U, Y^*)} \cdot \|\tilde{\boldsymbol{z}} - \boldsymbol{z}\|_{L^1_{\mathbb{P}}(\Xi; Y)} \leq \|B\|_{\mathcal{L}(U, Y^*)} \cdot \|\tilde{\boldsymbol{z}} - \boldsymbol{z}\|_{L^2_{\mathbb{P}}(\Xi; Y)}.$$

*Proof.* For $\boldsymbol{v} \in \boldsymbol{Y}$ we have

$$\langle \boldsymbol{B}u, \boldsymbol{v} \rangle_{\boldsymbol{Y}^*, \boldsymbol{Y}} = \int_{\Xi} \langle Bu, \boldsymbol{v}(\cdot, \xi) \rangle_{Y^*, Y} \, \mathrm{d}\mathbb{P} = \langle u, B^* \big( \textstyle\int_{\Xi} \boldsymbol{v}(\cdot, \xi) \, \mathrm{d}\mathbb{P} \big) \rangle_{U, U^*} = \langle u, \boldsymbol{B}^* \boldsymbol{v} \rangle_{U, U^*}.$$

We compute

$$\begin{aligned} \|\boldsymbol{B}^*(\boldsymbol{z} - \tilde{\boldsymbol{z}})\|_{U^*} &= \big\| B^* \big( \textstyle\int_{\Xi} \tilde{\boldsymbol{z}}(\cdot, \xi) - \boldsymbol{z}(\cdot, \xi) \, \mathrm{d}\mathbb{P} \big) \big\|_{U^*} \\ &\leq \|B\|_{\mathcal{L}(U, Y^*)} \cdot \int_{\Xi} \|(\tilde{\boldsymbol{z}}(\cdot, \xi) - \boldsymbol{z}(\cdot, \xi))\|_Y \, \mathrm{d}\mathbb{P} \\ &\leq \|B\|_{\mathcal{L}(U, Y^*)} \cdot \|\tilde{\boldsymbol{z}} - \boldsymbol{z}\|_{L^2_{\mathbb{P}}(\Xi; Y)}, \end{aligned}$$

where the last inequality follows from Proposition A.2. □

**Theorem 5.8.** *Given Assumptions 3.3 and 5.6 and a control $u \in U$, let $\boldsymbol{y} = \boldsymbol{S}(u) \in L^{r_f}_{\mathbb{P}}(\Xi; Y)$ be the exact state (cf. Corollary 3.15) and let $\tilde{\boldsymbol{y}} \in L^{r_f}_{\mathbb{P}}(\Xi; Y)$. Furthermore, let $\boldsymbol{z} = \boldsymbol{T}(u) \in L^{r_f}_{\mathbb{P}}(\Xi; Y)$ be the exact adjoint state (cf. Lemma 3.23), let $\hat{\boldsymbol{z}} \in L^{r_f}_{\mathbb{P}}(\Xi; Y)$ be the unique solution (cf. Proposition 3.24) of the perturbed adjoint equation*

$$\boldsymbol{A}\boldsymbol{z} + \boldsymbol{N}'(\tilde{\boldsymbol{y}})\boldsymbol{z} = -\boldsymbol{Q}^*(\boldsymbol{Q}\tilde{\boldsymbol{y}} - \hat{\boldsymbol{q}})_{\boldsymbol{H}},$$

*cf. (3.18) and (5.6), and let $\tilde{\boldsymbol{z}} \in L^{r_f}_{\mathbb{P}}(\Xi; Y)$.*

*Assume that* $\boldsymbol{J} : \boldsymbol{Y} \times U \to \mathbb{R}$ *is defined as in* (5.18)*, and that* $\hat{\boldsymbol{J}} : U \to \mathbb{R}$ *is defined as in* (3.15)*, and set* $\boldsymbol{m}'(0) := -\boldsymbol{B}^*\tilde{\boldsymbol{z}} + \boldsymbol{J}_u(\tilde{\boldsymbol{y}}, u)$*. Suppose that* $\tilde{r} \in (2, \infty)$ *(for* $n = 2$*) or even* $\tilde{r} \in (2, 6]$ *(for* $n = 3$*) and let* $c_{\tilde{r}}$ *be the Sobolev constant such that* $\|y\|_{L^{\tilde{r}}(\Omega)} \leq c_{\tilde{r}}\|y\|_Y$ *holds for every* $y \in Y$*. Let* $p_1 \in [1, \frac{r_f}{p-2}]$*,* $p_2 \in [1, r_f]$ *such that* $\frac{1}{p_1} + \frac{1}{p_2} = 1$*.*
*Then the following estimate holds true:*

$$\begin{aligned}
&\|\boldsymbol{m}'(0) - \hat{\boldsymbol{J}}'(u)\|_{U^*} \leq \qquad (5.19)\\
&\|\boldsymbol{B}^*(\tilde{\boldsymbol{z}} - \hat{\boldsymbol{z}})\|_{U^*} + \tfrac{1}{\underline{\kappa}}\|B\|_{\mathcal{L}(U,Y^*)}\|Q\|_{L^\infty_{\mathbb{P}}(\Xi;\mathcal{L}(Y,H))}\Big(\|\boldsymbol{Q}(\boldsymbol{y} - \tilde{\boldsymbol{y}})\|_{L^1_{\mathbb{P}}(\Xi,H)}\\
&\quad + \tfrac{1}{\underline{\kappa}}c_{\tilde{r}}^2\|\varphi'(\tilde{\boldsymbol{y}}) - \varphi'(\boldsymbol{y})\|_{L^{p_1}_{\mathbb{P}}(\Xi,L^{\tilde{r}/(\tilde{r}-2)}(\Omega))}\big(\|\boldsymbol{Q}\tilde{\boldsymbol{y}} - \hat{\boldsymbol{q}}\|_{L^{p_2}_{\mathbb{P}}(\Xi,H)} + \|\boldsymbol{Q}(\boldsymbol{y} - \tilde{\boldsymbol{y}})\|_{L^{p_2}_{\mathbb{P}}(\Xi,H)}\big)\Big).
\end{aligned}$$

*Proof.* Analogously to (5.9), it holds that

$$\|\boldsymbol{m}'(0) - \hat{\boldsymbol{J}}'(u)\|_{U^*} \leq \|\boldsymbol{B}^*(\hat{\boldsymbol{z}} - \tilde{\boldsymbol{z}})\|_{U^*} + \|\boldsymbol{B}^*(\boldsymbol{z} - \hat{\boldsymbol{z}})\|_{U^*}.$$

The second summand is estimated by Lemma 5.7: $\|\boldsymbol{B}^*(\boldsymbol{z}-\hat{\boldsymbol{z}})\|_{U^*} \leq \|B\|_{\mathcal{L}(U,Y^*)}\|\boldsymbol{z}-\hat{\boldsymbol{z}}\|_{L^1_{\mathbb{P}}(\Xi;Y)}$. Since $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$ are defined pointwise (almost everywhere), the estimate (5.13) yields

$$\begin{aligned}
\|\boldsymbol{z} - \hat{\boldsymbol{z}}\|_{L^1_{\mathbb{P}}(\Xi;Y)} \leq &\tfrac{1}{\underline{\kappa}}\|\boldsymbol{Q}^*\boldsymbol{Q}(\boldsymbol{y} - \tilde{\boldsymbol{y}})\|_{L^1_{\mathbb{P}}(\Xi;Y^*)} + \tfrac{1}{\underline{\kappa}^2}\|\boldsymbol{N}'(\tilde{\boldsymbol{y}}) - \boldsymbol{N}'(\boldsymbol{y})\|_{L^{p_1}_{\mathbb{P}}(\Xi;\mathcal{L}(Y,Y^*))}\\
&\cdot\Big(\|\boldsymbol{Q}^*(\boldsymbol{Q}\tilde{\boldsymbol{y}} - \hat{\boldsymbol{q}})\|_{L^{p_2}_{\mathbb{P}}(\Xi;Y^*)} + \|\boldsymbol{Q}^*\boldsymbol{Q}(\boldsymbol{y} - \tilde{\boldsymbol{y}})\|_{L^{p_2}_{\mathbb{P}}(\Xi;Y^*)}\Big).
\end{aligned}$$

with $p_1, p_2 \in [1, \infty]$, $\frac{1}{p_1} + \frac{1}{p_2} = 1$. Using $\|N'(y) - N'(y)\|_{\mathcal{L}(Y,Y^*)} \leq c_{\tilde{r}}^2\|\varphi'(y) - \varphi'(\tilde{y})\|_{L^{\tilde{r}/(\tilde{r}-2)}(\Omega)}$ (cf. (5.15)) results in (5.19). The admissible values of $p_1$ and $p_2$ ensure together with the regularity of $\boldsymbol{y}, \tilde{\boldsymbol{y}}$ and the growth of $\varphi'$ that every appearing quantity in (5.19) is finite. □

**Remark 5.9.** The parameters $p_1$ and $p_2$ in Theorem 5.8 make different estimates involving $\|\boldsymbol{Q}(\boldsymbol{y} - \tilde{\boldsymbol{y}})\|_{L^{p_2}_{\mathbb{P}}(\Xi;H)}$ with $2 \leq p_2 \leq r_f$ possible. For larger $p_2$, a weaker norm w. r. t. $\xi$ can be used to estimate the error in $\boldsymbol{N}'(\tilde{\boldsymbol{y}})$.

We see that in general it is sufficient to control the adjoint state error in the $L^2_{\mathbb{P}}(\Xi; Y)$-norm or even the $L^1_{\mathbb{P}}(\Xi; Y)$-norm. Error control in a weaker norm can be sufficient depending on the example: For the concrete definition (3.11) of $\boldsymbol{B}$ it holds that

$$\|\boldsymbol{B}^*(\boldsymbol{z} - \tilde{\boldsymbol{z}})\|_{U^*} = \|D^*\big(\textstyle\int_\Xi \tilde{\boldsymbol{z}}(\cdot, \xi) - \boldsymbol{z}(\cdot, \xi)\,\mathrm{d}\mathbb{P}\big)\|_{L^2(\Omega_u)},$$

identifying $L^2(\Omega_u)^* = L^2(\Omega_u)$. If, e. g., $D \equiv I : L^2(\Omega) \to L^2(\Omega)$, it would be sufficient to control the $L^1_{\mathbb{P}}(\Xi; L^2(\Omega))$-error. This is no longer true if we consider, e. g., a boundary control problem. Thus, we will control the $L^2_{\mathbb{P}}(\Xi; Y)$-error in the adjoint state to keep our algorithm flexible. Furthermore, this enables us to use the fact that the operator $\boldsymbol{A}$ is strongly monotone with constant $\underline{\kappa}$ on $L^2_{\mathbb{P}}(\Xi; Y)$ (but not strongly monotone on $L^p_{\mathbb{P}}(\Xi; Y)$ for $p > 2$).

We will see that, in certain cases, it can be sufficient to control the $L^2_{\mathbb{P}}(\Xi; Y)$-error in the computed state in order to bound the error caused by $\tilde{\boldsymbol{y}}$ entering the perturbed adjoint equation. Again, we need an estimate of the form

$$\|\varphi'(\tilde{\boldsymbol{y}}) - \varphi'(\boldsymbol{y})\|_{L^{p_1}_{\mathbb{P}}(\Xi;L^{\tilde{r}/(\tilde{r}-2)}(\Omega))} \leq C_{\varphi'}\|\tilde{\boldsymbol{y}} - \boldsymbol{y}\|_{L^{p_3}_{\mathbb{P}}(\Xi;Y)}$$

with a local Lipschitz constant $C_{\varphi'}$.

**Lemma 5.10.** *Let $\varphi : \mathbb{R} \to \mathbb{R}$ fulfill the respective conditions in Assumption 3.3, let $\boldsymbol{N} : \boldsymbol{Y} \to \boldsymbol{Y}^*$ be defined as in (3.11), and let $\boldsymbol{y}, \tilde{\boldsymbol{y}} \in \boldsymbol{Y}$ be given. Let $p_1$ be as in Theorem 5.8 and let $p_3, p_4 \in [1, \infty]$ such that $\frac{1}{p_3} + \frac{1}{p_4} = \frac{1}{p_1}$.*
*Then it holds that*

$$\begin{aligned} &\|\varphi'(\tilde{\boldsymbol{y}}) - \varphi'(\boldsymbol{y})\|_{L^{p_1}_{\mathbb{P}}(\Xi;L^{p/(p-2)}(\Omega))} \leq \\ &c_p \|\tilde{\boldsymbol{y}} - \boldsymbol{y}\|_{L^{p_3}_{\mathbb{P}}(\Xi;Y)} \big(a''_{\varphi''}\, \lambda(\Omega)^{(p-3)/p} + c''_{\varphi''} c_p^{p-3} \max\big\{\|\boldsymbol{y}\|_{L^{p_4(p-3)}_{\mathbb{P}}(\Xi;Y)}, \|\tilde{\boldsymbol{y}}\|_{L^{p_4(p-3)}_{\mathbb{P}}(\Xi;Y)}\big\}^{p-3}\big). \end{aligned} \tag{5.20}$$

*Proof.* As in the proof of Lemma 5.2 we estimate

$$\begin{aligned} &\|\varphi'(\tilde{\boldsymbol{y}}) - \varphi'(\boldsymbol{y})\|_{L^{p_1}_{\mathbb{P}}(\Xi;L^{\tilde{r}/(\tilde{r}-2)}(\Omega))} \\ &= \Big\| \int_0^1 \varphi''(\boldsymbol{y} + \tau(\tilde{\boldsymbol{y}} - \boldsymbol{y}))(\tilde{\boldsymbol{y}} - \boldsymbol{y})\, \mathrm{d}\tau \Big\|_{L^{p_1}_{\mathbb{P}}(\Xi;L^{\tilde{r}/(\tilde{r}-2)}(\Omega))} \\ &\leq \|\tilde{\boldsymbol{y}} - \boldsymbol{y}\|_{L^{p_3}_{\mathbb{P}}(\Xi;L^{\tilde{r}}(\Omega))} \cdot \sup_{\tau\in[0,1]} \|\varphi''(\boldsymbol{y} + \tau(\tilde{\boldsymbol{y}} - \boldsymbol{y}))\|_{L^{p_4}_{\mathbb{P}}(\Xi;L^{\tilde{r}/(\tilde{r}-3)}(\Omega))} \end{aligned}$$

for some $\tilde{r} \in (3, \infty)$, using $\frac{1}{p_3} + \frac{1}{p_4} = \frac{1}{p_1}$. The second factor can be estimated as

$$\begin{aligned} &\sup_{\tau\in[0,1]} \|\varphi''(\boldsymbol{y} + \tau(\tilde{\boldsymbol{y}} - \boldsymbol{y}))\|_{L^{p_4}_{\mathbb{P}}(\Xi;L^{\tilde{r}/(\tilde{r}-3)}(\Omega))} \\ \leq\ &\sup_{\tau\in[0,1]} \|a''_{\varphi''} + c''_{\varphi''} |\boldsymbol{y} + \tau(\tilde{\boldsymbol{y}} - \boldsymbol{y})|^{p-3}\|_{L^{p_4}_{\mathbb{P}}(\Xi;L^{\tilde{r}/(\tilde{r}-3)}(\Omega))} \\ =\ &a''_{\varphi''} \cdot \lambda(\Omega)^{(\tilde{r}-3)/\tilde{r}} + c''_{\varphi''} \cdot \max\big\{\|\boldsymbol{y}\|_{L^{p_4(p-3)}_{\mathbb{P}}(\Xi;L^{r_5}(\Omega))}, \|\tilde{\boldsymbol{y}}\|_{L^{p_4(p-3)}_{\mathbb{P}}(\Xi;L^{r_5}(\Omega))}\big\}^{p-3} \end{aligned}$$

with $r_5 = \frac{\tilde{r}(p-3)}{\tilde{r}-3}$ by $|\varphi''(t)| \leq a''_{\varphi''} + c''_{\varphi''}|t|^{p-3}$ with $c''_{\varphi''} \geq 0$ and $p > 3$. If we choose $\tilde{r} = p$ and use $H^1_0(\Omega) \subset L^p(\Omega)$, we obtain (5.20). □

Combining Theorem 5.8 and Lemma 5.10 and choosing $p_2 = p_3$, we see that we have to estimate the error $\|\tilde{\boldsymbol{y}} - \boldsymbol{y}\|_{L^{p_2}_{\mathbb{P}}(\Xi;Y)}$, compute the norm $\|\tilde{\boldsymbol{y}}\|_{L^{p_4(p-3)}_{\mathbb{P}}(\Xi;Y)}$ or bound it from above, and bound the norm $\|\boldsymbol{y}\|_{L^{p_4(p-3)}_{\mathbb{P}}(\Xi;Y)}$, e. g., by the a priori estimate (3.9).

For $r_f = \infty$ we can choose $p_2 = p_3 = 2$ and $p_4 = \infty$. Then it is enough to estimate the error in the inexact state in the $L^2_{\mathbb{P}}(\Xi;Y)$-norm as long as we can compute or bound $\|\tilde{\boldsymbol{y}}\|_{L^\infty_{\mathbb{P}}(\Xi;Y)}$ and $\|\boldsymbol{y}\|_{L^\infty_{\mathbb{P}}(\Xi;Y)}$.

For $r_f < \infty$, we have to choose $p_4 \in [\frac{r_f}{r_f-2}, \frac{r_f}{p-3}]$ and $p_2 = p_3 = \frac{2p_4}{p_4-1} \in [\frac{2r_f}{r_f-p+3}, r_f]$.[9] Observe that $p_1 = \frac{2p_4}{p_4+1} \leq \frac{2r_f}{r_f+p-3} \leq \frac{r_f}{p-\frac{3}{2}}$ follows then. On the one hand, for $r_f = p$ we can take $p_4 = \frac{p}{p-2}$. Then we have $p_2 = p_3 = p$ and the error in the computed state has to be estimated in the $\boldsymbol{Y}$-norm. On the other hand, for increasing $r_f$ and using $p_4 = \frac{r_f}{p-3}$, the exponents $p_2$ and $p_3$ get close to 2. Then it is enough to estimate the error in the computed state in a weaker norm than the $\boldsymbol{Y}$-norm, but bounds on the exact and the inexact state have to be computed in a stronger norm.

[9] Note that $1 < \frac{r_f}{r_f-2} \leq \frac{r_f}{r_f-3} \leq \frac{r_f}{p-3}$ and $2 < \frac{2r_f}{r_f-p+3} \leq \frac{2}{3} r_f$ hold for $p \in (3, \infty)$, $r_f \in [p, \infty)$.

### Criticality Measure Error

By Proposition 4.11, the control of the error in the approximate criticality measure reduces to controlling the inexactness of the approximate projection and the model gradient.

### Objective Function Evaluation Error

**Proposition 5.11.** *Let $\boldsymbol{J} : \boldsymbol{Y} \times U \to \mathbb{R}$ be defined as in (3.4) and let $\boldsymbol{y}, \tilde{\boldsymbol{y}} \in Y$, $u \in U$ be given. Then it holds that*

$$|\boldsymbol{J}(\boldsymbol{y}, u) - \boldsymbol{J}(\tilde{\boldsymbol{y}}, u)| \leq \tfrac{1}{2}\|\boldsymbol{Q}(\boldsymbol{y} - \tilde{\boldsymbol{y}})\|^2_{L^2_{\mathbb{P}}(\Xi;H)} + \|\boldsymbol{Q}\tilde{\boldsymbol{y}} - \hat{\boldsymbol{q}}\|_{L^2_{\mathbb{P}}(\Xi;H)}\|\boldsymbol{Q}(\boldsymbol{y} - \tilde{\boldsymbol{y}})\|_{L^2_{\mathbb{P}}(\Xi;H)}.$$

*Proof.* The estimate (5.17) (Proposition 5.5) and Hölder's inequality yield

$$\begin{aligned}
&|\boldsymbol{J}(\boldsymbol{y}, u) - \boldsymbol{J}(\tilde{\boldsymbol{y}}, u)| = \Big| \int_\Xi J[\xi](y(\xi), u) - J[\xi](\tilde{y}(\xi), u) \,\mathrm{d}\mathbb{P} \Big| \\
\leq &\int_\Xi \tfrac{1}{2}\|Q(\xi)(y(\xi) - \tilde{y}(\xi))\|^2_H + \|Q(\xi)\tilde{y}(\xi) - \hat{q}(\xi)\|_H \|Q(\xi)(y(\xi) - \tilde{y}(\xi))\|_H \,\mathrm{d}\mathbb{P} \\
\leq &\tfrac{1}{2}\|\boldsymbol{Q}(\boldsymbol{y} - \tilde{\boldsymbol{y}})\|^2_{L^2_{\mathbb{P}}(\Xi;H)} + \|\boldsymbol{Q}\tilde{\boldsymbol{y}} - \hat{\boldsymbol{q}}\|_{L^2_{\mathbb{P}}(\Xi;H)}\|\boldsymbol{Q}(\boldsymbol{y} - \tilde{\boldsymbol{y}})\|_{L^2_{\mathbb{P}}(\Xi;H)}.
\end{aligned}$$

□

We see that we have to compute or bound the $L^2_{\mathbb{P}}(\Xi; H)$-norm of $\boldsymbol{Q}\tilde{\boldsymbol{y}} - \hat{\boldsymbol{q}}$ and have to estimate $\|\boldsymbol{Q}(\boldsymbol{y} - \tilde{\boldsymbol{y}})\|_{L^2_{\mathbb{P}}(\Xi;H)}$. If, e. g., $Q(\xi) \equiv \iota : H^1_0(\Omega) \hookrightarrow L^2(\Omega) = H$, it is enough to estimate the $L^2_{\mathbb{P}}(\Xi; L^2(\Omega))$-norm of $\boldsymbol{y} - \tilde{\boldsymbol{y}}$. This is not true anymore if we consider a problem with boundary observation. Again, to have a flexible algorithm and to use strong monotonicity of $\boldsymbol{A}$, we estimate the $L^2_{\mathbb{P}}(\Xi; Y)$-error and use:

**Lemma 5.12.** *Under Assumption 5.6,*

$$\|\boldsymbol{Q}(\boldsymbol{y} - \tilde{\boldsymbol{y}})\|_{L^2_{\mathbb{P}}(\Xi;H)} \leq \|Q\|_{L^\infty_{\mathbb{P}}(\Xi;\mathcal{L}(Y,H))}\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|_{L^2_{\mathbb{P}}(\Xi;Y)}$$

*holds for all $\boldsymbol{y}, \tilde{\boldsymbol{y}} \in \boldsymbol{Y}$.*

*Proof.* This is a simple consequence of Hölder's inequality, cf. Proposition 3.1. It is important that $Q$ has a higher regularity w. r. t. $\xi$ than required for the operator $\boldsymbol{Q} : \boldsymbol{Y} \to \boldsymbol{H}$ to be well-defined. □

How the error in the computed state and adjoint state can be measured up to fixed, but possibly unknown constant factors, is discussed in the Chapter 7 for the example of a semilinear, elliptic PDE with stochastic coefficients from Section 3.2.

# 6. Discretization of the Model Problem

After discussing the functional analytic setting of the problem, a solution algorithm formulated in a Hilbert space, and the realization of the error estimates based on the error in the approximate solution of the state and the adjoint equation, it remains to discretize and adaptively solve the problem and the corresponding equations. The discretization is carried out almost exactly as in our paper [46], i. e., we use conforming finite element discretizations for the deterministic state and control spaces and polynomials for the spaces of random variables. Then the full tensor product of the respective finite-dimensional subspaces is built and used for a conforming stochastic Galerkin discretization. We use weighted Lagrange polynomials w. r. t. the Gaussian quadrature nodes as bases instead of the typical orthonormal polynomials with increasing degree. This has some useful consequences:

- These Lagrange polynomials are orthogonal and can be weighted such that they are orthonormal.
- A connection to stochastic collocation methods can be established, see [44, 46].
- Pointwise state constraints or certain nonsmooth risk measures can be handled by posing the constraint in every quadrature node in the discrete setting.
- Nonlinear dependence on the parameters and nonlinear operators can be approximated nicely. This was already done in [46, Example 3.6] and is used in this thesis for the discretization of the nonlinearity.

In [46], the discretization was fixed, but will now be adaptive. Hence, we will use sequences of nested discrete spaces with their respective bases and linear maps prolongating the coefficients in a coarser space to those in a finer space. For completeness, we review the discretization procedure described in [46] here, add some details about the discretization of all appearing nonlinearities, and point out necessary changes. The concrete adaptive approach is described in Chapter 7. Since constructing a finer space and prolongating the coefficients to it is a simple task for the considered FE and polynomial spaces, we describe only a fixed discretization in this chapter, having in mind that the mesh size and the degrees of the polynomials are parameters which will be adapted in the final implementation.

## 6.1. Space Discretization: Finite Elements

**Assumption 6.1.** For the discretization of the deterministic function spaces we assume that

- the open, bounded Lipschitz domain $\Omega \subset \mathbb{R}^2$ is polygonal and that
- the restriction $D|_{\mathrm{U}} : \mathrm{U} \to L^2(\Omega)$ of the given operator $D$ to any used discrete subspace $\mathrm{U} \subset U$ can be evaluated exactly.

**Remark 6.2.** We exclude the case $\Omega \subset \mathbb{R}^3$ here, which leads to a more compact presentation of a posteriori error estimates in Chapter 7. It is possible to generalize many results to the 3D case. The polygonal domain $\Omega \subset \mathbb{R}^2$ can be covered by a finite element triangulation so that we do not have to care about variational crimes caused by domain approximation.

The domain $\Omega$ is partitioned into a finite element mesh yielding a triangulation $\mathcal{T}$. Let $\mathrm{Y} \subset Y$ denote the discrete subspace of piecewise linear, globally continuous finite element functions with zero boundary data. Generalizations of this would be possible, but are skipped here for the ease of presentation and implementation. The nodal FE basis of the discrete, deterministic state space is $\{\phi_{k_0}\}_{k_0=1}^{d_0} \subset \mathrm{Y} \subset Y = H_0^1(\Omega)$ and the basis of discrete control space $\mathrm{U} \subset U = L^2(\Omega_u)$ is denoted by $\{\psi_\ell\}_{\ell=1}^{d_u}$. This can also be a nodal FE basis if $\Omega_u$ is a subset of $\Omega$ with positive measure or—for finite-dimensional controls—the standard basis of $\mathbb{R}^{d_u} = L^2([d_u])$. In the latter case, no discretization is required. The basis functions shall sum to one, i. e., $\sum_{k_0=1}^{d_0} \phi_{k_0}(x) = 1$ and $\sum_{k_u=1}^{d_u} \psi_{k_u}(\tilde{x}) = 1$ for all $x \in \Omega$, $\tilde{x} \in \Omega_u$ to perform mass lumping in a meaningful way later.

In addition to assuming that $D|_{\mathrm{U}} : L^2(\Omega_u) \to L^2(\Omega)$ can be evaluated exactly (Assumption 6.1), we match the discretizations of the state and the control space, i. e., we use the same grid on $\Omega$ and $\Omega_u$ if $\Omega_u$ is a subset of $\Omega$ with positive measure. This makes the computation of the gradient of the reduced objective function easier since typically $\hat{J}'(u) = B^*z + \gamma(u, \cdot)_U$. Then it is desirable that $B^*z$ and $u$ share the same grid. This is also important to be able to perform gradient-based updates of the control in an optimization method.

We define the following matrices:

- the stiffness matrix $\mathsf{K} \in \mathbb{R}^{d_0 \times d_0}$ for Y: $\mathsf{K}_{k_0 l_0} := (\nabla\phi_{k_0}, \nabla\phi_{l_0})_{L^2(\Omega)^n}$,
- the mass matrix $\mathsf{M} \in \mathbb{R}^{d_0 \times d_0}$ for Y: $\mathsf{M}_{k_0 l_0} := (\phi_{k_0}, \phi_{l_0})_{L^2(\Omega)}$,
- the lumped mass matrix $\mathsf{M_L} \in \mathbb{R}^{d_0 \times d_0}$ for Y: $(\mathsf{M_L})_{k_0 k_0} := \sum_{l_0=1}^{d_0} \mathsf{M}_{k_0 l_0} = \int_\Omega \phi_{k_0} \,\mathrm{d}x$ and $(\mathsf{M_L})_{k_0 l_0} = 0$ for $k_0 \neq l_0$,
- the mass matrix $\tilde{\mathsf{M}} \in \mathbb{R}^{d_u \times d_u}$ for U: $\tilde{\mathsf{M}}_{k_u l_u} := (\psi_{k_u}, \psi_{l_u})_{L^2(\Omega_u)}$,
- the lumped mass matrix $\tilde{\mathsf{M}}_\mathsf{L} \in \mathbb{R}^{d_u \times d_u}$ for U: $(\tilde{\mathsf{M}}_\mathsf{L})_{k_u k_u} := \sum_{l_u=1}^{d_u} \tilde{\mathsf{M}}_{k_u l_u} = \int_{\Omega_u} \psi_{k_u} \,\mathrm{d}\tilde{x}$ and $(\tilde{\mathsf{M}}_\mathsf{L})_{k_u l_u} = 0$ for $k_u \neq l_u$.

Let $\mathsf{y} \in \mathbb{R}^{d_0}$ and $\mathsf{u} \in \mathbb{R}^{d_u}$ be the coefficients, representing the discrete state y and control u, respectively. Inserting $\mathrm{y}(x) = \sum_{k_0=1}^{d_0} \mathsf{y}_{k_0} \phi_{k_0}(x)$ and $\mathrm{u}(x) = \sum_{k_u=1}^{d_u} \mathsf{u}_{k_u} \psi_{k_u}(x)$ into (3.8), and testing with $v \equiv \phi_{k_0}$ for $k_0 \in [d_0]$, the discrete version of the deterministic state equation reads

$$\mathsf{A}(\xi)\mathsf{y} + \mathsf{N}(\mathsf{y}) = \mathsf{B}\mathsf{u} + \mathsf{b}(\xi) \tag{6.1}$$

with

$$\begin{aligned} \mathsf{A}(\xi) &\in \mathbb{R}^{d_0 \times d_0}, & \mathsf{A}_{k_0 l_0}(\xi) &= (\kappa(\cdot, \xi)\nabla\phi_{l_0}, \nabla\phi_{k_0})_{L^2(\Omega)^n}, \\ \mathsf{N} : \mathbb{R}^{d_0} &\to \mathbb{R}^{d_0}, & \mathsf{N}(\mathsf{y}) &= \mathsf{M_L}\varphi(\mathsf{y}), \\ \mathsf{B} &\in \mathbb{R}^{d_0 \times d_u}, & \mathsf{B}_{k_0 l_u} &= (D\psi_{l_u}, \phi_{k_0})_{L^2(\Omega)}, \\ \mathsf{b}(\xi) &\in \mathbb{R}^{d_0}, & \mathsf{b}_{k_0}(\xi) &= (f(\xi), \phi_{k_0})_{L^2(\Omega)}. \end{aligned} \tag{6.2}$$

We will refer to $\mathsf{A}(\xi)$ as the *system matrix* in contrast to the *stiffness matrix*, which is induced by the $H_0^1(\Omega)$-norm. Due to the ease of implementation and interpolation, a quadrature error is allowed to occur in the discretization of the nonlinearity which is connected to mass lumping: The integral $\int_\Omega \varphi(\mathrm{y})\phi_{k_0}\,\mathrm{d}x$ is evaluated inexactly by a quadrature formula, the nodes of which are the finite element grid nodes and the weights of which are the respective entries of the lumped mass matrix. This quadrature formula is exact for integrals over a single linear finite element function. We obtain

$$\int_\Omega \varphi(\mathrm{y})\phi_{k_0}\,\mathrm{d}x \approx \varphi(\mathsf{y}_{k_0})(\mathsf{M}_\mathsf{L})_{k_0k_0}. \tag{6.3}$$

Since the trust-region algorithm (Algorithm 1) is formulated in the infinite-dimensional space $U$, it is desirable to not make additional errors by mass lumping in the objective function, but to evaluate the $U$- and $H$-inner product exactly. Let $H$ be discretized such that $Q(\xi)|_Y$ can be evaluated exactly and such that $\hat{q}(\xi)$ can be represented exactly. The discrete subspace H of $H$ is isomorphic to $\mathbb{R}^{d_H}$ ($d_H \in \mathbb{N}$) equipped with the inner product induced by the symmetric, positive definite matrix $\mathsf{M}_H \in \mathbb{R}^{d_H\times d_H}$. Let $\mathsf{Q}(\xi) \in \mathbb{R}^{d_H\times d_0}$ and $\hat{\mathsf{q}}(\xi) \in \mathbb{R}^{d_H}$ be the discrete versions of $Q(\xi)$ and $\hat{q}(\xi)$, respectively. Then, the discretized objective function from (3.4) reads

$$\mathsf{J}[\xi](\mathsf{y},\mathsf{u}) = \tfrac{1}{2}\|\mathsf{Q}(\xi)\mathsf{y} - \hat{\mathsf{q}}(\xi)\|^2_{\mathsf{M}_H} + \tfrac{\gamma}{2}\mathsf{u}^\top\tilde{\mathsf{M}}\mathsf{u}. \tag{6.4}$$

We note that under the stated assumptions, the evaluation of the objective function is exact so that the error in the reduced objective function depends only on the error in the discretized state and Proposition 5.5 can be applied. The discrete version of the deterministic adjoint equation (3.16) is

$$\mathsf{A}(\xi)\mathsf{z}(\xi) + \mathsf{M}_\mathsf{L}(\varphi'(\mathsf{y}(\xi)) \odot \mathsf{z}(\xi)) = -\mathsf{Q}(\xi)^\top\mathsf{M}_H(\mathsf{Q}(\xi)\mathsf{y}(\xi) - \hat{\mathsf{q}}(\xi)), \tag{6.5}$$

where again the quadrature formula using the finite element nodes has been applied. In fact, $\mathsf{N}'(\mathsf{y})\mathsf{z} = \mathsf{M}_\mathsf{L}(\varphi'(\mathsf{y}) \odot \mathsf{z})$ holds also in the discrete setting and we can identify $\mathsf{N}'(\mathsf{y}) = \mathsf{M}_\mathsf{L}\,\mathrm{diag}(\varphi'(\mathsf{y}))$. The gradient of the reduced, deterministic, discretized objective function is then given by

$$\nabla\hat{\mathsf{J}}[\xi](\mathsf{u}) = -\tilde{\mathsf{M}}^{-1}\mathsf{B}^\top\mathsf{z}(\xi) + \gamma\mathsf{u}, \tag{6.6}$$

cf. (3.21). Note that in typical situations it is not necessary to invert the mass matrix $\tilde{\mathsf{M}}$ to compute the reduced gradient: If, e. g., $\Omega_u \subset \Omega$ is a subset of positive measure, $B^* : H_0^1(\Omega) \to L^2(\Omega_u), z \mapsto z|_{\Omega_u}$ is the canonical embedding $\iota : H_0^1(\Omega) \hookrightarrow L^2(\Omega)$ combined with restriction of the function to $\Omega_u$, and the grids on $\Omega$ and $\Omega_u$ match, the application of $\tilde{\mathsf{M}}^{-1}\mathsf{B}$ consists of a simple extraction of components of the vector $\mathsf{z}(\xi)$ and/or adding zero components for the nodes on $\partial\Omega$. Such situations are favorable because the error in the discrete gradient then only depends on the error in the discrete adjoint state and it is sufficient to apply the error estimate from Theorem 5.1.

Once the equations (3.25) and (3.24) are discretized, the application of the Hessian operator to a direction can be computed via (3.23). For this purpose, it only remains to discretize the term in (3.24) involving the second derivative $N_{yy}$, which is again done by using the

FE nodes based quadrature. Then, if $\mathsf{s} \in \mathbb{R}^{d_u}$ represents a direction $s \in U$, and $\mathsf{y}(\xi)$, $\mathsf{z}(\xi)$ represent the current state and adjoint state, respectively, the application of the Hessian to this direction reads

$$\nabla^2 \hat{J}[\xi](\mathsf{u})\mathsf{s} = \tilde{\mathsf{M}}^{-1}\mathsf{B}^\top \mathsf{h}(\xi) + \gamma \mathsf{s}, \tag{6.7}$$

where $\mathsf{h}(\xi)$ solves

$$[\mathsf{A}(\xi) + \mathsf{N}'(\mathsf{y}(\xi))]\mathsf{h}(\xi) = \mathsf{Q}(\xi)^\top \mathsf{M}_H \mathsf{Q}(\xi)\mathsf{d}(\xi) + \mathsf{M}_\mathsf{L}\big(\mathsf{z}(\xi) \odot \varphi''(\mathsf{y}(\xi)) \odot \mathsf{d}(\xi)\big)$$

with $\mathsf{d}(\xi) = [\mathsf{A}(\xi) + \mathsf{N}'(\mathsf{y}(\xi))]^{-1}\mathsf{B}\mathsf{s}$.

## 6.2. Stochastic Discretization: Polynomial Chaos

We proceed with the discretization of the space $L^p_{\mathbb{P}}(\Xi)$ of vectors of independent random variables distributed on $\Xi := \bigtimes_{i=1}^m \Xi_i$ with the probability measure $\mathbb{P} := \bigotimes_{i=1}^m \mathbb{P}_i$.

**Assumption 6.3.** For the discretization of these spaces we assume the following:

- The sets $\Xi_i \subset \mathbb{R}$ are open and bounded[10] intervals for all $i \in [m]$.
- Each probability measure $\mathbb{P}_i$ does not consist of finitely many atoms such that discretization is really necessary.

For $i \in \{1, \dots, m\}$ and a fixed $d \in \mathbb{N}^m$, the spaces $L^p_{\mathbb{P}_i}(\Xi_i)$ are discretized by polynomials of degree $d_i - 1$. By Assumption 6.3, all polynomials of arbitrary degree defined on $\Xi_i$ are $\mathbb{P}_i$-integrable and the space of polynomials is dense in $L^p_{\mathbb{P}}(\Xi_i)$, see [106, Chap. 8]. Furthermore, there exist sets $\{\beta^{(i)}_{k_i}\}_{k_i=1}^\infty \subset L^2_{\mathbb{P}_i}(\Xi_i)$ of orthonormal polynomials w. r. t. the $L^2_{\mathbb{P}_i}(\Xi_i)$-inner product, where $\beta^{(i)}_{k_i}$ has degree $k_i - 1$ by [106, Thm. 8.5]. These sets are Hilbert bases of $L^2_{\mathbb{P}_i}(\Xi_i)$, respectively, and can be constructed by applying the Grad-Schmidt process to the monomial basis $\{1, \xi_i, \xi_i^2, \dots\}$ for example. We want to mention that some papers [40, 8] dealing with uncertainty quantification restrict the discussion to certain probability distributions, for which the "classical" orthonormal polynomials of increasing degree are well-known. Important examples are the Legendre polynomials for the uniform distribution, the Hermite polynomials for the normal distribution and the Jacobi polynomials for the beta distribution [106, Example 8.2]. In order to have a more general setting, we only assume that the orthonormal polynomials can be constructed and evaluated, e. g., by the three-term recurrence relation [106, Sec. 8.2]. In particular, we do not assume a purely continuous [44, 108, 28, 68] or a symmetric [38] distribution. Defining $\beta_k(\xi) := \prod_{i=1}^m \beta^{(i)}_{k_i}(\xi_i)$ we obtain a set $\{\beta_k\}_{k \in \mathbb{N}^m}$ of orthonormal polynomials which form a Hilbert basis of $L^2_{\mathbb{P}}(\Xi)$, see [116, Thm. 3.12(b)] and Section 2.2. Note that $k$ is an index vector.

Let $\{\mathsf{a}^{(i)}_{l_i}\}_{l_i=1}^{d_i} \subset \Xi_i$ be the $d_i$ pairwise distinct roots of the polynomial $\beta^{(i)}_{d_i+1}$ in ascending order, respectively. They exist and have the mentioned properties due to [106, Thm. 8.16] and are known as Gaussian quadrature nodes. Let $\{\mathsf{w}^{(i)}_{l_i}\}_{l_i=1}^{d_i}$ be the positive Gaussian quadrature

[10] Boundedness of the intervals is assumed because then the set of polynomials of arbitrary degree is a dense subset of $L^p_{\mathbb{P}_i}(\Xi_i)$. This condition can possibly be relaxed, cf. [106, Chap. 8].

weights associated to the nodes $\{\mathsf{a}_{l_i}^{(i)}\}_{l_i=1}^{d_i}$ for $i \in [m]$ and $l_i \in [d_i]$ defined by integration over the respective Lagrange polynomials, see also [106, Def. 9.2]. We use the Gaussian quadrature nodes to define weighted Lagrange polynomials $\{\theta_{k_i}^{(i)}\}_{k_i=1}^{d_i}$, which fulfill $\theta_{k_i}^{(i)}(\mathsf{a}_{l_i}^{(i)}) = \delta_{k_i l_i} \cdot \omega_{k_i}^{(i)}$ for some weights $\omega_{k_i}^{(i)} > 0$. If we choose $\omega_{k_i}^{(i)} = (\mathsf{w}_{k_i}^{(i)})^{-1/2}$ for $k_i \in \{1, \dots, d_i\}$, it follows from the exactness of Gaussian quadrature, that these weighted Lagrange polynomials are also orthonormal:

$$\int_{\Xi_i} \theta_{k_i}^{(i)} \theta_{l_i}^{(i)} \,\mathrm{d}\mathbb{P}_i = \sum_{\hat{\ell}_i=1}^{d_i} \mathsf{w}_{\hat{\ell}_i}^{(i)} \theta_{k_i}^{(i)}(\mathsf{a}_{\hat{\ell}_i}^{(i)}) \theta_{l_i}^{(i)}(\mathsf{a}_{\hat{\ell}_i}^{(i)}) = \sum_{\hat{\ell}_i=1}^{d_i} \mathsf{w}_{\hat{\ell}_i}^{(i)} \delta_{k_i \hat{\ell}_i} (\mathsf{w}_{k_i}^{(i)})^{-1/2} \delta_{l_i \hat{\ell}_i} (\mathsf{w}_{l_i}^{(i)})^{-1/2} = \delta_{k_i l_i}.$$

The product of two Lagrange polynomials has degree $2d_i - 2$ and is integrated exactly by Gaussian quadrature, which is exact up to degree $2d_i - 1$ [106, Thm. 9.9]. Defining $\theta_{k_i}^{(i)} :\equiv \beta_{k_i}^{(i)}$ for $k_i \geq d_i + 1$, we get that $\{\theta_{k_i}^{(i)}\}_{k_i=1}^{\infty}$ is also a Hilbert basis of $L^2_{\mathbb{P}_i}(\Xi_i)$. As before, the Hilbert basis $\{\theta_k\}_{k \in \mathbb{N}^m}$ with $\theta_k(\xi) := \prod_{i=1}^m \theta_{k_i}^{(i)}(\xi_i)$ can be defined. Writing $\mathsf{a}_l := \big(\mathsf{a}_{l_1}^{(1)}, \dots, \mathsf{a}_{l_m}^{(m)}\big)^\top$ and $\omega_l := \prod_{i=1}^m \omega_{l_i}^{(i)}$, it holds that $\theta_k(\mathsf{a}_l) = \delta_{kl} \cdot \omega_l$ for $k \leq d$ componentwise.

The state space $\boldsymbol{Y} = L^p_{\mathbb{P}}(\Xi; Y) \cong Y \otimes L^p_{\mathbb{P}_1}(\Xi_1) \otimes \cdots \otimes L^p_{\mathbb{P}_m}(\Xi_m)$ is discretized by the full tensor product of the respective finite-dimensional subspaces with the basis

$$\{\phi_{k_0} \otimes \theta_{k_1}^{(1)} \otimes \cdots \otimes \theta_{k_m}^{(m)}, k_i \in [d_i], i \in \{0, \dots, m\}\}.$$

This is a basis due to [54, Lem. 3.11], see Remark 2.7. A function $\mathbf{y} \in \mathbf{Y}$ belonging to the finite-dimensional space is represented by a coefficient tensor $\mathbf{y} \in \mathbb{R}^{d_0 \times \cdots \times d_m}$ corresponding to weighted values of the function $\mathbf{y}$ since nodal FE ansatz functions and Lagrange polynomials are used. This means that

$$\mathbf{y}(x, \xi_1, \dots, \xi_m) = \sum_{k_0=1}^{d_0} \sum_{k_1=1}^{d_1} \dots \sum_{k_m=1}^{d_m} \mathbf{y}(k_0, k_1, \dots, k_m) \phi_{k_0}(x) \theta_{k_1}^{(1)}(\xi_1) \cdots \theta_{k_m}^{(m)}(\xi_m) \tag{6.8}$$

and in particular

$$\mathbf{y}(x, \mathsf{a}_{k_1}^{(1)}, \dots, \mathsf{a}_{k_m}^{(m)}) = \sum_{k_0=1}^{d_0} \mathbf{y}(\mathring{k}) \phi_{k_0}(x) \boldsymbol{\omega}(k_1, \dots, k_m), \tag{6.9}$$

where we abbreviate $\mathring{k} = (k_0, k_1, \dots, k_m)$ here, as well as $\mathring{l} = (l_0, l_1, \dots, l_m)$ and $(\mathring{k}, \mathring{l}) = (k_0, k_1, \dots, k_m, l_0, l_1, \dots, l_m)$ etc. in the following in contrast to $l = (l_1, \dots, l_m)$ etc. The weight tensors $\boldsymbol{\omega}, \mathbf{w} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ are defined by $\boldsymbol{\omega}(k) := \prod_{i=1}^m \omega_{k_i}^{(i)}$ and $\mathbf{w}(k) := \prod_{i=1}^m \mathsf{w}_{k_i}^{(i)}$ and obviously have rank 1.

The discretization of the space $L^p_{\mathbb{P}_1}(\Xi_1) \otimes \cdots \otimes L^p_{\mathbb{P}_m}(\Xi_m)$ is obtained from the above considerations by setting $Y = \mathbb{R}$, $d_0 = 1$ and $\phi_1 = 1$. Let $\mathbf{p}$ be a polynomial of *coordinate degree* $d - \mathbb{1}$ belonging to the finite-dimensional subspace $\mathcal{P}_{d-\mathbb{1}}(\Xi) \subset L^p_{\mathbb{P}}(\Xi)$, i. e., $\mathbf{p}$ is a polynomial of degree $d_i - 1$ in the variable $\xi_i$ for all $i \in [d]$. We want to emphasize that this function is

represented by a tensor $\mathbf{p} \in \mathbb{R}^{d_1 \times \cdots \times d_m}$ given by

$$\mathbf{p}(k_1, \dots, k_m) = \mathbf{p}(\mathsf{a}_{k_1}^{(1)}, \dots, \mathsf{a}_{k_m}^{(m)})\, \boldsymbol{\omega}(k_1, \dots, k_m).^{-1} \tag{6.10}$$

due to (6.9). The standard $L^2_{\mathbb{P}}(\Xi)$-inner product of two functions $\mathbf{p}, \tilde{\mathbf{p}}$ is discretized by applying Gaussian quadrature:

$$\begin{aligned}(\mathbf{p}, \tilde{\mathbf{p}})_{L^2_{\mathbb{P}}(\Xi)} = \int_\Xi \mathbf{p}\, \tilde{\mathbf{p}} \,\mathrm{d}\mathbb{P} &= \sum_{k_1=1}^{d_1} \dots \sum_{k_m=1}^{d_m} \mathbf{w}(k_1, \dots, k_m)\, \mathbf{p}(\mathsf{a}_{k_1}^{(1)}, \dots, \mathsf{a}_{k_m}^{(m)})\, \tilde{\mathbf{p}}(\mathsf{a}_{k_1}^{(1)}, \dots, \mathsf{a}_{k_m}^{(m)}) \\ &= \langle \mathbf{w} \odot \boldsymbol{\omega} \odot \mathbf{p}, \boldsymbol{\omega} \odot \tilde{\mathbf{p}} \rangle =: \langle \mathbf{p}, \tilde{\mathbf{p}} \rangle_{\mathbf{w} \odot \boldsymbol{\omega}.^2}.\end{aligned} \tag{6.11}$$

We see that for the special choice of orthonormal polynomials with $\boldsymbol{\omega} = \mathbf{w}.^{-1/2}$ the inner product of two functions is computed by a simple Frobenius inner product of tensors. A special case is the expectation of a function $\mathbf{p}$, computed as an inner product with the function $\mathbb{1}$, which is constant one and represented by the tensor $\boldsymbol{\omega}.^{-1}$.

We now discretize the operators defined in (3.11) by testing with $\mathbf{v} \in \mathbf{Y}$ represented by $\mathbf{v} \in \mathbb{R}^{d_0 \times \cdots \times d_m}$ and using the discretized deterministic operators (6.2), see [46]. This gives

$$\begin{aligned}\langle \boldsymbol{B}\mathrm{u}, \mathbf{v} \rangle_{\boldsymbol{Y}^*, \boldsymbol{Y}} &= \int_\Xi (D\mathrm{u}, \mathbf{v}(\cdot, \xi))_{L^2(\Omega)} \,\mathrm{d}\mathbb{P} \\ &= \sum_{k_0, \dots, k_m = 1}^{d_0, \dots, d_m} \mathbf{v}(k_0, \dots, k_m) (D\mathrm{u}, \phi_{k_0})_{L^2(\Omega)} \int_\Xi \prod_{i=1}^m \theta_{k_i}^{(i)}(\xi_i) \,\mathrm{d}\mathbb{P} \\ &= \langle (\mathsf{B}\mathsf{u}) \otimes (\mathbf{w} \odot \boldsymbol{\omega}), \mathbf{v} \rangle =: \langle \mathbf{B}\mathsf{u}, \mathbf{v} \rangle_{\mathbb{1} \otimes (\mathbf{w} \odot \boldsymbol{\omega}.^2)}\end{aligned}$$

with $\mathbf{B}\mathsf{u} = (\mathsf{B}\mathsf{u}) \otimes \boldsymbol{\omega}.^{-1}$ and

$$\begin{aligned}\langle \boldsymbol{b}, \mathbf{v} \rangle_{\boldsymbol{Y}^*, \boldsymbol{Y}} &= \int_\Xi (f(\xi), \mathbf{v}(\cdot, \xi))_{L^2(\Omega)} \,\mathrm{d}\mathbb{P} \\ &= \sum_{k_0, \dots, k_m = 1}^{d_0, \dots, d_m} \mathbf{v}(k_0, \dots, k_m) \int_\Xi \mathsf{b}_{k_0}(\xi) \prod_{i=1}^m \theta_{k_i}^{(i)}(\xi_i) \,\mathrm{d}\mathbb{P} =: \langle \mathbf{b}, \mathbf{v} \rangle_{\mathbb{1} \otimes (\mathbf{w} \odot \boldsymbol{\omega}.^2)}\end{aligned}$$

with

$$\mathbf{b}(\mathring{k}) = \frac{1}{\prod_{i=1}^m \mathsf{w}_{k_i}^{(i)} (\omega_{k_i}^{(i)})^2} \left( \int_\Xi \mathsf{b}_{k_0}(\xi) \prod_{i=1}^m \theta_{k_i}^{(i)}(\xi_i) \,\mathrm{d}\mathbb{P} \right).$$

We assume that this tensor can be constructed and has sufficiently small rank. In particular, if $\mathsf{b}_{k_0}(\cdot)$ are polynomials of total degree at most $d$, Gaussian quadrature, which is exact up to coordinate degree $2d - \mathbb{1}$, can be applied to compute

$$\int_\Xi \mathsf{b}_{k_0}(\xi) \prod_{i=1}^m \theta_{k_i}^{(i)}(\xi_i) \,\mathrm{d}\mathbb{P} = \sum_{l_1, \dots, l_m = 1}^{d_1, \dots, d_m} \mathbf{w}(l) \mathsf{b}_{k_0}(\mathsf{a}_l) \prod_{i=1}^m \theta_{k_i}^{(i)}(\mathsf{a}_{l_i}^{(i)}) = \mathbf{w}(k) \boldsymbol{\omega}(k) \mathsf{b}_{k_0}(\mathsf{a}_k) \tag{6.12}$$

so that $\mathbf{b}(\bullet, k) = \boldsymbol{\omega}(k).^{-1}\mathsf{b}(\mathsf{a}_k)$ holds. Furthermore,

$$\begin{aligned}
\langle \boldsymbol{A}\mathbf{y}, \mathbf{v}\rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}} &= \int_\Xi (\kappa(\cdot,\xi)\nabla_x \mathbf{y}(\cdot,\xi), \nabla_x \mathbf{v}(\cdot,\xi))_{L^2(\Omega)^n}\,\mathrm{d}\mathbb{P} \\
&= \sum_{k_0,\dots,k_m=1}^{d_0,\dots,d_m} \sum_{l_0,\dots,l_m=1}^{d_0,\dots,d_m} \mathbf{y}(\mathring{l})\mathbf{v}(\mathring{k}) \int_\Xi (\mathsf{A}(\xi))_{k_0 l_0} \prod_{i=1}^m \theta_{l_i}^{(i)}(\xi_i)\theta_{k_i}^{(i)}(\xi_i)\,\mathrm{d}\mathbb{P} \\
&= \langle\langle \tilde{\mathbf{a}}, \mathbf{y}\rangle_{(m+2,\dots,2m+2),[m+1]}, \mathbf{v}\rangle_{\mathbb{1}\otimes(\mathbf{w}\odot\boldsymbol{\omega}.^2)} =: \langle \mathbf{A}\mathbf{y}, \mathbf{v}\rangle_{\mathbb{1}\otimes(\mathbf{w}\odot\boldsymbol{\omega}.^2)}
\end{aligned} \tag{6.13}$$

holds with the tensor $\tilde{\mathbf{a}} \in \mathbb{R}^{d_0\times d_1\times\cdots\times d_m\times d_0\times d_1\times\cdots\times d_m}$ defined by

$$\tilde{\mathbf{a}}(\mathring{k},\mathring{l}) := \frac{1}{\prod_{i=1}^m \mathsf{w}_{k_i}^{(i)}(\omega_{k_i}^{(i)})^2} \left( \int_\Xi (\mathsf{A}(\xi))_{k_0 l_0} \prod_{i=1}^m \theta_{l_i}^{(i)}(\xi_i)\theta_{k_i}^{(i)}(\xi_i)\,\mathrm{d}\mathbb{P} \right). \tag{6.14}$$

The nonlinear part of the equation is approximated by Gaussian quadrature/interpolation to simplify the implementation. We will see later that this relates the whole approach to stochastic collocation.

$$\begin{aligned}
\langle \boldsymbol{N}(\mathbf{y}), \mathbf{v}\rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}} &= \int_\Xi \int_\Omega \varphi(\mathbf{y}(x,\xi))\mathbf{v}(x,\xi)\,\mathrm{d}x\,\mathrm{d}\mathbb{P} \\
&= \sum_{k_0,\dots,k_m=1}^{d_0,\dots,d_m} \mathbf{v}(k_0,\dots,k_m) \int_\Xi \int_\Omega \varphi(\mathbf{y}(x,\xi))\phi_{k_0}(x)\,\mathrm{d}x \prod_{i=1}^m \theta_{k_i}^{(i)}(\xi_i)\,\mathrm{d}\mathbb{P} \\
&\approx \sum_{k_0,\dots,k_m=1}^{d_0,\dots,d_m} \mathbf{v}(\mathring{k}) \sum_{l_1,\dots,l_m=1}^{d_1,\dots,d_m} \mathbf{w}(l) \int_\Omega \varphi(\mathbf{y}(x,\mathsf{a}_l))\phi_{k_0}(x)\,\mathrm{d}x \prod_{i=1}^m \theta_{k_i}^{(i)}(\mathsf{a}_{l_i}^{(i)}) \\
&= \sum_{k_0,\dots,k_m=1}^{d_0,\dots,d_m} \mathbf{v}(\mathring{k})\,\mathbf{w}(k)\boldsymbol{\omega}(k) \int_\Omega \varphi\Big(\sum_{l_0=1}^{d_0} \mathbf{y}(l_0,k_1,\dots,k_m)\phi_{l_0}(x)\boldsymbol{\omega}(k)\Big)\phi_{k_0}(x)\,\mathrm{d}x \\
&\approx \sum_{k_0,\dots,k_m=1}^{d_0,\dots,d_m} \mathbf{v}(\mathring{k})\,\mathbf{w}(k)\boldsymbol{\omega}(k)\varphi\Big(\mathbf{y}(\mathring{k})\boldsymbol{\omega}(k)\Big)(\mathsf{M}_\mathsf{L})_{k_0 k_0} \\
&= \langle (\mathbb{1}\otimes\boldsymbol{\omega}.^{-1}) \odot (\mathsf{M}_\mathsf{L} \circ_1 \varphi((\mathbb{1}\otimes\boldsymbol{\omega})\odot\mathbf{y})), \mathbf{v}\rangle_{\mathbb{1}\otimes(\mathbf{w}\odot\boldsymbol{\omega}.^2)} \\
&= \langle \hat{\boldsymbol{\omega}}.^{-1} \odot (\mathsf{M}_\mathsf{L} \circ_1 \varphi(\hat{\boldsymbol{\omega}}\odot\mathbf{y})), \mathbf{v}\rangle_{\mathbb{1}\otimes(\mathbf{w}\odot\boldsymbol{\omega}.^2)} =: \langle \mathbf{N}(\mathbf{y}), \mathbf{v}\rangle_{\mathbb{1}\otimes(\mathbf{w}\odot\boldsymbol{\omega}.^2)},
\end{aligned} \tag{6.15}$$

where the first approximate equality is due to Gaussian quadrature in the parameter space and the second one is due to the FE nodes based quadrature, which is related to mass lumping, see (6.3). We define the rank-1-tensor $\hat{\boldsymbol{\omega}} := \mathbb{1}\otimes\boldsymbol{\omega}$ for readability purposes. Here we see the importance of using mass lumping and weighted Lagrange polynomials: It yields that the nonlinear function $\varphi : \mathbb{R} \to \mathbb{R}$ can be applied componentwise to a tensor in the discrete setting (6.15).

The discrete state equation reads

$$\mathbf{A}\mathbf{y} + \mathbf{N}(\mathbf{y}) = \mathbf{B}\mathsf{u} + \mathbf{b}. \tag{6.16}$$

**Assumption 6.4.** In order to be able to evaluate the objective function (3.4) on the finite-dimensional subspace exactly in a simple way, we need the following additional assumptions:

- The operator $\mathsf{Q}(\xi) = \mathsf{Q}$ is constant.
- The desired state $\hat{\mathsf{q}}(\cdot)$ is a polynomial of coordinate degree at most $d - \mathbb{1}$.

The discrete version of the objective function is

$$\begin{aligned}\mathsf{J}(\mathsf{y},\mathsf{u}) &:= \int_\Xi \mathsf{J}[\xi]\Big(\sum_{k_1,\dots,k_m=1}^{d_1,\dots,d_m} \mathsf{y}(\bullet,k_1,\dots,k_m)\theta_{k_1}^{(1)}(\xi_1)\cdots\theta_{k_m}^{(m)}(\xi_m),\mathsf{u}\Big)\,\mathrm{d}\mathbb{P}\\ &= \int_\Xi \tfrac{1}{2}\Big\|\sum_{k_1,\dots,k_m=1}^{d_1,\dots,d_m} \mathsf{Q}(\xi)\mathsf{y}(\bullet,k_1,\dots,k_m)\theta_{k_1}^{(1)}(\xi_1)\cdots\theta_{k_m}^{(m)}(\xi_m) - \hat{\mathsf{q}}(\xi)\Big\|_{\mathsf{M}_H}^2\,\mathrm{d}\mathbb{P} + \tfrac{\gamma}{2}\mathsf{u}^\top\tilde{\mathsf{M}}\mathsf{u},\end{aligned} \tag{6.17}$$

cf. (6.4). Since $\mathsf{Q}(\xi) = \mathsf{Q}$ is constant and $\hat{\mathsf{q}}(\cdot)$ is a polynomial of coordinate degree at most $d-\mathbb{1}$ (Assumption 6.4), the integrand in (6.17) has degree at most $2d-2\cdot\mathbb{1}$ and the integral can be evaluated exactly by Gaussian quadrature. This gives

$$\begin{aligned}\mathsf{J}(\mathsf{y},\mathsf{u}) &= \sum_{l_1,\dots,l_m}^{d_1,\dots,d_m} \mathbf{w}(l)\,\tfrac{1}{2}\big\|\boldsymbol{\omega}(l)\,(\mathsf{Q}\,\mathsf{y}(\bullet,l_1,\dots,l_m)) - \hat{\mathsf{q}}(\mathsf{a}_l)\big\|_{\mathsf{M}_H}^2 + \tfrac{\gamma}{2}\mathsf{u}^\top\tilde{\mathsf{M}}\mathsf{u}\\ &=: \langle \mathbf{M_H}(\mathbf{Qy}-\hat{\mathbf{q}}), \mathbf{Qy}-\hat{\mathbf{q}}\rangle + \tfrac{\gamma}{2}\mathsf{u}^\top\tilde{\mathsf{M}}\mathsf{u}\end{aligned} \tag{6.18}$$

with $\mathbf{Qy} = \mathsf{Q}\circ_1\mathbf{y}$, $\hat{\mathbf{q}}\in\mathbb{R}^{d_H\times d_1\times\cdots\times d_m}$ defined by $\hat{\mathbf{q}}(\bullet,l) = \hat{\mathsf{q}}(\mathsf{a}_l)\,\boldsymbol{\omega}(l)^{-1}$, and

$$\mathbf{M_H} : \mathbb{R}^{d_H\times d_1\times\cdots\times d_m}\to\mathbb{R}^{d_H\times d_1\times\cdots\times d_m},\ \mathbf{M_H}\hat{\mathbf{q}} := (\mathbb{1}\otimes(\mathbf{w}\odot\boldsymbol{\omega}.^2))\odot(\mathsf{M}_H\circ_1\hat{\mathbf{q}}). \tag{6.19}$$

Again, this evaluation is exact on the discrete subspace and the objective function evaluation error can be estimated using Proposition 5.11.

For the computation of the gradient of the reduced objective function, it remains to discretize the adjoint equation (3.18). The adjoint state $\mathbf{z}$ is represented by the tensor $\mathbf{z}$ analogously to the state. We have

$$\begin{aligned}\langle \boldsymbol{N}'(\mathbf{y})\mathbf{z},\mathbf{v}\rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}} &= \int_\Xi\int_\Omega \varphi'(\mathbf{y}(x,\xi))\mathbf{z}(x,\xi)\mathbf{v}(x,\xi)\,\mathrm{d}x\,\mathrm{d}\mathbb{P}\\ &\approx \sum_{k_0,\dots,k_m=1}^{d_0,\dots,d_m}\mathbf{v}(\mathring{k})\sum_{l_1,\dots,l_m=1}^{d_1,\dots,d_m}\mathbf{w}(l)\int_\Omega\varphi'(\mathbf{y}(x,\mathsf{a}_l))\mathbf{z}(x,\mathsf{a}_l)\phi_{k_0}(x)\,\mathrm{d}x\prod_{i=1}^m\theta_{k_i}^{(i)}(\mathsf{a}_{l_i}^{(i)})\\ &\approx \sum_{k_0,\dots,k_m=1}^{d_0,\dots,d_m}\mathbf{v}(\mathring{k})\,\mathbf{w}(k)\boldsymbol{\omega}(k)\varphi'\Big(\mathbf{y}(\mathring{k})\boldsymbol{\omega}(k)\Big)\mathbf{z}(\mathring{k})\boldsymbol{\omega}(k)(\mathsf{M}_\mathsf{L})_{k_0k_0}\\ &= \big\langle\hat{\boldsymbol{\omega}}.^{-1}\odot\big(\mathsf{M}_\mathsf{L}\circ_1\big(\varphi'(\hat{\boldsymbol{\omega}}\odot\mathbf{y})\odot(\hat{\boldsymbol{\omega}}\odot\mathbf{z})\big)\big),\mathbf{v}\big\rangle_{\mathbb{1}\otimes(\mathbf{w}\odot\boldsymbol{\omega}.^2)}\\ &= \langle\mathbf{N}'(\mathbf{y})\mathbf{z},\mathbf{v}\rangle_{\mathbb{1}\otimes(\mathbf{w}\odot\boldsymbol{\omega}.^2)},\end{aligned}$$

cf. (6.15), where $\mathbf{N}'$ is exactly the derivative of the discretized operator $\mathbf{N}$. Furthermore, the right-hand side is

$$\begin{aligned}
&\langle -\boldsymbol{Q}^*(\boldsymbol{Q}\mathbf{y} - \hat{\boldsymbol{q}}), \mathbf{v}\rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}} \\
&= \sum_{k_1,\dots,k_m=1}^{d_1,\dots,d_m} \mathbf{v}(\bullet,k)^\top \int_\Xi \Big(-\mathsf{Q}^\top \mathsf{M}_H \Big( \sum_{l_1,\dots,l_m=1}^{d_1,\dots,d_m} \mathsf{Q}\,\mathbf{y}(\bullet,l) \prod_{i=1}^m \theta_{l_i}^{(i)}(\xi_i) - \hat{\mathsf{q}}(\xi)\Big)\Big) \prod_{i=1}^m \theta_{k_i}^{(i)}(\xi_i)\,\mathrm{d}\mathbb{P} \\
&= \sum_{k_1,\dots,k_m=1}^{d_1,\dots,d_m} \mathbf{v}(\bullet,k)^\top \mathbf{w}(k)\boldsymbol{\omega}(k)\Big(-\mathsf{Q}^\top \mathsf{M}_H \Big(\mathsf{Q}\,\mathbf{y}(\bullet,k)\boldsymbol{\omega}(k) - \hat{\mathsf{q}}(\mathsf{a}_k)\Big)\Big) \\
&= \langle -(\mathsf{Q}^\top \mathsf{M}_H) \circ_1 (\mathbf{Q}\mathbf{y} - \hat{\mathbf{q}}), \mathbf{v}\rangle_{\mathbb{1}\otimes(\mathbf{w}\odot\boldsymbol{\omega}.^2)}, \qquad (6.20)
\end{aligned}$$

cf. (6.5). Given the solution $\mathbf{z}$ of the discretized adjoint equation

$$\mathbf{A}\mathbf{z} + \mathbf{N}'(\mathbf{y})\mathbf{z} = -(\mathsf{Q}^\top \mathsf{M}_H) \circ_1 (\mathbf{Q}\mathbf{y} - \hat{\mathbf{q}}), \qquad (6.21)$$

the gradient of the reduced, discretized objective function reads

$$\nabla \hat{\mathbf{J}}(\mathsf{u}) = -(\tilde{\mathsf{M}}^{-1}\mathsf{B}^\top)\langle \mathbf{z}, \mathbf{w}\odot\boldsymbol{\omega}\rangle_{[m]+1,[m]} + \gamma \mathsf{u}, \qquad (6.22)$$

cf. (6.6). Due to $\mathbf{w}\odot\boldsymbol{\omega} = (\mathbf{w}\odot\boldsymbol{\omega}.^2)\odot\boldsymbol{\omega}.^{-1}$, the term $\langle \mathbf{z}, \mathbf{w}\odot\boldsymbol{\omega}\rangle_{[m]+1,[m]}$ corresponds to computing the $L^2_{\mathbb{P}}(\Xi)$-inner product (induced by $\mathbf{w}\odot\boldsymbol{\omega}.^2$) of $\mathbf{z}$ and the function $\mathbb{1}$ (represented by $\boldsymbol{\omega}.^{-1}$), i. e., computing the expectation of $\mathbf{z}$, see also (6.11) and the paragraph below this equation. In [46] we used differently scaled formulations of (6.21) and (6.22), which coincide to the ones given here in the case $\mathbf{w} = \boldsymbol{\omega}.^{-1/2}$ (orthonormal polynomials). Here the form (6.21) of the adjoint equation is closer to the continuous setting as derived in (6.20). As in (6.6), the matrix $(\tilde{\mathsf{M}}^{-1}\mathsf{B}^\top)$ can often be applied without inverting the mass matrix $\tilde{\mathsf{M}}$ so that Theorem 5.8 can be applied to estimate the gradient error.

In analogy to (6.7), we obtain the discrete Hessian

$$\nabla^2 \hat{\mathbf{J}}(\mathsf{u})\mathsf{s} = (\tilde{\mathsf{M}}^{-1}\mathsf{B}^\top)\langle \mathbf{h}, \mathbf{w}\odot\boldsymbol{\omega}\rangle_{[m]+1,[m]} + \gamma \mathsf{s}, \qquad (6.23)$$

where $\mathbf{h} \in \mathbb{R}^{d_0\times\cdots\times d_m}$ solves

$$[\mathbf{A} + \mathbf{N}'(\mathbf{y})]\mathbf{h} = (\mathsf{Q}^\top \mathsf{M}_H) \circ_1 (\mathbf{Q}\mathbf{d}) + \hat{\boldsymbol{\omega}}.^{-1} \odot \big(\mathsf{M}_\mathsf{L} \circ_1 \big((\hat{\boldsymbol{\omega}}\odot\mathbf{z})\odot\varphi''(\hat{\boldsymbol{\omega}}\odot\mathbf{y})\odot(\hat{\boldsymbol{\omega}}\odot\mathbf{d})\big)\big) \qquad (6.24)$$

with $\mathbf{d} = [\mathbf{A} + \mathbf{N}'(\mathbf{y})]^{-1}\mathbf{B}\mathsf{s}$ and the current state $\mathbf{y}$ and adjoint state $\mathbf{z}$. The second summand at the right hand side of (6.24) is exactly the derivative $\mathbf{N}''(\mathbf{y})$ applied to $\mathbf{z}$ and $\mathbf{d}$. In (6.24) we could get rid of the componentwise multiplication by $\hat{\boldsymbol{\omega}}.^{-1}$ by, e. g., dropping the multiplication of $\mathbf{d}$ by $\hat{\boldsymbol{\omega}}$. But we want to emphasize that this approach is related to interpolation: The tensor $\hat{\boldsymbol{\omega}}\odot\mathbf{y}$ for example contains exactly the function values of the state at the FE nodes and Gaussian quadrature grid points. Therefore, pointwise operations, such as the application of $\varphi''$ or the multiplication of functions, carry over to componentwise operations on tensors in the discrete setting. Multiplying a tensor of functions values by $\hat{\boldsymbol{\omega}}.^{-1}$ again transforms it back to the representation with weighted Lagrange polynomials.

### Application of the Operators to Low-Rank Tensors

To make the solution of the state and the adjoint equation and further computations efficient, we represent all tensors, such as $\mathbf{y}$ and $\mathbf{z}$, in a low-rank format (TT or HT, see Subsection 2.1.2). As already noted in [40], this is related to reduced basis methods [29, 28]. Let for example $\mathbf{y} \in \mathbb{R}^{d_0 \times \cdots \times d_m}$ be given in the HT format. Then each leaf matrix $\mathsf{U}_i \in \mathbb{R}^{d_i \times r_i}$ (see Subsection 2.1.2) contains a generating system of a lower-dimensional subspace of $\mathbb{R}^{d_i}$, typically even a basis of an $r_i$-dimensional subspace. These bases can be interpreted as bases of subspaces of the discrete spaces Y, $\mathcal{P}_{d_1-1}(\Xi_1)$, …, $\mathcal{P}_{d_m-1}(\Xi_m)$, respectively.

Equations (6.16) and (6.21) will be solved by a low-rank tensor solver such as AMEn, which chooses the tensor rank adaptively. In particular, the basis $\mathsf{U}_0$ for the subspace of the deterministic discrete state space Y may be adapted by the low-rank tensor solver during the solution process and it may even change its size. This is in contrast to the standard reduced basis method, where the basis is chosen a priori based on solution snapshots. A second difference is that also reduced bases for the polynomial spaces $\mathcal{P}_{d_i-1}(\Xi_i)$ appear in the low-rank tensor.

In order to be able to apply such a low-rank tensor solver, it is important that all operators are efficiently applicable to low-rank tensors. This shall be discussed in the following. We see that the nonlinear function $\mathbf{N}$ as well as its derivatives $\mathbf{N}'$, $\mathbf{N}''$ can be applied to low-rank tensors as long as the elementwise application of the functions $\varphi$, $\varphi'$, and $\varphi''$ can be implemented efficiently because the rest of the computations consists of multiplications with rank-1-tensors, $i$-mode matrix multiplications, and componentwise multiplications to apply $\mathbf{N}'(\mathbf{y})$ and $\mathbf{N}''(\mathbf{y})$. In particular, the componentwise multiplication by a rank-1-tensor can be written as $i$-mode multiplications with diagonal matrices. In our case, it holds that

$$\mathbf{N}(\mathbf{y}) = \mathsf{M}_\mathsf{L} \circ_1 \operatorname{diag}(\omega^{(1)}.^{-1}) \circ_2 \cdots \operatorname{diag}(\omega^{(m)}.^{-1}) \circ_{m+1} \varphi(\operatorname{diag}(\omega^{(1)}) \circ_2 \cdots \operatorname{diag}(\omega^{(m)}) \circ_{m+1} \mathbf{y}). \tag{6.25}$$

Analogously, the operator $\mathbf{M}_{\boldsymbol{H}}$ defined in (6.19) can be implemented efficiently using $i$-mode matrix products only. More $i$-mode matrix products, outer and inner products appear in form of the operator $\mathbf{Q}$ used in (6.18) and (6.21), the operator $\mathbf{B}$, and in (6.20), (6.22), and (6.23). There, the contraction with a rank-1-tensor can be carried out by

$$\langle \mathbf{z}, \mathbf{w} \odot \boldsymbol{\omega} \rangle_{[m]+1,[m]} = \big( (\mathsf{w}^{(1)} \odot \omega^{(1)})^\top \circ_2 \ldots (\mathsf{w}^{(m)} \odot \omega^{(m)})^\top \circ_{m+1} \mathbf{z} \big)(\bullet, 1, \ldots, 1),$$

where the indexing at the end is done to remove the tensor modes of dimension one. It only remains to investigate the application of the operator $\mathbf{A}$ to low-rank tensors. Then all operations in (6.16), (6.21), (6.22), and (6.23) can be realized in an efficient way.

By (6.13), the application of the operator $\mathbf{A}$ can be performed as contraction with the tensor $\tilde{\mathbf{a}}$ defined in (6.14). If this tensor can be represented or approximated in the used low-rank format, this contraction can be computed easily. More concretely, it can also represent a TT matrix as described in Subsection 2.1.3. The contraction $\langle \tilde{\mathbf{a}}, \mathbf{y} \rangle_{(m+2,\ldots,2m+2),[m+1]}$ is exactly the application of a TT matrix to a TT tensor as depicted in Figure 2.6. Another favorable option is having the operator $\mathbf{A}$ of small CP rank. We can achieve this by using additional structure.

**Example 6.5** ([46, Example 3.3, Example 3.6])**.** In light of Example 3.17 and Lemma 3.18, we consider the specific form $\kappa(x,\xi) = \kappa_0(x)(1+\sum_{i=1}^m \xi_i\eta_i(x))$ of the coefficient function. The definitions of the operators $A_0, A_i \in \mathcal{L}(Y,Y^*)$ from Lemma 3.18 carry over to the discrete setting via

$$\begin{aligned}\mathsf{A}_0 &\in \mathbb{R}^{d_0\times d_0}, \quad (\mathsf{A}_0)_{k_0l_0} := (\kappa_0\nabla\phi_{l_0}, \nabla\phi_{k_0})_{L^2(\Omega)^n},\\ \mathsf{A}_i &\in \mathbb{R}^{d_0\times d_0}, \quad (\mathsf{A}_i)_{k_0l_0} := (\eta_i\kappa_0\nabla\phi_{l_0}, \nabla\phi_{k_0})_{L^2(\Omega)^n},\end{aligned}$$

cf. (6.2), and it holds that $\mathsf{A}(\xi) = \mathsf{A}_0 + \sum_{i=1}^m \xi_i\mathsf{A}_i$. Writing $\theta_{l_j}^{(j)}\theta_{k_j}^{(j)} = \theta_{l_j}^{(j)}(\xi_j)\theta_{k_j}^{(j)}(\xi_j)$ for brevity, we obtain

$$\begin{aligned}&\int_\Xi (\mathsf{A}(\xi))_{k_0l_0}\prod_{j=1}^m \theta_{l_j}^{(j)}(\xi_j)\theta_{k_j}^{(j)}(\xi_j)\,\mathrm{d}\mathbb{P}\\
&= (\mathsf{A}_0)_{k_0l_0}\prod_{j=1}^m\int_{\Xi_j}\theta_{l_j}^{(j)}\theta_{k_j}^{(j)}\,\mathrm{d}\mathbb{P}_j + \sum_{i=1}^m(\mathsf{A}_i)_{k_0l_0}\int_{\Xi_i}\xi_i\theta_{l_i}^{(i)}\theta_{k_i}^{(i)}\,\mathrm{d}\mathbb{P}_i\prod_{j\neq i}\int_{\Xi_j}\theta_{l_j}^{(j)}\theta_{k_j}^{(j)}\,\mathrm{d}\mathbb{P}_j\\
&= (\mathsf{A}_0)_{k_0l_0}\prod_{j=1}^m \mathsf{w}_{l_j}^{(j)}(\omega_{l_j}^{(j)})^2\delta_{l_jk_j} + \sum_{i=1}^m(\mathsf{A}_i)_{k_0l_0}\mathsf{a}_{l_i}^{(i)}\mathsf{w}_{l_i}^{(i)}(\omega_{l_i}^{(i)})^2\delta_{l_ik_i}\prod_{j\neq i}\mathsf{w}_{l_j}^{(j)}(\omega_{l_j}^{(j)})^2\delta_{l_jk_j}\\
&= \prod_{j=1}^m \mathsf{w}_{l_j}^{(j)}(\omega_{l_j}^{(j)})^2\Big((\mathsf{A}_0)_{k_0l_0} + \sum_{i=1}^m(\mathsf{A}_i)_{k_0l_0}\mathsf{a}_{l_i}^{(i)}\Big)\prod_{j=1}^m\delta_{l_jk_j}\end{aligned}$$

using that Gaussian quadrature is exact up to degree $2d_i - 1$. Therefore, (6.14) becomes

$$\tilde{\mathbf{a}}(\mathring{k},\mathring{l}) = \Big((\mathsf{A}_0)_{k_0l_0} + \sum_{i=1}^m(\mathsf{A}_i)_{k_0l_0}\mathsf{a}_{l_i}^{(i)}\Big)\prod_{j=1}^m\delta_{l_jk_j}.$$

This gives

$$\begin{aligned}\mathbf{A}\mathbf{y} = \langle\tilde{\mathbf{a}},\mathbf{y}\rangle_{(m+2,\dots,2m+2),[m+1]} &= \sum_{l_0,\dots,l_m=1}^{d_0,\dots,d_m}\tilde{\mathbf{a}}(\mathring{k},\mathring{l})\mathbf{y}(\mathring{l})\\
&= \sum_{l_0=1}^{d_0}\Big((\mathsf{A}_0)_{k_0l_0} + \sum_{i=1}^m(\mathsf{A}_i)_{k_0l_0}\mathsf{a}_{k_i}^{(i)}\Big)\mathbf{y}(l_0,k_1,\dots,k_m)\\
&= \mathsf{A}_0\circ_1\mathbf{y} + \sum_{i=1}^m\mathsf{A}_i\circ_1\tilde{\mathsf{S}}_i\circ_{i+1}\mathbf{y}\end{aligned}$$

with $\tilde{\mathsf{S}}_i = \mathrm{diag}(\mathsf{a}^{(i)})$ in analogy to Lemma 3.18. We see that the operator $\mathbf{A}$ admits CP rank $m+1$ and can be implemented using simple $i$-mode matrix multiplications and summation.

In [46] an additional example is given, where a similar structure is derived for an operator with additional domain parametrizations via interpolation. As mentioned at the end of Section 3.2, we skip this example in this thesis because the a posteriori error estimator derived in Chapter 7 will also rely on the structure of the coefficient function used in Example 6.5.

### Relation to Stochastic Collocation

With the structure of the operator $\mathbf{A}$ from Example 6.5 and the approximation of the nonlinearity (6.15) and the right-hand side (6.12) by Gaussian quadrature, we obtain that the described stochastic Galerkin discretization of the state equation (6.16) yields in fact a completely decoupled system, one nonlinear equation for each combination $(\mathsf{a}_{k_1}^{(1)}, \dots, \mathsf{a}_{k_m}^{(m)})$ of parameter realizations:

$$\begin{aligned}
(\mathbf{A}\mathbf{y})(\bullet, k) &= (\mathsf{A}_0 + \sum_{i=1}^{m} \mathsf{a}_{k_i}^{(i)} \mathsf{A}_i)\mathbf{y}(\bullet, k) = \boldsymbol{\omega}(k)^{-1}\, \mathsf{A}(\mathsf{a}_{k_i}^{(i)})(\boldsymbol{\omega}(k)\mathbf{y}(\bullet, k)), \\
(\mathbf{N}(\mathbf{y}))(\bullet, k) &= (\hat{\boldsymbol{\omega}}.^{-1} \odot (\mathsf{M}_\mathsf{L} \circ_1 \varphi(\hat{\boldsymbol{\omega}} \odot \mathbf{y})))(\bullet, k) = \boldsymbol{\omega}(k)^{-1}\, \mathsf{M}_\mathsf{L}\, \varphi(\boldsymbol{\omega}(k)\mathbf{y}(\bullet, k)), \\
(\mathbf{B}\mathsf{u})(\bullet, k) &= ((\mathsf{B}\mathsf{u}) \otimes \boldsymbol{\omega}.^{-1})(\bullet, k) = \boldsymbol{\omega}(k)^{-1}\, \mathsf{B}\mathsf{u}, \\
\mathbf{b}(\bullet, k) &= \boldsymbol{\omega}(k).^{-1}\mathsf{b}(\mathsf{a}_k).
\end{aligned}$$

The vector $\boldsymbol{\omega}(k)\mathbf{y}(\bullet, k)$ represents the FE function $\mathbf{y}(\cdot, \mathsf{a}_{k_1}^{(1)}, \dots, \mathsf{a}_{k_m}^{(m)})$, see (6.9). This relates the approach to stochastic collocation, cf. [44], where one would have $\boldsymbol{\omega}(k) = 1$ for all $k$ and then get the same weight 1 for all equations obtained by inserting the collocation points into (6.1). Taking orthonormal polynomials as basis, these single equations are weighted differently, namely by $\boldsymbol{\omega}(k)^{-1} = \mathbf{w}(k)^{1/2}$, in the tensor version.

## 6.3. Choice of the Trust-Region Model

We describe the choice of the trust-region model in the stochastic setting only because the deterministic setting works analogously. The main difference is that the stochastic setting requires further approximations of the Hessian to have efficient computations. Let $\mathrm{U} \subset U$ be the current discrete subspace and let $\mathrm{u}^k \in \mathrm{U}$ be the current iterate, represented by the vector $\mathsf{u}^k \in \mathbb{R}^{d_u}$. The tensor $\tilde{\mathbf{y}}^k$ shall solve the state equation (6.16) with $\mathsf{u} = \mathsf{u}^k$ approximately and let $\tilde{\mathbf{z}}^k$ be an approximate solution of the adjoint equation (6.21) with $\mathbf{y} = \tilde{\mathbf{y}}^k$. We have to keep in mind that we apply iterative low-rank tensor solvers to these equations so that we cannot expect to obtain exact solutions.

In order to be able to apply the error estimates from Chapter 5, we choose a quadratic trust-region model $m_k$ for any direction $\mathrm{s} \in \mathrm{U}$ represented by the vector $\mathsf{s}$ as follows:

$$m_k(\mathrm{s}) = \mathsf{m}_k(\mathsf{s}) := \nabla \mathsf{m}_k(0)^\top \tilde{\mathsf{M}}\mathsf{s} + \tfrac{1}{2}\mathsf{s}^\top \tilde{\mathsf{M}} \nabla^2 \mathsf{m}_k(0)\mathsf{s}$$

The space $\mathbb{R}^{d_u}$ is equipped with the inner product induced by the mass matrix $\tilde{\mathsf{M}}$ and derivatives, such as $\nabla \mathsf{m}_k$, are computed w. r. t. this inner product. To approximate the true gradient sufficiently well, we choose

$$\nabla \mathsf{m}_k(0) = -(\tilde{\mathsf{M}}^{-1}\mathsf{B}^\top)\langle \tilde{\mathbf{z}}^k, \mathbf{w} \odot \boldsymbol{\omega} \rangle_{[m]+1,[m]} + \gamma \mathsf{u}^k$$

in light of (6.22) and Theorem 5.8. In principle, we could choose $\nabla^2 m_k(0)$ to be computed as in (6.23), but this evaluation requires two stochastic PDE solves, which is on the one hand costly and on the other hand inexact due to the low-rank tensor solver so that the Hessian

model would be noisy. To overcome this issue, we approximate the operator $\mathbf{A} + \mathbf{N}'(\tilde{\mathbf{y}}^k)$ by an operator of CP rank 1, which can be inverted easily. We do this by using a reference operator acting only on the space mode.

To approximate the operator $\mathbf{A}$, we can choose $\mathsf{A}_{\mathrm{ref}} := A(\bar{\xi})$, where $\bar{\xi} = \int_\Xi \xi \,\mathrm{d}\mathbb{P}$, or $\mathsf{A}_{\mathrm{ref}} := \mathbb{E}[\mathsf{A}(\cdot)]$ for example. If $\mathsf{A} : \Xi \to \mathbb{R}^{d_0 \times d_0}$ is affine as in Example 6.5, the two choices coincide. We can approximate

$$\mathbf{A} \approx \mathbf{A}_{\mathrm{ref}} := \mathsf{A}_{\mathrm{ref}} \otimes \mathsf{I} \cdots \otimes \mathsf{I}$$

and have $\mathbf{A}_{\mathrm{ref}}^{-1} = \mathsf{A}_{\mathrm{ref}}^{-1} \otimes \mathsf{I} \cdots \otimes \mathsf{I}$, i. e., the inverse can be applied to low-rank tensors by computing a simple $i$-mode matrix product. In Chapter 7, it will be shown that it makes sense to use the operator $\mathbf{A}_{\mathrm{ref}}$ as a preconditioner for the solution of equations of the form (6.16) and (6.21). Then, the operator $\mathbf{A}_{\mathrm{ref}}^{-1}$ has to be applied many times so that it is suitable to increase efficiency by, e. g., computing a sparse Cholesky decomposition of $\mathsf{A}_{\mathrm{ref}}$ once so that applying the inverse $\mathsf{A}_{\mathrm{ref}}^{-1}$ to a vector consists only of permutations as well as forward and backward substitutions.

For the approximation of the operator $\mathbf{N}'(\mathbf{y})$ we introduce the parameter dependent vector

$$\tilde{\mathsf{y}}(\xi) := \sum_{k_1=1}^{d_1} \dots \sum_{k_m=1}^{d_m} \tilde{\mathbf{y}}(\bullet, k_1, \dots, k_m) \theta_{k_1}^{(1)}(\xi_1) \cdots \theta_{k_m}^{(m)}(\xi_m)$$

cf. (6.8), and propose three options:

- Firstly, we can take $\mathsf{N}'_{\mathrm{ref}}(\tilde{\mathbf{y}}) := \mathsf{N}'(\langle \tilde{\mathbf{y}}, \mathbf{w} \odot \boldsymbol{\omega} \rangle_{[m]+1,[m]})$, i. e., we evaluate $\mathsf{N}'$ at the expected value $\mathbb{E}[\tilde{\mathsf{y}}(\cdot)]$ of the current state. This is the option we use in the implementation because after having computed the expected value, the code from the deterministic setting can be reused.

- Secondly, and similarly, $\mathsf{N}'_{\mathrm{ref}}(\tilde{\mathbf{y}}) := \mathsf{N}'(\langle \tilde{\mathbf{y}}, \boldsymbol{\theta}(\bar{\xi}) \rangle_{[m]+1,[m]})$ can be taken, where $\boldsymbol{\theta}(\bar{\xi})(k) = \prod_{i=1}^m \theta_{k_i}^{(i)}(\bar{\xi}_i)$ meaning that $\mathsf{N}'$ is evaluated at the "reference state" $\tilde{\mathsf{y}}(\bar{\xi})$.

- A third option would be to define $\mathsf{N}'_{\mathrm{ref}}(\tilde{\mathbf{y}}) := \mathsf{M}_{\mathsf{L}} \operatorname{diag}(\langle \tilde{\boldsymbol{\varphi}}', \mathbf{w} \rangle_{[m]+1,[m]}) \approx \mathbb{E}[\mathsf{N}'(\tilde{\mathsf{y}}(\cdot))]$ with $\tilde{\boldsymbol{\varphi}}' \approx \varphi'(\hat{\boldsymbol{\omega}} \odot \tilde{\mathbf{y}})$. This option seems to be costly because the nonlinear function $\varphi'$ has to be evaluated on the full tensor $\tilde{\mathbf{y}}$, but the quantity can be reused from the solution of the adjoint equation.

Using one of these options, the the final rank-1-approximation of the stochastic, linearized PDE operator is given by

$$\mathbf{A} + \mathbf{N}'(\tilde{\mathbf{y}}^k) \approx (\mathsf{A}_{\mathrm{ref}} + \mathsf{N}'_{\mathrm{ref}}(\tilde{\mathbf{y}}^k)) \otimes \mathsf{I} \otimes \cdots \otimes \mathsf{I} =: \mathbf{A}_{\mathrm{ref}} + \mathbf{N}'_{\mathrm{ref}}(\tilde{\mathbf{y}}^k).$$

With this approximation, the tensor $\mathbf{d}$ from (6.24) is inexactly computed via

$$\mathbf{d} \approx \tilde{\mathbf{d}}^k := (\mathbf{A}_{\mathrm{ref}} + \mathbf{N}'_{\mathrm{ref}}(\tilde{\mathbf{y}}^k))^{-1} \mathbf{B}\mathsf{s} = ((\mathsf{A}_{\mathrm{ref}} + \mathsf{N}'_{\mathrm{ref}}(\tilde{\mathbf{y}}^k))^{-1} \mathsf{B}\mathsf{s}) \otimes \boldsymbol{\omega}.^{-1}.$$

Given this rank-1-tensor, the tensor $\mathbf{h}$ from (6.24) can be approximated via

$$\mathbf{h} \approx \tilde{\mathbf{h}}^k := (\mathbf{A}_{\mathrm{ref}} + \mathbf{N}'_{\mathrm{ref}}(\tilde{\mathbf{y}}^k))^{-1} \Big[ (\mathsf{Q}^\top \mathsf{M}_H \mathsf{Q}) \circ_1 \tilde{\mathbf{d}}^k + \mathsf{M}_{\mathsf{L}} \circ_1 \left( \tilde{\boldsymbol{\zeta}}^k \odot \tilde{\mathbf{d}}^k \right) \Big],$$

where $\tilde{\boldsymbol{\zeta}}^k \approx \hat{\boldsymbol{\omega}} \odot \tilde{\mathbf{z}}^k \odot \varphi''(\hat{\boldsymbol{\omega}} \odot \tilde{\mathbf{y}}^k)$. Using this approximation in (6.23) defines the model Hessian, applied to a direction $\mathsf{s}$:

$$\begin{aligned}\nabla^2 \mathsf{m}_k(0)\mathsf{s} &= (\tilde{\mathsf{M}}^{-1}\mathsf{B}^\top)\langle \tilde{\mathbf{h}}_k, \mathbf{w} \odot \boldsymbol{\omega}\rangle_{[m]+1,[m]} + \gamma \mathsf{s} \\ &= (\tilde{\mathsf{M}}^{-1}\mathsf{B}^\top)(\tilde{\mathsf{A}}_{\mathrm{ref}}^{(k)})^{-1}\Big[(\mathsf{Q}^\top \mathsf{M}_H \mathsf{Q}) + \mathsf{M}_\mathsf{L}\ \mathrm{diag}(\langle \tilde{\boldsymbol{\zeta}}^k, \mathbf{w}\rangle_{[m]+1,[m]})\Big](\tilde{\mathsf{A}}_{\mathrm{ref}}^{(k)})^{-1}\mathsf{B}\mathsf{s} + \gamma \mathsf{s}\end{aligned} \tag{6.26}$$

using $\langle \boldsymbol{\omega}.^{-1}, \mathbf{w} \odot \boldsymbol{\omega}\rangle = 1$ (sum of quadrature weights) and with $\tilde{\mathsf{A}}_{\mathrm{ref}}^{(k)} := \mathsf{A}_{\mathrm{ref}} + \mathsf{N}'_{\mathrm{ref}}(\tilde{\mathbf{y}}^k)$. This Hessian operator is symmetric on $\mathbb{R}^{d_u}$ w. r. t. the inner product induced by $\tilde{\mathsf{M}}$. In contrast to [46], we use the concrete rank-1-property of all involved operators here to derive that after forming $\langle \tilde{\boldsymbol{\zeta}}^k, \mathbf{w}\rangle_{[m]+1,[m]}$, the application of the model Hessian can be done by pure matrix calculus. In our implementation we choose

$$\langle \tilde{\boldsymbol{\zeta}}^k, \mathbf{w}\rangle_{[m]+1,[m]} = \langle \tilde{\mathbf{z}}^k, \mathbf{w} \odot \boldsymbol{\omega}\rangle_{[m]+1,[m]} \odot \varphi''(\langle \tilde{\mathbf{y}}^k, \mathbf{w} \odot \boldsymbol{\omega}\rangle_{[m]+1,[m]})$$

for simplicity, i. e., we work with the expected state and adjoint state and reuse the deterministic Hessian computation.

## 6.4. Solution of the Discrete Semismooth Newton System

As pointed out in Section 4.4, a semismooth Newton method can be applied to improve the step found by a linesearch. In this section, we describe how this method is implemented in the discrete setting for the special choice $U_{\mathrm{ad}} := \{u \in U : \mathrm{u_l}(x) \le u(x) \le \mathrm{u_u}(x) \text{ for a. e. } x \in \Omega_u\}$, where $\mathrm{u_l}$, $\mathrm{u_u} \in \mathrm{U} \subset L^\infty(\Omega_u)$ are functions living in the finite-dimensional subspace and fulfilling $\mathrm{u_l}(x) \le \mathrm{u_u}(x)$ for (almost) every $x \in \Omega_u$. This description follows our work [46]. The exact projection onto the feasible set is given by

$$P_{U_{\mathrm{ad}}} : U \to U_{\mathrm{ad}},\ (P_{U_{\mathrm{ad}}}(u))(x) = \min\{\max\{\mathrm{u_l}(x), u(x)\}, \mathrm{u_u}(x)\}. \tag{6.27}$$

It is known [111] that this superposition operator is semismooth as a mapping from $L^q(\Omega_u)$ to $L^2(\Omega_u)$ for any $q > 2$. Since $u^k \in U_{\mathrm{ad}} \subset L^\infty(\Omega_u) \subset L^q(\Omega_u)$ and if $\iota \boldsymbol{B}^*$ maps $\boldsymbol{Y}$ into $L^q(\Omega_u)$ continuously, where $\iota : L^2(\Omega_u)^* \to L^2(\Omega_u)$ denotes the Riesz representation operator, the residual $R$ defined in (4.45) is semismooth w. r. t. the direction $\bar{s}$ as noted in Section 4.4. In our concrete case with $\boldsymbol{B}$ from (3.11), i. e., $\iota \boldsymbol{B}^* \boldsymbol{v} = D^* \int_\Xi \boldsymbol{v}(\cdot, \xi)\, \mathrm{d}\mathbb{P}$, the requirement on $\iota \boldsymbol{B}^*$ is met if $D^* \in \mathcal{L}(L^2(\Omega), L^2(\Omega_u))$ (using Riesz representatives) maps $Y \subset L^2(\Omega)$ continuously into $L^q(\Omega_u)$. If, e. g., $\Omega_u = \Omega$ and $D \equiv I : L^2(\Omega_u) \to L^2(\Omega)$ one can use the Sobolev embedding $H_0^1(\Omega) \hookrightarrow L^p(\Omega)$ and $q \le p$ to show this property.

In the flavor of (6.27), it is desirable to have a componentwise formulation of the projection also on the discrete subspace so that it can be computed efficiently. If we equip $\mathbb{R}^{d_u}$ with the inner product induced by the *lumped* mass matrix $\tilde{\mathsf{M}}_\mathsf{L}$, this holds true because this matrix is diagonal with positive entries. Let $\mathrm{u_l}, \mathrm{u_u}$ be represented by the vectors $\mathsf{u_l}, \mathsf{u_u} \in \mathbb{R}^{d_u}$ which fulfill equivalently $\mathsf{u_l} \le \mathsf{u_u}$ componentwise and define $\mathsf{U}_{\mathrm{ad}} := \{\mathsf{u} \in \mathbb{R}^{d_u} : \mathsf{u_l} \le \mathsf{u} \le \mathsf{u_u}\}$. The projection onto this box w. r. t. the $\tilde{\mathsf{M}}_\mathsf{L}$-inner product then reads

$$\mathsf{P}_{\mathsf{U}_{\mathrm{ad}}} : \mathbb{R}^{d_u} \to \mathsf{U}_{\mathrm{ad}},\ (\mathsf{P}_{\mathsf{U}_{\mathrm{ad}}}(\mathsf{u}))_j = \min\{\max\{(\mathsf{u_l})_j, \mathsf{u}_j\}, (\mathsf{u_u})_j\} \ \text{ for every } j \in [d_u]. \tag{6.28}$$

Given discrete versions $\mathsf{u}^k$ and $\nabla \mathsf{m}_k(0)$ of the current control and the model gradient, respectively, we formulate the discrete version of (4.45) w. r. t. the $\tilde{\mathsf{M}}_\mathsf{L}$-inner product. This means that we use the discrete projection (6.28) and a discrete version $\tilde{\nabla}^2 \mathsf{m}_k(0)$ of the model Hessian w. r. t. the $\tilde{\mathsf{M}}_\mathsf{L}$-inner product, which is obtained by replacing the mass matrix $\tilde{\mathsf{M}}$ by the lumped mass matrix $\tilde{\mathsf{M}}_\mathsf{L}$ in (6.26). This leads to a favorable structure in the semismooth Newton system as we will see later. In analogy to (4.44), we define

$$\hat{\mathsf{H}}(\mathsf{u}^k) := \mathsf{B}^\top (\tilde{\mathsf{A}}_{\text{ref}}^{(k)})^{-1} \Big[ (\mathsf{Q}^\top \mathsf{M}_H \mathsf{Q}) + \mathsf{M}_\mathsf{L} \operatorname{diag}(\langle \tilde{\boldsymbol{\zeta}}^k, \mathbf{w} \rangle_{[m]+1,[m]}) \Big] (\tilde{\mathsf{A}}_{\text{ref}}^{(k)})^{-1} \mathsf{B},$$

where $\tilde{\mathsf{A}}_{\text{ref}}^{(k)}$ and $\tilde{\boldsymbol{\zeta}}^k$ are computed from $\mathsf{u}^k$ as in Section 6.3. Then the alternative Hessian is given by $\tilde{\nabla}^2 \mathsf{m}_k(0)\mathsf{s} = \tilde{\mathsf{M}}_\mathsf{L}^{-1} \hat{\mathsf{H}}(\mathsf{u}^k)\mathsf{s} + \gamma \mathsf{s}$. Inverting the diagonal matrix $\tilde{\mathsf{M}}_\mathsf{L}$ is cheap compared to the inversion of $\tilde{\mathsf{M}}$. With these definitions, (4.45) becomes

$$\mathsf{R}(\bar{\mathsf{s}}) := \bar{\mathsf{s}} + \mathsf{u}^k - \mathsf{P}_{\mathsf{U}_{\text{ad}}} \big( \mathsf{u}^k - \tau_{\text{n},k} \nabla \mathsf{m}_k(0) - \tau_{\text{n},k} \tilde{\mathsf{M}}_\mathsf{L}^{-1} \hat{\mathsf{H}}(\mathsf{u}^k) \bar{\mathsf{s}} \big) = 0. \tag{6.29}$$

with $\tau_{\text{n},k} := (\gamma + \frac{\mathfrak{c}_{\text{n},k}}{\Delta_k})^{-1} > 0$. The discrete residual $\mathsf{R}$ is also semismooth w. r. t. the direction $\mathsf{s}$. We choose

$$\mathrm{D}\mathsf{P}_{\mathsf{U}_{\text{ad}}}(\mathsf{u}) \in \mathbb{R}^{d_u \times d_u} \text{ diagonal}, \quad (\mathrm{D}\mathsf{P}_{\mathsf{U}_{\text{ad}}}(\mathsf{u}))_{jj} = \begin{cases} 0 & \text{if } \mathsf{u}_j < (\mathsf{u}_\text{l})_j, \\ 1 & \text{if } (\mathsf{u}_\text{l})_j \le \mathsf{u}_j \le (\mathsf{u}_\text{u})_j, \\ 0 & \text{if } (\mathsf{u}_\text{u})_j < \mathsf{u}_j \end{cases}$$

as an element of the generalized Jacobian of $\mathsf{P}_{\mathsf{U}_{\text{ad}}}$. Note that if $\mathsf{u}_j \in \{(\mathsf{u}_\text{l})_j, (\mathsf{u}_\text{u})_j\}$, we are in principle free to take $(\mathrm{D}\mathsf{P}_{\mathsf{U}_{\text{ad}}}(\mathsf{u}))_{jj} \in [0,1]$, but choosing this value from $\{0,1\}$ will make a block elimination possible. With this choice, the discrete semismooth Newton system for the iterative solution of (6.29) is

$$(\mathsf{s}^{\ell+1} - \mathsf{s}^\ell) + \tau_{\text{n},k} \, \mathrm{D}\mathsf{P}_{\mathsf{U}_{\text{ad}}} \big( \mathsf{u}^k - \tau_{\text{n},k} \nabla \mathsf{m}_k(0) - \tau_{\text{n},k} \tilde{\mathsf{M}}_\mathsf{L}^{-1} \hat{\mathsf{H}}(\mathsf{u}^k) \mathsf{s}^\ell \big) \, \tilde{\mathsf{M}}_\mathsf{L}^{-1} \hat{\mathsf{H}}(\mathsf{u}^k) (\mathsf{s}^{\ell+1} - \mathsf{s}^\ell) = -\mathsf{R}(\mathsf{s}^\ell), \tag{6.30}$$

where $\mathsf{s}^\ell$ is the current approximation of the solution $\bar{\mathsf{s}}$ of (6.29) and $\mathsf{s}^{\ell+1} - \mathsf{s}^\ell$ is the semismooth Newton step. Since the typically large and dense matrix $\hat{\mathsf{H}}(\mathsf{u}^k)$ should never be formed explicitly, we have to apply an iterative method such as GMRES or CG working with the application of the matrix only to solve (6.30). This method should be formulated on the function space equivalent, i. e., work on $\mathbb{R}^{d_u}$ equipped with the $\tilde{\mathsf{M}}_\mathsf{L}$-inner product. Analogously and in favor of a simple implementation, we can multiply (6.30) by $\tilde{\mathsf{M}}_\mathsf{L}$ from the right and use the fact that diagonal matrices commute to obtain the equivalent, preconditioned system

$$\tilde{\mathsf{M}}_\mathsf{L}(\mathsf{s}^{\ell+1} - \mathsf{s}^\ell) + \tau_{\text{n},k} \, \mathrm{D}\mathsf{P}_{\mathsf{U}_{\text{ad}}} \big( \mathsf{u}^k - \tau_{\text{n},k} \nabla \mathsf{m}_k(0) - \tau_{\text{n},k} \tilde{\mathsf{M}}_\mathsf{L}^{-1} \hat{\mathsf{H}}(\mathsf{u}^k) \mathsf{s}^\ell \big) \, \hat{\mathsf{H}}(\mathsf{u}^k) (\mathsf{s}^{\ell+1} - \mathsf{s}^\ell) = -\tilde{\mathsf{M}}_\mathsf{L} \, \mathsf{R}(\mathsf{s}^\ell), \tag{6.31}$$

which can be solved by iterative methods formulated on $\mathbb{R}^{d_u}$ equipped with the standard Euclidean inner product. In particular, to apply the CG method successfully we symmetrize (6.31) by a block elimination. For this purpose, the index sets

$$\mathcal{I} = \mathcal{I}_k(\mathsf{s}^\ell) := \Big\{ j \in [\tilde{N}] : (\mathrm{D}\mathsf{P}_{\mathsf{U}_{\text{ad}}} \big( \mathsf{u}^k - \tau_{\text{n},k} \nabla \mathsf{m}_k(0) - \tau_{\text{n},k} \tilde{\mathsf{M}}_\mathsf{L}^{-1} \hat{\mathsf{H}}(\mathsf{u}^k) \mathsf{s}^\ell \big))_{jj} = 0 \Big\}$$

and $\mathcal{A} = \mathcal{A}_k(\mathsf{s}^\ell) := [d_u] \setminus \mathcal{I}_k(\mathsf{s}^\ell)$ are distinguished to define the partial solution $(\mathsf{s}^{\ell+1} - \mathsf{s}^\ell)_\mathcal{I} = -\mathsf{R}(\mathsf{s}^\ell)_\mathcal{I}$ of (6.30) or equivalently (6.31). We insert it into (6.31) to obtain the smaller system

$$\big((\tilde{\mathsf{M}}_\mathsf{L})_{\mathcal{A}\mathcal{A}} + \tau_{\mathrm{n},k}\hat{\mathsf{H}}(\mathsf{u}^k)_{\mathcal{A}\mathcal{A}}\big)(\mathsf{s}^{\ell+1} - \mathsf{s}^\ell)_\mathcal{A} = \tau_{\mathrm{n},k}\hat{\mathsf{H}}(\mathsf{u}^k)_{\mathcal{A}\mathcal{I}}\mathsf{R}(\mathsf{s}^\ell)_\mathcal{I} - (\tilde{\mathsf{M}}_\mathsf{L})_{\mathcal{A}\mathcal{A}}\mathsf{R}(\mathsf{s}^\ell)_\mathcal{A}. \tag{6.32}$$

The matrix $(\tilde{\mathsf{M}}_\mathsf{L})_{\mathcal{A}\mathcal{A}} + \tau_{\mathrm{n},k}\hat{\mathsf{H}}(\mathsf{u}^k)_{\mathcal{A}\mathcal{A}}$ is always symmetric w. r. t. the Euclidean inner product and positive definite if e. g., the entries of the vector $\langle\tilde{\boldsymbol{\zeta}}^k, \mathbf{w}\rangle_{[m]+1,[m]}$ are large enough to make the matrix

$$(\mathsf{Q}^\top \mathsf{M}_H \mathsf{Q}) + \mathsf{M}_\mathsf{L}\ \mathrm{diag}(\langle\tilde{\boldsymbol{\zeta}}^k, \mathbf{w}\rangle_{[m]+1,[m]})$$

(see (6.26)) positive definite. In practice, we do not experience a problem when applying CG to the system (6.32). In general, one would have to apply an adequate iterative solver such as MINRES if CG fails to compute an accurate enough solution. The matrix $\hat{\mathsf{H}}(\mathsf{u}^k)_{\mathcal{A}\mathcal{A}}$ should not be computed explicitly, but should be applied to a direction $\mathsf{s}_\mathcal{A}$ using the application of the full operator $\hat{\mathsf{H}}(\mathsf{u}^k)$ as follows:

$$\hat{\mathsf{H}}(\mathsf{u}^k)_{\mathcal{A}\mathcal{A}}\mathsf{s}_\mathcal{A} = \begin{pmatrix}\hat{\mathsf{H}}(\mathsf{u}^k)_{\mathcal{A}\mathcal{A}} & \hat{\mathsf{H}}(\mathsf{u}^k)_{\mathcal{A}\mathcal{I}}\end{pmatrix}\begin{pmatrix}\mathsf{s}_\mathcal{A}\\ 0\end{pmatrix} = \left(\hat{\mathsf{H}}(\mathsf{u}^k)\begin{pmatrix}\mathsf{s}_\mathcal{A}\\ 0\end{pmatrix}\right)_\mathcal{A}.$$

# 7. A Posteriori Error Estimation and Adaptive Solution of a Class of Parametric PDEs Using Low-Rank Tensors

Following [22, 23, 38, 39, 40] we consider the parametric, elliptic, nonlinear operator equation

$$A(\xi)y(\xi) + N(y(\xi)) = b(\xi), \tag{7.1}$$

where $\xi$ is a vector of independent random variables distributed on $\Xi := \bigtimes_{i=1}^{m} \Xi_i$ with the probability measure $\mathbb{P} := \bigotimes_{i=1}^{m} \mathbb{P}_i$ as in Section 6.2. Assumption 6.3 shall still be valid. The right-hand side $b(\xi) \in Y^*$ is given and strongly measurable w. r. t. $\xi$. Let $A(\xi) : Y \to Y^*$ be a linear, self-adjoint, and bounded operator between a Hilbert space and its dual for almost every $\xi$ and let $N : Y \to Y^*$ be well-defined, continuous and monotone, but possibly nonlinear. We assume that $\xi \mapsto A(\xi)$ is also strongly measurable and that $A(\xi)$ is uniformly bounded and boundedly invertible:

$$\|A(\xi)\|_{\mathcal{L}(Y,Y^*)} \le c_{\max}, \quad \|A(\xi)^{-1}\|_{\mathcal{L}(Y^*,Y)} \le c_{\min}^{-1}.$$

This ensures, together with measurability, that $A \in L^\infty_{\mathbb{P}}(\Xi; \mathcal{L}(Y, Y^*))$ and thus $\boldsymbol{A} \in \mathcal{L}(\boldsymbol{Y}, \boldsymbol{Y}^*)$ by Proposition 3.1 for the operator

$$\langle \boldsymbol{A}\boldsymbol{y}, \boldsymbol{v}\rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}} := \int_\Xi \langle A(\xi)\boldsymbol{y}(\cdot,\xi), \boldsymbol{v}(\cdot,\xi)\rangle_{Y^*,Y} \,\mathrm{d}\mathbb{P} \tag{7.2}$$

and the spaces $\boldsymbol{Y} := L^p_{\mathbb{P}}(\Xi; Y)$, $\boldsymbol{Y}^* := L^{p^*}_{\mathbb{P}}(\Xi; Y^*)$ with adequately chosen $p > 3$. The operator $\hat{\boldsymbol{A}} \in \mathcal{L}(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{Y}}^*)$ defined exactly as $\boldsymbol{A}$ but with $\hat{\boldsymbol{Y}} := L^2_{\mathbb{P}}(\Xi; Y)$ and $\hat{\boldsymbol{Y}}^* := L^2_{\mathbb{P}}(\Xi; Y^*)$ is also well-defined and even boundedly invertible.

The ellipticity of the operator $A(\xi)$ gives rise to the inner product

$$(y, v)_{A(\xi)} := \langle A(\xi)y, v\rangle_{Y^*,Y}$$

and the corresponding energy norm $\|y\|_{A(\xi)} = \sqrt{(y,y)_{A(\xi)}}$.

For error estimation, we consider a linear, self-adjoint and elliptic "reference" operator $A_{\mathrm{ref}} \in \mathcal{L}(Y, Y^*)$. This can be the nominal operator $A_{\mathrm{ref}} = A(\bar{\xi})$, where $\bar{\xi} := \int_\Xi \xi \,\mathrm{d}\mathbb{P}$, or any other norm-inducing operator, such as $A_{\mathrm{ref}} = -\Delta$ for the $H^1_0(\Omega)$-norm. Then, $(y, v)_{A_{\mathrm{ref}}} := \langle A_{\mathrm{ref}} y, v\rangle_{Y^*,Y}$ defines an alternative inner product on the space $Y$ and the equivalence estimate

$$\lambda\langle A(\xi)v, v\rangle_{Y^*,Y} \le \langle A_{\mathrm{ref}} v, v\rangle_{Y^*,Y} \le \Lambda\langle A(\xi)v, v\rangle_{Y^*,Y} \tag{7.3}$$

holds for a. e. $\xi \in \Xi$, every $v \in Y$ and some constants $\lambda, \Lambda \in (0, \infty)$, $\lambda \leq \Lambda$, due to the uniform ellipticity of $A(\xi)$. We define the alternative norm $\|y\|_{A_{\mathrm{ref}}} := \sqrt{\langle A_{\mathrm{ref}} y, y\rangle_{Y^*,Y}}$ on $Y$ which is equivalent to the usual norm $\|\cdot\|_Y$. Analogously defining the inner products and norms on $Y^*$ induced by the operators $A(\xi)^{-1}$ and $A_{\mathrm{ref}}^{-1}$, we get the inverse estimates

$$\tfrac{1}{\Lambda}\|\hat{b}\|^2_{A(\xi)^{-1}} \leq \langle \hat{b}, A_{\mathrm{ref}}^{-1}\hat{b}\rangle_{Y^*,Y} = \|\hat{b}\|^2_{A_{\mathrm{ref}}^{-1}} \leq \tfrac{1}{\lambda}\|\hat{b}\|^2_{A(\xi)^{-1}} \tag{7.4}$$

for every $\hat{b} \in Y^*$ by Proposition A.3.

Now let $y(\xi) \in Y$ be the unique and exact solution of (7.1), and let $\tilde{\mathrm{y}}(\xi) \in \mathrm{Y} \subset Y$ be an inexact solution living in a subspace Y of $Y$ (e. g., a finite element subspace), fulfilling

$$A(\xi)\tilde{\mathrm{y}}(\xi) + N(\tilde{\mathrm{y}}(\xi)) - b(\xi) =: r(\xi), \tag{7.5}$$

where $r(\xi) \in Y^*$ is the residual. We assume that this residual can be evaluated exactly in the sense that $\langle r(\xi), \mathrm{v}^+\rangle_{Y^*,Y}$ can be evaluated for every given $\mathrm{v}^+ \in \mathrm{Y}^+ \supset \mathrm{Y}$ in some finite-dimensional (e. g., FE) subspace $\mathrm{Y}^+ \subset Y$.

**Lemma 7.1.** *Let $y(\xi) \in Y$ be the solution of* (7.1) *and let $\tilde{\mathrm{y}}(\xi) \in \mathrm{Y}$. Under the standing assumptions, it holds that*

$$\|\tilde{\mathrm{y}}(\xi) - y(\xi)\|_{A_{\mathrm{ref}}} \leq \Lambda \|r(\xi)\|_{A_{\mathrm{ref}}^{-1}} \tag{7.6}$$

*with $r(\xi)$ defined as in* (7.5).

*Proof.* We can estimate using the monotonicity of $N$:

$$\begin{aligned}
&\|\tilde{\mathrm{y}}(\xi) - y(\xi)\|^2_{A(\xi)} \\
&\leq \langle A(\xi)(\tilde{\mathrm{y}}(\xi) - y(\xi)), \tilde{\mathrm{y}}(\xi) - y(\xi)\rangle_{Y^*,Y} + \langle N(\tilde{\mathrm{y}}(\xi)) - N(y(\xi)), \tilde{\mathrm{y}}(\xi) - y(\xi)\rangle_{Y^*,Y} \\
&= \langle r(\xi), \tilde{\mathrm{y}}(\xi) - y(\xi)\rangle_{Y^*,Y} = \langle A(\xi)A(\xi)^{-1} r(\xi), \tilde{\mathrm{y}}(\xi) - y(\xi)\rangle_{Y^*,Y} \\
&= (A(\xi)^{-1} r(\xi), \tilde{\mathrm{y}}(\xi) - y(\xi))_{A(\xi)} \leq \|A(\xi)^{-1} r(\xi)\|_{A(\xi)} \|\tilde{\mathrm{y}}(\xi) - y(\xi)\|_{A(\xi)}.
\end{aligned}$$

This gives $\|\tilde{\mathrm{y}}(\xi) - y(\xi)\|_{A(\xi)} \leq \|r(\xi)\|_{A(\xi)^{-1}}$ and

$$\tfrac{1}{\sqrt{\Lambda}}\|\tilde{\mathrm{y}}(\xi) - y(\xi)\|_{A_{\mathrm{ref}}} \leq \|\tilde{\mathrm{y}}(\xi) - y(\xi)\|_{A(\xi)} \leq \|r(\xi)\|_{A(\xi)^{-1}} \leq \sqrt{\Lambda}\|r(\xi)\|_{A_{\mathrm{ref}}^{-1}}$$

by the norm estimates (7.3) and (7.4), which yields the result. □

**Remark 7.2.** In the case $N \equiv 0$, we even have $\tilde{\mathrm{y}}(\xi) - y(\xi) = A(\xi)^{-1} r(\xi)$ and thus $\|\tilde{\mathrm{y}}(\xi) - y(\xi)\|_{A(\xi)} = \|r(\xi)\|_{A(\xi)^{-1}}$. Then the lower bound $\sqrt{\lambda}\|r(\xi)\|_{A_0^{-1}} \leq \|r(\xi)\|_{A(\xi)^{-1}}$ is useful.

For the computation of $\|r(\xi)\|_{A_{\mathrm{ref}}^{-1}}$ we define $w(\xi) \in Y$ to be the unique solution of the equation $A_{\mathrm{ref}} w(\xi) = r(\xi)$. Then it holds that

$$\|r(\xi)\|^2_{A_{\mathrm{ref}}^{-1}} = \langle r(\xi), A_{\mathrm{ref}}^{-1} r(\xi)\rangle_{Y^*,Y} = \langle A_{\mathrm{ref}} w(\xi), w(\xi)\rangle_{Y^*,Y} = \|w(\xi)\|^2_{A_{\mathrm{ref}}}. \tag{7.7}$$

We compute a discrete solution $\mathrm{w}(\xi) \in \mathrm{Y}$ fulfilling

$$\langle A_{\mathrm{ref}} \mathrm{w}(\xi), \mathrm{v}\rangle_{Y^*,Y} = \langle r(\xi), \mathrm{v}\rangle_{Y^*,Y} \text{ for all } \mathrm{v} \in \mathrm{Y}, \tag{7.8}$$

i. e., $\mathrm{w}(\xi) = \mathrm{A}_{\mathrm{ref}}^{-1} r(\xi)$, where $\mathrm{A}_{\mathrm{ref}} : \mathrm{Y} \to Y^*$ is the restriction of the operator $A_{\mathrm{ref}}$ onto the space Y and its inverse is defined in the sense of (7.8). We assume that this equation is solved exactly having our application in mind. If this it not the case, the algebraic error caused by the inexact solution of the discrete equation has to be incorporated additionally.

**Lemma 7.3.** *Let $y(\xi), \tilde{\mathrm{y}}(\xi), r(\xi)$ be as in Lemma 7.1, and let $w(\xi) = A_{\mathrm{ref}}^{-1} r(\xi)$ and $\mathrm{w}(\xi) \in \mathrm{Y}$ defined by (7.8). Then, under the standing assumptions,*

$$\|\tilde{\mathrm{y}}(\xi) - y(\xi)\|_{A_{\mathrm{ref}}}^2 \leq \Lambda^2 \big( \|\mathrm{w}(\xi)\|_{A_{\mathrm{ref}}}^2 + \|w(\xi) - \mathrm{w}(\xi)\|_{A_{\mathrm{ref}}}^2 \big) \tag{7.9}$$

*holds.*

*Proof.* Combining (7.6) and (7.7) results in

$$\|\tilde{\mathrm{y}}(\xi) - y(\xi)\|_{A_{\mathrm{ref}}}^2 \leq \Lambda^2 \big( \|\mathrm{w}(\xi) + w(\xi) - \mathrm{w}(\xi)\|_{A_{\mathrm{ref}}}^2 \big) = \Lambda^2 \big( \|\mathrm{w}(\xi)\|_{A_{\mathrm{ref}}}^2 + \|w(\xi) - \mathrm{w}(\xi)\|_{A_{\mathrm{ref}}}^2 \big),$$

where the last equality is due to (7.8) and $\mathrm{w}(\xi) \in \mathrm{Y}$, cf. the proof of [38, Thm. 5.1]. □

The first summand in (7.9) turns out to be the purely algebraic error contribution caused by solving a discretized version of (7.1) inexactly:

$$\begin{aligned} \|\mathrm{w}(\xi)\|_{A_{\mathrm{ref}}}^2 &= \langle r(\xi), \mathrm{A}_{\mathrm{ref}}^{-1} r(\xi) \rangle_{Y^*,Y} \\ &= \langle A(\xi)\tilde{\mathrm{y}}(\xi) + N(\tilde{\mathrm{y}}(\xi)) - b(\xi), \mathrm{A}_{\mathrm{ref}}^{-1} (A(\xi)\tilde{\mathrm{y}}(\xi) + N(\tilde{\mathrm{y}}(\xi)) - b(\xi)) \rangle_{Y^*,Y}. \end{aligned}$$

The second summand will be estimated by standard a posteriori error estimates for (7.8).

## 7.1. Realization of the Deterministic a Posteriori Error Estimator

We discuss the realization of a deterministic a posteriori error estimator for the estimation of the term $\|w(\xi) - \mathrm{w}(\xi)\|_{A_{\mathrm{ref}}}$ for the example from Section 3.2 with the deterministic state equation (3.8) and the adjoint equation (3.16). This is an adaption of the ideas presented in [114, Chap. 1] and [2, Chap. 2] to our setting.

We define in analogy to (3.10), but with slight differences:

$$\begin{aligned} \langle A(\xi)y(\xi), v \rangle_{Y^*,Y} &:= \int_\Omega \kappa(\cdot,\xi) \nabla y(\xi) \cdot \nabla v + \chi(\cdot,\xi) y(\xi) v \,\mathrm{d}x, \\ \langle N(y(\xi)), v \rangle_{Y^*,Y} &:= \int_\Omega \hat{\varphi}(y(\xi)) v \,\mathrm{d}x, \\ \langle b(\xi), v \rangle_{Y^*,Y} &:= \int_\Omega \hat{f}(\xi) v \,\mathrm{d}x, \\ \langle A_{\mathrm{ref}} y(\xi), v \rangle_{Y^*,Y} &:= \int_\Omega \kappa_{\mathrm{ref}} \nabla y(\xi) \cdot \nabla v + \chi_{\mathrm{ref}} y(\xi) v \,\mathrm{d}x \end{aligned} \tag{7.10}$$

with $\chi \in L^\infty_{\lambda \otimes \mathbb{P}}(\Omega \times \Xi)$, $\chi(x,\xi) \geq 0$ for a. e. $(x,\xi) \in \Omega \times \Xi$ and the deterministic reference coefficients $\kappa_{\mathrm{ref}}$ (uniformly positive) and $\chi_{\mathrm{ref}}$ (nonnegative). The functions $\hat{\varphi} \in \mathcal{C}^2(\mathbb{R}, \mathbb{R})$ and $\hat{f}(\xi) \in L^2(\Omega)$ shall fulfill the same properties as $\varphi$ and $f$ in Assumption 3.3.

**Remark 7.4.** The setting covers all important equations and operators:

- For $\chi \equiv 0$, $\hat{\varphi} \equiv \varphi$, and $\hat{f}(\xi) = Du + f(\xi)$ we get the state equation (3.8).
- If, e. g., $Q(\xi) \equiv \iota : H_0^1(\Omega) \hookrightarrow L^2(\Omega)$, we get the adjoint equation (3.16) for $\chi(\cdot, \xi) = \varphi'(\tilde{\mathrm{y}}(\xi))$, $\hat{\varphi} \equiv 0$ and $\hat{f}(\xi) = -(Q(\xi)\tilde{\mathrm{y}}(\xi) - \hat{\mathrm{q}}(\xi))$.
- For $\kappa_{\mathrm{ref}}(x) = \kappa(x, \bar{\xi})$ and $\chi_{\mathrm{ref}} \equiv 0$, we obtain the nominal operator $A(\bar{\xi})$. For $\kappa_{\mathrm{ref}} \equiv 1$ and $\chi_{\mathrm{ref}} \equiv 0$ we get the $H_0^1(\Omega)$-norm inducing operator. Changing to $\chi_{\mathrm{ref}} \equiv 1$, the $H^1(\Omega)$-norm is induced.

**Assumption 7.5.**

- We consider a two-dimensional, polygonal domain $\Omega \subset \mathbb{R}^2$, cf. Assumption 6.1.
- As described in Section 6.1 we discretize the space $Y = H_0^1(\Omega)$ by linear finite elements (continuous on $\overline{\Omega}$) on a triangulation $\mathcal{T}$. The discrete subspace is denoted by $\mathrm{Y} \subset Y$.
- The coefficient functions $\kappa(\cdot, \xi)$ and $\kappa_{\mathrm{ref}}$ are assumed to be piecewise constant on the triangles.

**Remark 7.6.** These assumptions are only made to simplify the explanations and results in this section. Quadrilateral elements, cf. [2], higher order FE functions, or not piecewise constant coefficient functions could be included quite simply, but would result in more complicated formulas or distinctions of cases.

With $v \in Y$ we get

$$\begin{aligned}
&(w(\xi) - \mathrm{w}(\xi), v)_{A_{\mathrm{ref}}} = \langle r(\xi) - A_{\mathrm{ref}}\mathrm{w}(\xi), v\rangle_{Y^*,Y} = \\
&= \sum_{T\in\mathcal{T}} \Big( \int_T \kappa(\cdot,\xi)\nabla\tilde{\mathrm{y}}(\xi)\cdot\nabla v + \chi(\cdot,\xi)\tilde{\mathrm{y}}(\xi)v + \hat{\varphi}(\tilde{\mathrm{y}}(\xi))v - \hat{f}(\xi)v \\
&\qquad\qquad - \kappa_{\mathrm{ref}}\nabla\mathrm{w}(\xi)\cdot\nabla v - \chi_{\mathrm{ref}}\mathrm{w}(\xi)v \,\mathrm{d}x\Big) \\
&= \sum_{T\in\mathcal{T}} \Big( \int_T -\mathrm{div}(\kappa(\cdot,\xi)\nabla\tilde{\mathrm{y}}(\xi))v + \mathrm{div}(\kappa_{\mathrm{ref}}\nabla\mathrm{w}(\xi))v \\
&\qquad\qquad + \big(\chi(\cdot,\xi)\tilde{\mathrm{y}}(\xi) + \hat{\varphi}(\tilde{\mathrm{y}}(\xi)) - \hat{f}(\xi) - \chi_{\mathrm{ref}}\mathrm{w}(\xi)\big)v \,\mathrm{d}x \\
&\qquad\qquad + \int_{\partial T} \big(\kappa(\cdot,\xi)\nabla\tilde{\mathrm{y}}(\xi) - \kappa_{\mathrm{ref}}\nabla\mathrm{w}(\xi)\big)\cdot\nu_T \; v \,\mathrm{d}S\Big),
\end{aligned}$$

where $\nu_T$ is the outer unit normal of the triangle $T$. Since $\kappa(\cdot, \xi)$ and $\kappa_{\mathrm{ref}}$ are piecewise constant on the triangles and we use piecewise linear ansatz functions (Assumption 7.5), $\mathrm{div}(\kappa_{\mathrm{ref}}\nabla\mathrm{w}(\xi)) \equiv 0 \equiv \mathrm{div}(\kappa(\cdot,\xi)\nabla\tilde{\mathrm{y}}(\xi))$ holds true on each element $T$. Therefore, using $(w(\xi) - \mathrm{w}(\xi), \mathrm{v})_{A_{\mathrm{ref}}} = 0$, we get

$$\begin{aligned}
&(w(\xi) - \mathrm{w}(\xi), v)_{A_{\mathrm{ref}}} = (w(\xi) - \mathrm{w}(\xi), v - \mathrm{v})_{A_{\mathrm{ref}}} \\
&= \sum_{T\in\mathcal{T}} \Big( \int_T \big(\chi(\cdot,\xi)\tilde{\mathrm{y}}(\xi) + \hat{\varphi}(\tilde{\mathrm{y}}(\xi)) - \hat{f}(\xi) - \chi_{\mathrm{ref}}\mathrm{w}(\xi)\big)(v - \mathrm{v}) \,\mathrm{d}x \\
&\qquad + \int_{\partial T} \big(\kappa(\cdot,\xi)\tfrac{\partial}{\partial\nu_T}\tilde{\mathrm{y}}(\xi) - \kappa_{\mathrm{ref}}\tfrac{\partial}{\partial\nu_T}\mathrm{w}(\xi)\big)(v - \mathrm{v}) \,\mathrm{d}S\Big)
\end{aligned} \tag{7.11}$$

for arbitrary $\mathrm{v} \in \mathrm{Y}$. Since $\tilde{\mathrm{y}}$ is bounded and continuous on $\overline{\Omega}$ and $\hat{\varphi} : \mathbb{R} \to \mathbb{R}$ is continuous, $\hat{\varphi}(\tilde{\mathrm{y}}(\xi))$ belongs to $L^2(\Omega)$. Due to $\chi(\cdot,\xi), \chi_{\mathrm{ref}} \in L^\infty(\Omega)$, we also have $\chi(\cdot,\xi)\tilde{\mathrm{y}}(\xi), \chi_{\mathrm{ref}}\mathrm{w}(\xi) \in L^2(\Omega)$. The integrals over the triangle boundaries $\partial T$ are considered edge-wise: If an edge $E$ belongs to $\partial\Omega$, this part of the integral vanishes. Interior edges also appear in the integral over the boundary of a neighboring triangle. Thus, the sum over all triangle boundary integrals can be collected to integrals over all interior edges $E \in \mathcal{E}^0$, denoting by $\mathcal{E}$ the set of all edges and by $\mathcal{E}^0$ the set of all interior edges. There, the normal jumps involving the discrete solutions $\tilde{\mathrm{y}}$ and w appear:

$$[\![\phi]\!]_E(x) := \lim_{t\to 0+} \phi(x + t\nu_E) - \lim_{t\to 0+} \phi(x - t\nu_E),$$

where $\nu_E$ is a unit normal vector corresponding to $E$, cf. [114, Sec. 1.1].

We estimate

$$\begin{aligned}(w(\xi) - \mathrm{w}(\xi), v)_{A_{\mathrm{ref}}} \leq & \sum_{T\in\mathcal{T}} \|\chi(\cdot,\xi)\tilde{\mathrm{y}}(\xi) + \hat{\varphi}(\tilde{\mathrm{y}}(\xi)) - \hat{f}(\xi) - \chi_{\mathrm{ref}}\mathrm{w}(\xi)\|_{L^2(T)} \|v - \mathrm{v}\|_{L^2(T)} \\ & + \sum_{E\in\mathcal{E}^0} \|[\![(\kappa(\cdot,\xi)\nabla\tilde{\mathrm{y}} - \kappa_{\mathrm{ref}}\nabla\mathrm{w}(\xi))\cdot\nu_E]\!]_E\|_{L^2(E)} \|v - \mathrm{v}\|_{L^2(E)}\end{aligned} \tag{7.12}$$

If we insert the Clément interpolant $\mathrm{v} \in \mathrm{Y}$ of $v$ and use that $(w - \mathrm{w}, v)_{A_{\mathrm{ref}}} \leq c \cdot \eta(\mathrm{w}) \cdot \|v\|_{A_{\mathrm{ref}}}$ for all $v \in Y$ yields $\|w - \mathrm{w}\|_{A_{\mathrm{ref}}} \leq c \cdot \eta(\mathrm{w})$, a similar estimation as in [114, Sec. 1.2] or [2, Sec. 2.2] can be performed, giving

$$\|w(\xi) - \mathrm{w}(\xi)\|_{A_{\mathrm{ref}}} \leq c_{\mathcal{T}} c_{A_{\mathrm{ref}}} \Big( \sum_{T\in\mathcal{T}} \eta_T(\tilde{\mathrm{y}}(\xi))^2 + \sum_{E\in\mathcal{E}^0} \eta_E(\tilde{\mathrm{y}}(\xi))^2 \Big)^{1/2} \tag{7.13}$$

with

$$\eta_T(\tilde{\mathrm{y}}(\xi)) := h_T \|\chi(\cdot,\xi)\tilde{\mathrm{y}}(\xi) + \hat{\varphi}(\tilde{\mathrm{y}}(\xi)) - \hat{f}(\xi) - \chi_{\mathrm{ref}}\mathrm{w}(\xi)\|_{L^2(T)} \tag{7.14}$$

and

$$\eta_E(\tilde{\mathrm{y}}(\xi)) := h_E^{1/2} \|[\![(\kappa(\cdot,\xi)\nabla\tilde{\mathrm{y}}(\xi) - \kappa_{\mathrm{ref}}\nabla\mathrm{w}(\xi))\cdot\nu_E]\!]_E\|_{L^2(E)}. \tag{7.15}$$

The constant $c_{\mathcal{T}} > 0$ depends only on the smallest angle in the triangulation $\mathcal{T}$ and the coercivity constant $c_{A_{\mathrm{ref}}} > 0$ is chosen such that $\|y\|_{H^1(\Omega)} \leq c_{A_{\mathrm{ref}}} \|y\|_{A_{\mathrm{ref}}}$ holds for all $y \in Y$. The diameters of the triangles and edges are denoted by $h_T$ and $h_E$, respectively.

**Remark 7.7.** This estimate can possibly be refined if the local properties of the interpolation operator are considered more carefully by including the coefficient function. Then the constant $c_{A_{\mathrm{ref}}}$ could be improved.

The overall error estimator looks as follows:

**Theorem 7.8.** *Let $\tilde{\mathrm{y}}(\xi) \in \mathrm{Y}$ be given and let $y(\xi) \in Y$ be the exact solution of* (7.1)*, where the respective operators are defined as in* (7.10)*. Furthermore, let Assumption 7.5 hold and define* $\mathrm{w}(\xi)$ *by* (7.8) *and* (7.5)*.*

*Then,*

$$\|\tilde{y}(\xi) - y(\xi)\|_{A_{\mathrm{ref}}}^2 \leq \Lambda^2 \|\mathrm{w}(\xi)\|_{A_{\mathrm{ref}}}^2 + \Lambda^2 c_{\mathcal{T}}^2 c_{A_{\mathrm{ref}}}^2 \Big( \sum_{T \in \mathcal{T}} \eta_T(\tilde{y}(\xi))^2 + \sum_{E \in \mathcal{E}^0} \eta_E(\tilde{y}(\xi))^2 \Big) \tag{7.16}$$

*holds with* $\Lambda$ *from* (7.3)*,* $c_{\mathcal{T}}$ *and* $c_{A_{\mathrm{ref}}}$ *from* (7.13)*, and* $\eta_T(\tilde{y}(\xi))$ *and* $\eta_E(\tilde{y}(\xi))$ *as in* (7.14)*,* (7.15)*.*

*Proof.* Combining (7.9) and (7.13) yields the desired result. □

## 7.2. A Posteriori Error Estimation in $L^2_{\mathbb{P}}(\Xi)$

In this subsection we discuss an a posteriori error estimator for $Y = \mathbb{R}$ and $\hat{\varphi} \equiv 0$, i. e., a linear equation on $\mathbb{R}$, discretized by a stochastic Galerkin approach. This is only a simple model problem. The essential ideas will be used later to combine both error estimators to one for the PDE with stochastic coefficients.

Let $A(\xi) : \mathbb{R} \to \mathbb{R}$ be a linear, uniformly coercive and bounded operator mapping from $\mathbb{R}$ to $\mathbb{R}$, i. e., we identify $A(\xi) \in [\underline{\kappa}, \overline{\kappa}] \subset \mathbb{R}_{>0}$, and let $b \in L^2_{\mathbb{P}}(\Xi)$. We want to solve $A(\xi)y(\xi) = b(\xi)$ by a stochastic Galerkin method with polynomial chaos. Let $y \in L^2_{\mathbb{P}}(\Xi)$ be the unique weak solution fulfilling

$$(Ay, v)_{L^2_{\mathbb{P}}(\Xi)} = (b, v)_{L^2_{\mathbb{P}}(\Xi)} \quad \text{for all } v \in L^2_{\mathbb{P}}(\Xi),$$

i. e., $y(\xi) = \frac{b(\xi)}{A(\xi)}$ almost surely. We choose polynomials of degree at most $d_i - 1 \in \mathbb{N}_0$ as discretization subspace $\mathcal{P}_{d_i-1}(\Xi_i) \subset L^2_{\mathbb{P}}(\Xi_i)$ and take the full tensor product

$$\mathcal{P}_{d-\mathbb{1}}(\Xi) := \bigotimes_{i=1}^{m} \mathcal{P}_{d_i-1}(\Xi_i) \subset \bigotimes_{i=1}^{m} L^2_{\mathbb{P}_i}(\Xi_i) = L^2_{\mathbb{P}}(\Xi),$$

where $d \in \mathbb{N}^m$ is a vector, cf. Section 6.2. Let $\mathrm{y} \in \mathcal{P}_{d-\mathbb{1}}(\Xi)$ be the Galerkin solution fulfilling

$$(A\mathrm{y}, \mathrm{v})_{L^2_{\mathbb{P}}(\Xi)} = (b, \mathrm{v})_{L^2_{\mathbb{P}}(\Xi)} \quad \text{for all } \mathrm{v} \in \mathcal{P}_{d-\mathbb{1}}(\Xi).$$

Let $\{\beta_{k_i}^{(i)}\}_{k_i=1}^{\infty} \subset L^2_{\mathbb{P}_i}(\Xi_i)$ be sets of orthonormal polynomials w. r. t. the $L^2_{\mathbb{P}_i}(\Xi_i)$ inner product, where $\beta_{k_i}^{(i)}$ has degree $k_i - 1$, and let $\{\beta_k\}_{k \in \mathbb{N}^m}$ be the corresponding orthonormal tensor product polynomials which form a Hilbert basis of $L^2_{\mathbb{P}}(\Xi)$, see Section 6.2. Then, for every $v \in L^2_{\mathbb{P}}(\Xi)$ there exist unique coefficients $(\mu_k)_k \in \ell^2(\mathbb{N}^m)$ such that $v(\xi) = \sum_{k \in \mathbb{N}^m} \mu_k \beta_k(\xi)$ holds almost surely and the series converges in the $L^2_{\mathbb{P}}(\Xi)$ sense.

We have

$$\big(A(y - \mathrm{y}), v\big)_{L^2_{\mathbb{P}}(\Xi)} = \big(b - A\mathrm{y}, \textstyle\sum_{k \in \mathbb{N}^m} \mu_k \beta_k\big)_{L^2_{\mathbb{P}}(\Xi)} = \displaystyle\sum_{k \not\leq d} \mu_k \big(b - A\mathrm{y}, \beta_k\big)_{L^2_{\mathbb{P}}(\Xi)}.$$

The summands for $k \leq d$ (componentwise) cancel out due to Galerkin orthogonality. This corresponds to testing with $\mathrm{v}(\xi) := \sum_{k \leq d} \mu_k \beta_k(\xi)$. Note that we write $k \not\leq d$ and not $k > d$ since we compare index vectors here.

Now we use an approximation $\tilde{A} \in \mathcal{P}_{\tilde{d}}(\Xi)$ of $A$ with $\tilde{d} \in \mathbb{N}_0^m$ and an approximation $\tilde{b} \in \mathcal{P}_{d+\tilde{d}-\mathbb{1}}(\Xi)$ of $b$ and compute

$$\big(b - A\mathrm{y}, v\big)_{L^2_{\mathbb{P}}(\Xi)} = \big(\tilde{b} - \tilde{A}\mathrm{y}, v\big)_{L^2_{\mathbb{P}}(\Xi)} + \big((\tilde{A} - A)\mathrm{y}, v\big)_{L^2_{\mathbb{P}}(\Xi)} + \big(b - \tilde{b}, v\big)_{L^2_{\mathbb{P}}(\Xi)}.$$

Since $\tilde{b} - \tilde{A}\mathrm{y} \in \mathcal{P}_{d+\tilde{d}-\mathbb{1}}(\Xi)$, we have

$$\begin{aligned}\sum_{k \not\leq d} \mu_k \big(\tilde{b} - \tilde{A}\mathrm{y}, \beta_k\big)_{L^2_{\mathbb{P}}(\Xi)} &= \sum_{k \not\leq d}^{k \leq d+\tilde{d}} \mu_k \big(\tilde{b} - \tilde{A}\mathrm{y}, \beta_k\big)_{L^2_{\mathbb{P}}(\Xi)} \\ &\leq \Big(\textstyle\sum_{k \not\leq d}^{k \leq d+\tilde{d}} \mu_k^2\Big)^{1/2} \Big(\sum_{k \not\leq d}^{k \leq d+\tilde{d}} \big(\tilde{b} - \tilde{A}\mathrm{y}, \beta_k\big)^2_{L^2_{\mathbb{P}}(\Xi)}\Big)^{1/2}\end{aligned}$$

due to orthogonality. The sum $\sum_{k \not\leq d}^{k \leq d+\tilde{d}} \big(\tilde{b} - \tilde{A}\mathrm{y}, \beta_k\big)^2_{L^2_{\mathbb{P}}(\Xi)}$ is finite and can be evaluated simply, e. g., by using a quadrature formula of high enough order or analytically. Inserting $v = y - \mathrm{y}$ and using $\|(\mu_k)_k\|_{\ell^2(\mathbb{N})^m} = \|v\|_{L^2_{\mathbb{P}}(\Xi)}$ gives

$$\begin{aligned}&\|y - \mathrm{y}\|_A^2 = \big(\tilde{b} - \tilde{A}\mathrm{y}, y - \mathrm{y}\big)_{L^2_{\mathbb{P}}(\Xi)} + \big((\tilde{A} - A)\mathrm{y}, y - \mathrm{y}\big)_{L^2_{\mathbb{P}}(\Xi)} + \big(b - \tilde{b}, y - \mathrm{y}\big)_{L^2_{\mathbb{P}}(\Xi)} \leq \\ &\Big(\big(\textstyle\sum_{k \not\leq d}^{k \leq d+\tilde{d}} \big(\tilde{b} - \tilde{A}\mathrm{y}, \beta_k\big)^2_{L^2_{\mathbb{P}}(\Xi)}\big)^{1/2} + \|(\tilde{A} - A)\mathrm{y}\|_{L^2_{\mathbb{P}}(\Xi)} + \|b - \tilde{b}\|_{L^2_{\mathbb{P}}(\Xi)}\Big)\|y - \mathrm{y}\|_{L^2_{\mathbb{P}}(\Xi)}.\end{aligned}$$

By $\|v\|_{L^2_{\mathbb{P}}(\Xi)} \leq \frac{1}{\sqrt{\underline{\kappa}}}\|v\|_A$ we get

$$\|y - \mathrm{y}\|_A \leq \frac{1}{\sqrt{\underline{\kappa}}}\Big(\big(\textstyle\sum_{k \not\leq d}^{k \leq d+\tilde{d}} \big(\tilde{b} - \tilde{A}\mathrm{y}, \beta_k\big)^2_{L^2_{\mathbb{P}}(\Xi)}\big)^{1/2} + \|(\tilde{A} - A)\mathrm{y}\|_{L^2_{\mathbb{P}}(\Xi)} + \|b - \tilde{b}\|_{L^2_{\mathbb{P}}(\Xi)}\Big).$$

We see that even for $\tilde{d} = \mathbb{1}$ there are $2^m - 1$ different basis functions $\beta_k$ to test with. Therefore, it is convenient to use further structure of the involved polynomials. If, e. g., $\mathrm{y} \in \mathcal{P}_{d-\mathbb{1}}(\Xi)$, $\tilde{b} \in \mathcal{P}_{d-\mathbb{1}}(\Xi)$ and $\tilde{A}$ is affine, i. e., $\tilde{A}(\xi) = \tilde{A}_0 + \sum_{i=1}^m \xi_i \tilde{A}_i$, we obtain for $k \not\leq d$, $k \leq d + \mathbb{1}$:

$$\begin{aligned}\big(\tilde{b} - \tilde{A}\mathrm{y}, \beta_k\big)_{L^2_{\mathbb{P}}(\Xi)} &= -\big(\sum_{i=1}^m \xi_i \tilde{A}_i \mathrm{y}(\xi), \beta_k\big)_{L^2_{\mathbb{P}}(\Xi)} \\ &= \begin{cases} -\big(\xi_i \tilde{A}_i \mathrm{y}(\xi), \beta_k\big)_{L^2_{\mathbb{P}}(\Xi)} & \text{if } k = d + e_i, \\ 0 & \text{otherwise.}\end{cases}\end{aligned}$$

This gives

$$\sum_{k \not\leq d}^{k \leq d+\mathbb{1}} \big(\tilde{b} - \tilde{A}\mathrm{y}, \beta_k\big)^2_{L^2_{\mathbb{P}}(\Xi)} = \sum_{i=1}^m \big(\xi_i \tilde{A}_i \mathrm{y}(\xi), \beta_{d+e_i}\big)^2_{L^2_{\mathbb{P}}(\Xi)},$$

which can be interpreted as a decomposition of the error into contributions for each parameter.

## 7.3. Combination of Both Error Estimators

We combine the deterministic FE a posteriori error estimator (Section 7.1) with the one for the polynomial discretization of $L^2_{\mathbb{P}}(\Xi)$ (Section 7.2) to obtain an overall error estimator for the solution of the weak formulation of (7.1). For this purpose, we return to the notation from Section 7.1.

We consider the weak formulation of (7.1) w. r. t. the parameters:

$$\langle \boldsymbol{A}\boldsymbol{y} + \boldsymbol{N}(\boldsymbol{y}), \boldsymbol{v}\rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}} = \langle \boldsymbol{b}, \boldsymbol{v}\rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}} \quad \text{for all } \boldsymbol{v} \in \boldsymbol{Y}. \tag{7.17}$$

The operator $\boldsymbol{A}$ is defined in (7.2). Analogously, we have

$$\langle \boldsymbol{N}(\boldsymbol{y}), \boldsymbol{v}\rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}} := \int_\Xi \langle N(\boldsymbol{y}(\cdot,\xi)), \boldsymbol{v}(\cdot,\xi)\rangle_{Y^*,Y} \,\mathrm{d}\mathbb{P} \quad \text{and}$$

$$\langle \boldsymbol{b}, \boldsymbol{v}\rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}} := \int_\Xi \langle b(\xi), \boldsymbol{v}(\cdot,\xi)\rangle_{Y^*,Y} \,\mathrm{d}\mathbb{P} \quad \text{for all } \boldsymbol{v} \in \boldsymbol{Y}.$$

The exact solution $\boldsymbol{y}$ of (7.17) is contained in $\boldsymbol{Y} = L^p_{\mathbb{P}}(\Xi; Y)$ with $p \in (3, \infty)$, which we require for the weak formulation to be well-defined and for the nonlinear operator $\boldsymbol{N}$ to be twice continuously differentiable, but we will provide only error estimation in $\hat{\boldsymbol{Y}} = L^2_{\mathbb{P}}(\Xi; Y)$, because we want to make use of (Galerkin) orthogonality.

Analogously to (7.2) we define the reference operator $\boldsymbol{A}_{\mathrm{ref}} \in \mathcal{L}(\boldsymbol{Y}, \boldsymbol{Y}^*)$ by

$$\langle \boldsymbol{A}_{\mathrm{ref}}\boldsymbol{y}, \boldsymbol{v}\rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}} := \int_\Xi \langle A_{\mathrm{ref}}\boldsymbol{y}(\cdot,\xi), \boldsymbol{v}(\cdot,\xi)\rangle_{Y^*,Y} \,\mathrm{d}\mathbb{P}$$

and its version $\hat{\boldsymbol{A}}_{\mathrm{ref}} \in \mathcal{L}(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{Y}}^*)$. The operators $\hat{\boldsymbol{A}}$, $\hat{\boldsymbol{A}}_{\mathrm{ref}}$, and their inverses induce inner products and respective norms on $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{Y}}^*$, respectively. The estimates (7.3) and (7.4) carry over to this setting.

For every $\tilde{\mathbf{y}} \in \mathbf{Y} \subset \boldsymbol{Y}$ we can conclude

$$\|\tilde{\mathbf{y}} - \boldsymbol{y}\|_{\hat{\boldsymbol{A}}_{\mathrm{ref}}} \leq \Lambda \|\boldsymbol{r}\|_{\hat{\boldsymbol{A}}^{-1}_{\mathrm{ref}}} \tag{7.18}$$

in analogy to Lemma 7.1 with $\boldsymbol{r} = \boldsymbol{A}\tilde{\mathbf{y}} + \boldsymbol{N}(\tilde{\mathbf{y}}) - \boldsymbol{b}$. Note that estimating the error in $L^2$ enables us to use the coercivity of the operators $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{A}}_{\mathrm{ref}}$. We need $\boldsymbol{r} \in L^2_{\mathbb{P}}(\Xi; Y^*)$, which can be concluded using integrability properties of $\tilde{\mathbf{y}}$ and the regularity of the data, see Sections 3.2 and 3.4, and thus $\boldsymbol{w} = \hat{\boldsymbol{A}}^{-1}_{\mathrm{ref}}\boldsymbol{r} \in \hat{\boldsymbol{Y}}$. For the estimation of the total error we compute $\|\mathbf{w}\|^2_{\hat{\boldsymbol{A}}_{\mathrm{ref}}}$ and have to estimate $\|\boldsymbol{w} - \mathbf{w}\|^2_{\hat{\boldsymbol{A}}_{\mathrm{ref}}}$, cf. Lemma 7.3. Thus, it makes sense to minimize the error term $\|\mathbf{w}\|^2_{\hat{\boldsymbol{A}}_{\mathrm{ref}}}$ by an iterative solver used for the discretized equation. We have that $\mathbf{w} \in \mathrm{Y} \otimes \mathcal{P}_{d-\mathbb{1}}(\Xi) \subset L^2_{\mathbb{P}}(\Xi; Y)$ solves

$$\langle \hat{\boldsymbol{A}}_{\mathrm{ref}}\mathbf{w}, \mathbf{v}\rangle_{\hat{\boldsymbol{Y}}^*,\hat{\boldsymbol{Y}}} = \langle \boldsymbol{r}, \mathbf{v}\rangle_{\hat{\boldsymbol{Y}}^*,\hat{\boldsymbol{Y}}} \quad \text{for all } \mathbf{v} \in \mathrm{Y} \otimes \mathcal{P}_{d-\mathbb{1}}(\Xi). \tag{7.19}$$

We write an arbitrary test function in the form $\boldsymbol{v}(x,\xi) = \sum_{k\in\mathbb{N}^m} v^*_k(x)\vartheta_k(\xi)$ with some

Hilbert basis $\{\vartheta_k\}_{k\in\mathbb{N}^m} \subset L^2_{\mathbb{P}}(\Xi)$ and $v^*_k \in Y$ for all $k$. We have

$$\begin{aligned}
&\langle \hat{\boldsymbol{A}}_{\mathrm{ref}}(\boldsymbol{w}-\mathbf{w}), \boldsymbol{v}\rangle_{L^2_{\mathbb{P}}(\Xi;Y^*),L^2_{\mathbb{P}}(\Xi;Y)} \\
&= \langle \boldsymbol{r} - \hat{\boldsymbol{A}}_{\mathrm{ref}}\mathbf{w}, \textstyle\sum_{k\in\mathbb{N}^m} v^*_k\vartheta_k\rangle_{\hat{\boldsymbol{Y}}^*,\hat{\boldsymbol{Y}}} \\
&= \sum_{k\in\mathbb{N}^m} \langle \boldsymbol{r} - \hat{\boldsymbol{A}}_{\mathrm{ref}}\mathbf{w}, v^*_k\vartheta_k\rangle_{\hat{\boldsymbol{Y}}^*,\hat{\boldsymbol{Y}}} \\
&= \sum_{k\le d} \langle \boldsymbol{r} - \hat{\boldsymbol{A}}_{\mathrm{ref}}\mathbf{w}, (v^*_k - \mathrm{v}^*_k)\vartheta_k\rangle_{\hat{\boldsymbol{Y}}^*,\hat{\boldsymbol{Y}}} + \sum_{k\not\le d} \langle \boldsymbol{r} - \hat{\boldsymbol{A}}_{\mathrm{ref}}\mathbf{w}, v^*_k\vartheta_k\rangle_{\hat{\boldsymbol{Y}}^*,\hat{\boldsymbol{Y}}}
\end{aligned} \tag{7.20}$$

with $\mathrm{v}^*_k \in \mathrm{Y}$.

**Assumption 7.9.** We need the following prerequisites in addition to Assumption 7.5:

- The coefficient function is affine in $\xi$, i. e., $\kappa(x,\xi) = \kappa_0(x) + \sum_{i=1}^m \xi_i\kappa_i(x)$. This form can originate in a truncated Karhunen-Loève expansion. The functions $\kappa_i$ ($i \in \{0,\dots,m\}$) are assumed to be piecewise constant on the triangles. We define $\langle A_i y, v\rangle_{Y^*,Y} := \int_\Omega \kappa_i \nabla y \cdot \nabla v \,\mathrm{d}x$ for $i \in \{0,\dots,m\}$.
- Let $\tilde{\mathbf{y}} \in \mathrm{Y} \otimes \mathcal{P}_{d-\mathbb{1}}(\Xi)$. We assume that the nonlinear part $\hat{\varphi}(\tilde{\mathbf{y}})$ of the residual can be approximated sufficiently well by a function $\tilde{\boldsymbol{\varphi}} \in L^2(\Omega)\otimes\mathcal{P}_{d-\mathbb{1}}(\Xi)$, e. g., by interpolation. Moreover, we assume that the function $\chi\cdot\tilde{\mathbf{y}}$ can be approximated sufficiently well by a function $\tilde{\boldsymbol{\chi}} \in L^2(\Omega)\otimes\mathcal{P}_{d-\mathbb{1}}(\Xi)$.
- The right-hand side $\boldsymbol{b}$ can be identified with a function $\hat{\boldsymbol{f}} \in L^2(\Omega)\otimes\mathcal{P}_{d-\mathbb{1}}(\Xi)$.

Again, one can include the interpolation error as follows:

$$\begin{aligned}
&\langle \boldsymbol{r} - \hat{\boldsymbol{A}}_{\mathrm{ref}}\mathbf{w}, \boldsymbol{v}\rangle_{\hat{\boldsymbol{Y}}^*,\hat{\boldsymbol{Y}}} = \\
&\int_\Xi\int_\Omega \kappa\,\nabla_x\tilde{\mathbf{y}}\cdot\nabla_x\boldsymbol{v} + \chi\cdot\tilde{\mathbf{y}}\cdot\boldsymbol{v} + \varphi(\tilde{\mathbf{y}})\cdot\boldsymbol{v} - \hat{\boldsymbol{f}}\cdot\boldsymbol{v} - \kappa_{\mathrm{ref}}\,\nabla_x\tilde{\mathbf{w}}\cdot\nabla_x\boldsymbol{v}\,\mathrm{d}x\,\mathrm{d}\mathbb{P} = \\
&\int_\Xi\int_\Omega \kappa\,\nabla_x\tilde{\mathbf{y}}\cdot\nabla_x\boldsymbol{v} + \tilde{\boldsymbol{\chi}}\cdot\boldsymbol{v} + \tilde{\boldsymbol{\varphi}}\cdot\boldsymbol{v} - \hat{\boldsymbol{f}}\cdot\boldsymbol{v} - \kappa_{\mathrm{ref}}\,\nabla_x\tilde{\mathbf{w}}\cdot\nabla_x\boldsymbol{v}\,\mathrm{d}x\,\mathrm{d}\mathbb{P} \\
&+ \int_\Xi\int_\Omega (\chi\cdot\tilde{\mathbf{y}} - \tilde{\boldsymbol{\chi}} + \varphi(\tilde{\mathbf{y}}) - \tilde{\boldsymbol{\varphi}})\cdot\boldsymbol{v}\,\mathrm{d}x\,\mathrm{d}\mathbb{P}.
\end{aligned} \tag{7.21}$$

In the following, we will neglect $\chi\cdot\tilde{\mathbf{y}} - \tilde{\boldsymbol{\chi}}$ and $\varphi(\tilde{\mathbf{y}}) - \tilde{\boldsymbol{\varphi}}$ for the ease of presentation and implementation. Moreover, we allow for different orthonormal polynomial bases in the derivation of the error estimates.

### Conversion between Polynomial Basis Representations

To discretize the space $L^2_{\mathbb{P}_i}(\Xi_i)$, we choose orthonormal polynomials in tensor product form, as already described in Section 6.2 and used in Section 7.2. Typically, the orthonormal basis $\{\beta^{(i)}_{k_i}\}_{k_i=1}^\infty$, where $\beta^{(i)}_{k_i}$ has degree $k_i - 1$, and the tensor product basis $\{\beta_k\}_{k\in\mathbb{N}^m}$ of $L^2_{\mathbb{P}}(\Xi)$ is used. Alternatively, we have the Hilbert basis $\{\theta_k\}_{k\in\mathbb{N}^m}$, which results from replacing the

first $d_i$ univariate polynomials in $\{\beta_{k_i}^{(i)}\}_{k_i=1}^{\infty}$ by weighted, orthonormal Lagrange polynomials w. r. t. the Gaussian quadrature nodes.

Now consider a function $\mathbf{w} \in \mathrm{Y} \otimes \mathcal{P}_d(\Xi)$ written as

$$\mathbf{w}(x,\xi) = \sum_{k \le d} \mathrm{w}_k^*(x)\beta_k(\xi) = \sum_{k \le d} \mathrm{w}_k(x)\theta_k(\xi), \tag{7.22}$$

where $\mathrm{w}_k^*, \mathrm{w}_k \in \mathrm{Y}$ are the coefficients. On the one hand, inserting $\xi = \mathsf{a}_l$ (Gaussian nodes) yields

$$\mathrm{w}_l(x) = \omega_l^{-1} \cdot \sum_{k \le d} \mathrm{w}_k^*(x)\beta_k(\mathsf{a}_l).$$

On the other hand, testing (7.22) with $\beta_l$ gives

$$\mathrm{w}_l^*(x) = \sum_{k \le d} \Big(\mathrm{w}_k(x)\big(\theta_k, \beta_l\big)_{L^2_{\mathbb{P}}(\Xi)}\Big) = \sum_{k \le d} \Big(\mathrm{w}_k(x) \prod_{i=1}^{m} \big(\theta_{k_i}^{(i)}, \beta_{l_i}^{(i)}\big)_{L^2_{\mathbb{P}_i}(\Xi_i)}\Big).$$

Defining the orthonormal matrices $\mathsf{Q}^{(i)} \in \mathbb{R}^{d_i+1 \times d_i+1}$ by $\mathsf{Q}^{(i)}_{k_i l_i} = \big(\theta_{k_i}^{(i)}, \beta_{l_i}^{(i)}\big)_{L^2_{\mathbb{P}_i}(\Xi_i)}$, we see that the coefficients $\mathrm{w}_l^*$ $(l \le d)$ of one orthonormal basis in tensor product form can be computed by applying an orthonormal rank-1-operator $\mathsf{Q}^{(1)} \otimes \cdots \otimes \mathsf{Q}^{(m)}$ to the coefficients $\mathrm{w}_k$ $(k \le d)$ of another orthonormal tensor product basis.

### Discretization with Lagrange Polynomials

For a simpler approximation of the nonlinearity, it is beneficial to use the Lagrange basis $\{\theta_k\}_{k \in \mathbb{N}_0^m}$, as seen in Section 6.2. Still, we can convert the coefficients to representations in other bases if necessary. Furthermore, many results in this subsection still hold for arbitrary orthonormal polynomials. We will point out where we use special properties of the chosen Lagrange basis.

The polynomial basis representations of $\tilde{\mathbf{y}}$ and $\mathbf{w}$ shall be

$$\tilde{\mathbf{y}}(x,\xi) = \sum_{l \le d} \tilde{\mathrm{y}}_l(x)\theta_l(\xi), \quad \mathbf{w}(x,\xi) = \sum_{l \le d} \mathrm{w}_l(x)\theta_l(\xi)$$

with $\tilde{\mathrm{y}}_l, \mathrm{w}_l \in \mathrm{Y}$. Be aware of the fact that these sums are actually large since $l$ is an index vector. Analogously, we write $\hat{\boldsymbol{f}}$, $\tilde{\boldsymbol{\varphi}}$, and $\tilde{\boldsymbol{\chi}}$ with the deterministic functions $f_l, \tilde{\varphi}_l, \tilde{\chi}_l \in L^2(\Omega)$ and an arbitrary test function $\boldsymbol{v} \in \hat{\boldsymbol{Y}}$ as $\boldsymbol{v}(x,\xi) = \sum_{k \in \mathbb{N}^m} v_k(x)\theta_k(\xi)$.

Moreover, we will use transformed representations $\tilde{\mathbf{y}}(x,\xi) = \sum_{l \le d} \tilde{\mathrm{y}}_l^*(x)\vartheta_l(\xi)$ as well as $\mathbf{v}(x,\xi) = \sum_{k \in \mathbb{N}^m} v_k^*(x)\vartheta_k(\xi)$, where $\{\vartheta_{k_i}^{(i)}\}_{k_i=1}^{\infty}$ are alternative, still to be specified orthonormal bases of $L^2_{\mathbb{P}_i}(\Xi_i)$ with $\vartheta_{k_i}^{(i)} \equiv \theta_{k_i}^{(i)} \equiv \beta_{k_i}^{(i)}$ for $k_i \ge d_i + 1$ and $\tilde{\mathrm{y}}_l^* \in \mathrm{Y}$, $v_k^* \in Y$. These coefficients can simply be computed from $\tilde{\mathrm{y}}_l$ and $v_k$ by an orthonormal transformation. Especially for $\boldsymbol{v}$, we split up the representation into

$$\boldsymbol{v}(x,\xi) = \sum_{k \le d} v_k(x)\theta_k(\xi) + \sum_{k \not\le d} v_k^*(x)\vartheta_k(\xi).$$

For $k \leq d$, we obtain

$$
\begin{aligned}
&\langle \boldsymbol{r} - \hat{\boldsymbol{A}}_{\mathrm{ref}}\mathbf{w}, v_k\theta_k\rangle_{L^2_{\mathbb{P}}(\Xi;Y^*),L^2_{\mathbb{P}}(\Xi;Y)} \\
&= \int_\Xi \langle A(\xi)\tilde{\mathbf{y}}(\cdot,\xi) + N(\tilde{\mathbf{y}}(\cdot,\xi)) - b(\xi) - A_{\mathrm{ref}}\mathbf{w}(\cdot,\xi), v_k\rangle_{Y^*,Y}\,\theta_k(\xi)\,\mathrm{d}\mathbb{P} \\
&\approx \sum_{l\leq d}\int_\Xi \Big(\int_\Omega (\kappa_0 + \xi_1\kappa_1 + \ldots + \xi_m\kappa_m)\nabla\tilde{\mathrm{y}}_l\cdot\nabla v_k + \tilde{\chi}_l v_k + \tilde{\varphi}_l v_k \\
&\qquad\qquad - f_l v_k - \kappa_{\mathrm{ref}}\nabla \mathrm{w}_l\cdot\nabla v_k - \chi_{\mathrm{ref}}\mathrm{w}_l v_k\,\mathrm{d}x\Big)\theta_l(\xi)\theta_k(\xi)\,\mathrm{d}\mathbb{P} \\
&= \sum_{l\leq d}\Big(\int_\Omega \kappa_0\nabla\tilde{\mathrm{y}}_l\cdot\nabla v_k + \tilde{\chi}_l v_k + \tilde{\varphi}_l v_k \\
&\qquad\qquad - f_l v_k - \kappa_{\mathrm{ref}}\nabla \mathrm{w}_l\cdot\nabla v_k - \chi_{\mathrm{ref}}\mathrm{w}_l v_k\,\mathrm{d}x\Big)\Big(\int_\Xi \theta_l(\xi)\theta_k(\xi)\,\mathrm{d}\mathbb{P}\Big) \\
&\quad + \sum_{l\leq d}\sum_{i=1}^m \Big(\int_\Omega \kappa_i\nabla\tilde{\mathrm{y}}_l\cdot\nabla v_k\,\mathrm{d}x\Big)\Big(\int_\Xi \xi_i\theta_l(\xi)\theta_k(\xi)\,\mathrm{d}\mathbb{P}\Big) \\
&= \int_\Omega (\kappa_0 + \mathsf{a}^{(1)}_{k_1}\kappa_1 + \ldots + \mathsf{a}^{(m)}_{k_m}\kappa_m)\nabla\tilde{\mathrm{y}}_k\cdot\nabla v_k + \tilde{\chi}_k v_k \\
&\qquad + \tilde{\varphi}_k v_k - f_k v_k - \kappa_{\mathrm{ref}}\nabla \mathrm{w}_k\cdot\nabla v_k - \chi_{\mathrm{ref}}\mathrm{w}_k v_k\,\mathrm{d}x.
\end{aligned}
\tag{7.23}
$$

This is due to orthogonality of the polynomials $\{\theta_k\}_{k\in\mathbb{N}^m}$ and the fact

$$
\begin{aligned}
\int_\Xi \xi_i\theta_l(\xi)\theta_k(\xi)\,\mathrm{d}\mathbb{P} &= \int_\Xi \xi_i \prod_{j=1}^m \theta^{(j)}_{l_j}(\xi_j)\theta^{(j)}_{k_j}(\xi_j)\,\mathrm{d}\mathbb{P} \\
&= \int_{\Xi_i} \xi_i\theta^{(i)}_{l_i}(\xi_i)\theta^{(i)}_{k_i}(\xi_i)\,\mathrm{d}\mathbb{P}_i \cdot \prod_{j=1,j\neq i}^m \delta_{l_j k_j} \\
&= \begin{cases} \mathsf{a}^{(i)}_{k_i} & \text{if } k = l, \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
$$

Recall that $\{\mathsf{a}^{(i)}_{k_i}\}_{k_i=1}^{d_i} \subset \Xi_i$ are the Gaussian quadrature nodes w. r. t. $\mathbb{P}_i$. For a different polynomial basis we would obtain different values here and possibly no decoupling of the deterministic equations which we have to solve.

For $k \not\leq d$ (yielding $k \neq l$) we get

$$
\begin{aligned}
&\langle \boldsymbol{r} - \hat{\boldsymbol{A}}_{\mathrm{ref}}\mathbf{w}, v_k^*\vartheta_k\rangle_{L^2_{\mathbb{P}}(\Xi;Y^*),L^2_{\mathbb{P}}(\Xi;Y)} \\
&\approx \sum_{l\leq d}\sum_{i=1}^m \Big(\int_\Omega \kappa_i\nabla\tilde{\mathrm{y}}_l^*\cdot\nabla v_k^*\,\mathrm{d}x\Big)\Big(\int_\Xi \xi_i\vartheta_l(\xi)\vartheta_k(\xi)\,\mathrm{d}\mathbb{P}\Big) \\
&= \begin{cases} \sum_{l_i=1}^{d_i} \mathsf{c}^{(i)}_{l_i}\int_\Omega \kappa_i\nabla\tilde{\mathrm{y}}^*_{k_1,\ldots,l_i,\ldots,k_m}\cdot\nabla v_k^*\,\mathrm{d}x & \text{if } \begin{cases} k_i = d_i + 1 \text{ for an } i\in[m], \\ k_j \leq d_j \text{ for } j\in[m]\setminus\{i\}, \end{cases} \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
\tag{7.24}
$$

In this computation we used the following properties of any orthonormal polynomial basis $\{\vartheta_l\}_{l\in\mathbb{N}^m}$ in tensor product from, where $\vartheta_{l_i}^{(i)} \perp \mathcal{P}_{l_i-2}(\Xi_i)$ for all $l_i \geq d_i + 1$:

$$\begin{aligned}\int_\Xi \xi_i \vartheta_l(\xi)\vartheta_k(\xi)\,\mathrm{d}\mathbb{P} &= \int_{\Xi_i} \xi_i \vartheta_{l_i}^{(i)}(\xi_i)\vartheta_{k_i}^{(i)}(\xi_i)\,\mathrm{d}\mathbb{P}_i \cdot \prod_{j=1,j\neq i}^{m} \delta_{l_j k_j} \\ &= \begin{cases} \mathsf{c}_{l_i}^{(i)} & \text{if } k_i = d_i + 1 \text{ and } k_j = l_j \text{ for all } j \in [m]\setminus\{i\} \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

with $\mathsf{c}_{l_i}^{(i)} := \int_{\Xi_i} \xi_i \vartheta_{l_i}^{(i)}(\xi_i)\vartheta_{d_i+1}^{(i)}(\xi_i)\,\mathrm{d}\mathbb{P}_i$ holds for any $l \leq d$ and $k \not\leq d$.

**Example 7.10.** For $\vartheta_{l_i}^{(i)} \equiv \beta_{l_i}^{(i)}$ it even holds $\mathsf{c}_{l_i}^{(i)} = 0$ for all $l_i \leq d_i - 1$. Then we get

$$\sum_{l_i=1}^{d_i} \mathsf{c}_{l_i}^{(i)} \int_\Omega \kappa_i \nabla \tilde{\mathrm{y}}^*_{k_1,\dots,l_i,\dots,k_m} \cdot \nabla v_k^* \,\mathrm{d}x = \mathsf{c}_{d_i}^{(i)} \int_\Omega \kappa_i \nabla \tilde{\mathrm{y}}^*_{k_1,\dots,d_i,\dots,k_m} \cdot \nabla v_k^* \,\mathrm{d}x$$

in (7.24).

From (7.20), (7.23), and (7.24) we get

$$\begin{aligned}&\langle \hat{\boldsymbol{A}}_{\mathrm{ref}}(\boldsymbol{w}-\mathbf{w}), \boldsymbol{v}\rangle_{\hat{\boldsymbol{Y}}^*,\hat{\boldsymbol{Y}}} \\ &= \sum_{k\leq d} \langle \boldsymbol{r} - \hat{\boldsymbol{A}}_{\mathrm{ref}}\mathbf{w}, (v_k - \mathrm{v}_k)\theta_k\rangle_{\hat{\boldsymbol{Y}}^*,\hat{\boldsymbol{Y}}} + \sum_{k\not\leq d} \langle \boldsymbol{r} - \hat{\boldsymbol{A}}_{\mathrm{ref}}\mathbf{w}, v_k^*\vartheta_k\rangle_{\hat{\boldsymbol{Y}}^*,\hat{\boldsymbol{Y}}} \\ &\approx \sum_{k\leq d} \int_\Omega \big(\kappa(\cdot,\mathsf{a}_k)\nabla\tilde{\mathrm{y}}_k - \kappa_{\mathrm{ref}}\nabla \mathrm{w}_k\big)\cdot\nabla(v_k - \mathrm{v}_k) + \big(\tilde{\chi}_k + \tilde{\varphi}_k - f_k - \chi_{\mathrm{ref}}\mathrm{w}_k\big)(v_k - \mathrm{v}_k)\,\mathrm{d}x \\ &\qquad + \sum_{i=1}^{m}\sum_{k\leq d} \mathsf{c}_{k_i}^{(i)} \int_\Omega \kappa_i \nabla\tilde{\mathrm{y}}_k^* \cdot \nabla v^*_{k_1,\dots,d_i+1,\dots,d_m}\,\mathrm{d}x. \end{aligned} \tag{7.25}$$

Recall that we write $\mathsf{a}_k := \big(\mathsf{a}_{k_1}^{(1)},\dots,\mathsf{a}_{k_m}^{(m)}\big)^\top$.

We see that the first term at the end of (7.25) can be treated as in Section 7.1 and constitutes the deterministic part of the error. The second term can be viewed as the stochastic part of the error, where the $i$-th summand corresponds to the discretization error of $L^2_{\mathbb{P}_i}(\Xi_i)$.

The first term in (7.25) is estimated as in (7.11), (7.12) and is bounded as in (7.13) using the Clément interpolant $\mathrm{v}_k$ of $v_k$:

$$\begin{aligned}&\sum_{k\leq d} \int_\Omega \big(\kappa(\cdot,\mathsf{a}_k)\nabla\tilde{\mathrm{y}}_k - \kappa_{\mathrm{ref}}\nabla \mathrm{w}_k\big)\cdot\nabla(v_k - \mathrm{v}_k) + \big(\tilde{\chi}_k + \tilde{\varphi}_k - f_k - \chi_{\mathrm{ref}}\mathrm{w}_k\big)(v_k - \mathrm{v}_k)\,\mathrm{d}x \\ &= \sum_{k\leq d}\sum_{T\in\mathcal{T}} \Big( \int_T \big(\tilde{\chi}_k + \tilde{\varphi}_k - f_k - \chi_{\mathrm{ref}}\mathrm{w}_k\big)(v_k - \mathrm{v}_k)\,\mathrm{d}x \\ &\qquad + \int_{\partial T} \big(\kappa(\cdot,\mathsf{a}_k)\tfrac{\partial}{\partial\nu_T}\tilde{\mathrm{y}}_k - \kappa_{\mathrm{ref}}\tfrac{\partial}{\partial\nu_T}\mathrm{w}_k\big)(v_k - \mathrm{v}_k)\,\mathrm{d}S\Big)\end{aligned}$$

$$
\begin{aligned}
&\le \sum_{T\in\mathcal{T}}\sum_{k\le d}\|\tilde\chi_k+\tilde\varphi_k-f_k-\chi_{\mathrm{ref}}\mathrm{w}_k\|_{L^2(T)}\|v_k-\mathrm{v}_k\|_{L^2(T)}\\
&\quad+\sum_{E\in\mathcal{E}^0}\sum_{k\le d}\|[\![(\kappa(\cdot,\mathsf{a}_k)\nabla\tilde{\mathrm{y}}_k-\kappa_{\mathrm{ref}}\nabla\mathrm{w}_k)\cdot\nu_E]\!]_E\|_{L^2(E)}\|v_k-\mathrm{v}_k\|_{L^2(E)}\\
&\le c_{\mathcal{T}}c_{A_{\mathrm{ref}}}\Big(\sum_{T\in\mathcal{T}}h_T^2\sum_{k\le d}\|\tilde\chi_k+\tilde\varphi_k-f_k-\chi_{\mathrm{ref}}\mathrm{w}_k\|^2_{L^2(T)}\\
&\quad+\sum_{E\in\mathcal{E}^0}h_E\sum_{k\le d}\|[\![(\kappa(\cdot,\mathsf{a}_k)\nabla\tilde{\mathrm{y}}_k-\kappa_{\mathrm{ref}}\nabla\mathrm{w}_k)\cdot\nu_E]\!]_E\|^2_{L^2(E)}\Big)^{\frac12}\Big(\sum_{k\le d}\|v_k\|^2_{A_{\mathrm{ref}}}\Big)^{\frac12}\\
&= c_{\mathcal{T}}c_{A_{\mathrm{ref}}}\Big(\sum_{T\in\mathcal{T}}\boldsymbol{\eta}_T(\tilde{\mathbf{y}})^2+\sum_{E\in\mathcal{E}^0}\boldsymbol{\eta}_E(\tilde{\mathbf{y}})^2\Big)^{\frac12}\Big(\sum_{k\le d}\|v_k\|^2_{A_{\mathrm{ref}}}\Big)^{\frac12}
\end{aligned}
\tag{7.26}
$$

with

$$
\boldsymbol{\eta}_T(\tilde{\mathbf{y}}):=h_T\Big(\sum_{k\le d}\|\tilde\chi_k+\tilde\varphi_k-f_k-\chi_{\mathrm{ref}}\mathrm{w}_k\|^2_{L^2(T)}\Big)^{\frac12}
\tag{7.27}
$$

and

$$
\boldsymbol{\eta}_E(\tilde{\mathbf{y}}):=h_E^{1/2}\Big(\sum_{k\le d}\|[\![(\kappa(\cdot,\mathsf{a}_k)\nabla\tilde{\mathrm{y}}_k-\kappa_{\mathrm{ref}}\nabla\mathrm{w}_k)\cdot\nu_E]\!]_E\|^2_{L^2(E)}\Big)^{\frac12}.
\tag{7.28}
$$

Note that $\tilde{\mathbf{y}}$ enters in these definitions via $\mathrm{w}_k$ and $\tilde{\mathrm{y}}_k$, and indirectly via $\tilde\chi_k$ and $\tilde\varphi_k$.

The second term in (7.25) can be treated as follows:

$$
\begin{aligned}
&\sum_{i=1}^m\sum_{k\le d}\mathsf{c}_{k_i}^{(i)}\int_\Omega\kappa_i\nabla\tilde{\mathrm{y}}_k^*\cdot\nabla v^*_{k_1,\dots,d_i+1,\dots,k_m}\,\mathrm{d}x\\
&=\sum_{i=1}^m\sum_{k\le d}\mathsf{c}_{k_i}^{(i)}\langle A_i\tilde{\mathrm{y}}_k^*,v^*_{k_1,\dots,d_i+1,\dots,k_m}\rangle_{Y^*,Y}\\
&\le\sum_{i=1}^m\sum_{k_j\le d_j,j\ne i}\Big\|\sum_{k_i\le d_i}(\mathsf{c}_{k_i}^{(i)}\cdot A_i\tilde{\mathrm{y}}_k^*)\Big\|_{A_{\mathrm{ref}}^{-1}}\cdot\|v^*_{k_1,\dots,d_i+1,\dots,k_m}\|_{A_{\mathrm{ref}}}\\
&\le\Big(\sum_{i=1}^m\sum_{k_j\le d_j,j\ne i}\Big\|A_i\sum_{k_i\le d_i}(\mathsf{c}_{k_i}^{(i)}\tilde{\mathrm{y}}_k^*)\Big\|^2_{A_{\mathrm{ref}}^{-1}}\Big)^{\frac12}\Big(\sum_{i=1}^m\sum_{k_j\le d_j,j\ne i}\|v^*_{k_1,\dots,d_i+1,\dots,k_m}\|^2_{A_{\mathrm{ref}}}\Big)^{\frac12}.
\end{aligned}
\tag{7.29}
$$

**Example 7.11.** For $\vartheta_k\equiv\beta_k$ we have $\mathsf{c}_{k_i}^{(i)}=0$ for all $k_i\le d_i-1$, and the estimate (7.29) becomes

$$
\begin{aligned}
&\sum_{i=1}^m\sum_{k\le d}\mathsf{c}_{k_i}^{(i)}\int_\Omega\kappa_i\nabla\tilde{\mathrm{y}}_k^*\cdot\nabla v^*_{k_1,\dots,d_i+1,\dots,k_m}\,\mathrm{d}x\le\\
&\Big(\sum_{i=1}^m\sum_{k_j\le d_j,j\ne i}(\mathsf{c}_{d_i}^{(i)})^2\|A_i\tilde{\mathrm{y}}^*_{k_1,\dots,d_i,\dots,k_m}\|^2_{A_{\mathrm{ref}}^{-1}}\Big)^{\frac12}\Big(\sum_{i=1}^m\sum_{k_j\le d_j,j\ne i}\|v^*_{k_1,\dots,d_i+1,\dots,k_m}\|^2_{A_{\mathrm{ref}}}\Big)^{\frac12}.
\end{aligned}
$$

For general $\vartheta_k$, we focus on the term $\sum_{k_i \le d_i} (\mathsf{c}_{k_i}^{(i)} \tilde{\mathrm{y}}_k^*) =: \tilde{\mathrm{y}}_{k,i}^*$. It holds that

$$\sum_{k_i \le d_i} (\mathsf{c}_{k_i}^{(i)} \tilde{\mathrm{y}}_k^*) = \int_{\Xi_i} \xi_i \beta_{d_i+1}^{(i)}(\xi_i) \sum_{k_i \le d_i} \vartheta_{k_i}^{(i)}(\xi_i) \tilde{\mathrm{y}}_k^* \,\mathrm{d}\mathbb{P}_i$$

due to $\vartheta_{d_i+1}^{(i)} \equiv \beta_{d_i+1}^{(i)}$, where the appearing integral is in the sense of Bochner. The term $\sum_{k_i \le d_i} \vartheta_{k_i}^{(i)}(\xi_i) \tilde{\mathrm{y}}_k^*$ is independent of the concrete choice of the basis $\{\vartheta_{k_i}^{(i)}\}_{k_i \in \mathbb{N}}$.

It still remains to compute

$$\|A_i \tilde{\mathrm{y}}_{k,i}^*\|_{A_{\mathrm{ref}}^{-1}}^2 = \langle A_{\mathrm{ref}}^{-1} A_i \tilde{\mathrm{y}}_{k,i}^*, A_i \tilde{\mathrm{y}}_{k,i}^* \rangle_{Y,Y^*} = \langle \hat{y}_{k,i}^*, A_i \tilde{\mathrm{y}}_{k,i}^* \rangle_{Y,Y^*} = \int_\Omega \kappa_i \nabla \tilde{\mathrm{y}}_{k,i}^* \cdot \nabla \hat{y}_{k,i}^* \,\mathrm{d}x$$

where we define $\hat{y}_{k,i}^* := A_{\mathrm{ref}}^{-1} A_i \tilde{\mathrm{y}}_{k,i}$, i. e.,

$$\int_\Omega \kappa_{\mathrm{ref}} \nabla \hat{y}_{k,i}^* \cdot \nabla v + \chi_{\mathrm{ref}} \hat{y}_{k,i}^* v \,\mathrm{d}x = \int_\Omega \kappa_i \nabla \tilde{\mathrm{y}}_{k,i}^* \cdot \nabla v \,\mathrm{d}x$$

holds for all $v \in Y$. One could now compute a discrete version $\hat{\mathrm{y}}_{k,i}^*$ of $\hat{y}_{k,i}^*$ and additionally estimate the error $\|\hat{y}_{k,i}^* - \hat{\mathrm{y}}_{k,i}^*\|_{A_{\mathrm{ref}}}$ with the techniques from Section 7.1. Alternatively, we can estimate

$$\begin{aligned}
&\int_\Omega \kappa_i \nabla \tilde{\mathrm{y}}_{k,i}^* \cdot \nabla \hat{y}_{k,i}^* \,\mathrm{d}x \\
&\le \Big( \int_\Omega |\kappa_i| \nabla \tilde{\mathrm{y}}_{k,i}^* \cdot \nabla \tilde{\mathrm{y}}_{k,i}^* \,\mathrm{d}x \Big)^{\frac{1}{2}} \cdot \Big( \int_\Omega |\kappa_i| \nabla \hat{y}_{k,i}^* \cdot \nabla \hat{y}_{k,i}^* \,\mathrm{d}x \Big)^{\frac{1}{2}} \\
&\le \Big( \int_\Omega |\kappa_i| \nabla \tilde{\mathrm{y}}_{k,i}^* \cdot \nabla \tilde{\mathrm{y}}_{k,i}^* \,\mathrm{d}x \Big)^{\frac{1}{2}} \cdot \| \tfrac{\kappa_i}{\kappa_{\mathrm{ref}}} \|_{L^\infty(\Omega)}^{\frac{1}{2}} \cdot \Big( \int_\Omega \kappa_{\mathrm{ref}} \nabla \hat{y}_{k,i}^* \cdot \nabla \hat{y}_{k,i}^* \,\mathrm{d}x \Big)^{\frac{1}{2}} \\
&\le \Big( \int_\Omega |\kappa_i| \nabla \tilde{\mathrm{y}}_{k,i}^* \cdot \nabla \tilde{\mathrm{y}}_{k,i}^* \,\mathrm{d}x \Big)^{\frac{1}{2}} \cdot \| \tfrac{\kappa_i}{\kappa_{\mathrm{ref}}} \|_{L^\infty(\Omega)}^{\frac{1}{2}} \cdot \Big( \int_\Omega \kappa_{\mathrm{ref}} \nabla \hat{y}_{k,i}^* \cdot \nabla \hat{y}_{k,i}^* + \chi_{\mathrm{ref}} (\hat{y}_{k,i}^*)^2 \,\mathrm{d}x \Big)^{\frac{1}{2}} \\
&= \Big( \int_\Omega |\kappa_i| \nabla \tilde{\mathrm{y}}_{k,i}^* \cdot \nabla \tilde{\mathrm{y}}_{k,i}^* \,\mathrm{d}x \Big)^{\frac{1}{2}} \cdot \| \tfrac{\kappa_i}{\kappa_{\mathrm{ref}}} \|_{L^\infty(\Omega)}^{\frac{1}{2}} \cdot \Big( \int_\Omega \kappa_i \nabla \tilde{\mathrm{y}}_{k,i}^* \cdot \nabla \hat{y}_{k,i}^* \,\mathrm{d}x \Big)^{\frac{1}{2}}.
\end{aligned}$$

It follows that

$$\|A_i \tilde{\mathrm{y}}_{k,i}^*\|_{A_{\mathrm{ref}}^{-1}}^2 = \int_\Omega \kappa_i \nabla \tilde{\mathrm{y}}_{k,i}^* \cdot \nabla \hat{y}_{k,i}^* \,\mathrm{d}x \le \| \tfrac{\kappa_i}{\kappa_{\mathrm{ref}}} \|_{L^\infty(\Omega)} \Big( \int_\Omega |\kappa_i| \nabla \tilde{\mathrm{y}}_{k,i}^* \cdot \nabla \tilde{\mathrm{y}}_{k,i}^* \,\mathrm{d}x \Big). \tag{7.30}$$

Thus, we define

$$\boldsymbol{\zeta}_i(\tilde{\mathbf{y}}) := \| \tfrac{\kappa_i}{\kappa_{\mathrm{ref}}} \|_{L^\infty(\Omega)}^{1/2} \Big( \sum_{k_j \le d_j, j \ne i} \Big( \int_\Omega |\kappa_i| \big( \sum_{k_i \le d_i} \mathsf{c}_{k_i}^{(i)} \nabla \tilde{\mathrm{y}}_k^* \big) \cdot \big( \sum_{k_i \le d_i} \mathsf{c}_{k_i}^{(i)} \nabla \tilde{\mathrm{y}}_k^* \big) \,\mathrm{d}x \Big) \Big)^{\frac{1}{2}} \tag{7.31}$$

in order to formulate the following error estimation theorem.

**Theorem 7.12.** *Let $\tilde{\mathbf{y}} \in \mathrm{Y} \otimes \mathcal{P}_d(\Xi)$ be given such that $\boldsymbol{r} = \boldsymbol{A}\tilde{\mathbf{y}} + \boldsymbol{N}(\tilde{\mathbf{y}}) - \boldsymbol{b} \in L^2_{\mathbb{P}}(\Xi; Y^*)$. Suppose that Assumptions 7.5 and 7.9 are satisfied with the definitions* (7.10)*. Then,*

$$\begin{aligned}\|\tilde{\mathbf{y}} - \boldsymbol{y}\|^2_{\hat{\boldsymbol{A}}_{\mathrm{ref}}} \leq \Lambda^2 \Bigg( & c_{\mathcal{T}} c_{A_{\mathrm{ref}}} \Big( \sum_{T \in \mathcal{T}} \boldsymbol{\eta}_T(\tilde{\mathbf{y}})^2 + \sum_{E \in \mathcal{E}^0} \boldsymbol{\eta}_E(\tilde{\mathbf{y}})^2 \Big)^{\frac{1}{2}} + \Big( \sum_{i=1}^{m} \boldsymbol{\zeta}_i(\tilde{\mathbf{y}})^2 \Big)^{\frac{1}{2}} \\ & + \Big\| \big( \chi \cdot \tilde{\mathbf{y}} - \tilde{\boldsymbol{\chi}} + \varphi(\tilde{\mathbf{y}}) - \tilde{\boldsymbol{\varphi}}, \cdot \big)_{L^2_{\mathbb{P}}(\Xi; L^2(\Omega))} \Big\|_{\hat{\boldsymbol{A}}^{-1}_{\mathrm{ref}}} \Bigg)^2 + \Lambda^2 \|\mathbf{w}\|^2_{\hat{\boldsymbol{A}}_{\mathrm{ref}}},\end{aligned} \tag{7.32}$$

*where* $\mathbf{w}$ *is defined by* (7.19)*,* $\boldsymbol{\eta}_T$*,* $\boldsymbol{\eta}_E$ *are defined by* (7.27)*,* (7.28)*, and* $\boldsymbol{\zeta}_i$ *is defined in* (7.31)*.*

*Proof.* From (7.18) and with $\boldsymbol{w} = \hat{\boldsymbol{A}}^{-1}_{\mathrm{ref}} \boldsymbol{r} \in \hat{\boldsymbol{Y}}$ and $\mathbf{w}$ defined by (7.19) we conclude

$$\|\tilde{\mathbf{y}} - \boldsymbol{y}\|^2_{\hat{\boldsymbol{A}}_{\mathrm{ref}}} \leq \Lambda^2 \big( \|\boldsymbol{w} - \mathbf{w}\|^2_{\hat{\boldsymbol{A}}_{\mathrm{ref}}} + \|\mathbf{w}\|^2_{\hat{\boldsymbol{A}}_{\mathrm{ref}}} \big) \tag{7.33}$$

due to Galerkin orthogonality.

For an arbitrary $\boldsymbol{v} \in \hat{\boldsymbol{Y}}$, $\boldsymbol{v}(x, \xi) = \sum_{k \in \mathbb{N}^m} v_k(x) \theta_k(\xi) = \sum_{k \in \mathbb{N}^m} v_k^*(x) \vartheta_k(\xi)$ with $v_k, v_k^* \in Y$ for all $k$, we have that

$$\|\boldsymbol{v}\|^2_{\hat{\boldsymbol{A}}_{\mathrm{ref}}} = \int_{\Xi} \big\langle A_{\mathrm{ref}} \sum_{k \in \mathbb{N}^m} v_k(\cdot) \theta_k(\xi), \sum_{l \in \mathbb{N}^m} v_l(x) \theta_l(\xi) \big\rangle_{Y^*, Y} \, \mathrm{d}\mathbb{P} = \sum_{k \in \mathbb{N}^m} \|v_k\|^2_{A_{\mathrm{ref}}}$$

and $\|\boldsymbol{v}\|^2_{\hat{\boldsymbol{A}}_{\mathrm{ref}}} = \sum_{k \in \mathbb{N}^m} \|v_k^*\|^2_{A_{\mathrm{ref}}}$ because the polynomials $\{\theta_k\}_k$, $\{\vartheta_k\}_k$ are orthonormal.

Combining (7.21), (7.25), (7.26), (7.29), (7.30) and the fact that

$$\langle \hat{\boldsymbol{A}}_{\mathrm{ref}}(\boldsymbol{w} - \mathbf{w}), \boldsymbol{v} \rangle_{\hat{\boldsymbol{Y}}^*, \hat{\boldsymbol{Y}}} = \big( \boldsymbol{w} - \mathbf{w}, \boldsymbol{v} \big)_{\hat{\boldsymbol{A}}_{\mathrm{ref}}} \leq c \cdot \boldsymbol{\eta}(\mathbf{w}) \cdot \|\boldsymbol{v}\|_{\hat{\boldsymbol{A}}_{\mathrm{ref}}} \text{ for all } \boldsymbol{v} \in \hat{\boldsymbol{Y}}$$

yields $\|\boldsymbol{w} - \mathbf{w}\|_{\hat{\boldsymbol{A}}_{\mathrm{ref}}} \leq c \cdot \boldsymbol{\eta}(\mathbf{w})$, gives

$$\begin{aligned}\|\boldsymbol{w} - \mathbf{w}\|_{\hat{\boldsymbol{A}}_{\mathrm{ref}}} \leq \; & c_{\mathcal{T}} c_{A_{\mathrm{ref}}} \Big( \sum_{T \in \mathcal{T}} \boldsymbol{\eta}_T(\tilde{\mathbf{y}})^2 + \sum_{E \in \mathcal{E}^0} \boldsymbol{\eta}_E(\tilde{\mathbf{y}})^2 \Big)^{\frac{1}{2}} + \Big( \sum_{i=1}^{m} \boldsymbol{\zeta}_i(\tilde{\mathbf{y}})^2 \Big)^{\frac{1}{2}} \\ & + \Big\| \big( \chi \cdot \tilde{\mathbf{y}} - \tilde{\boldsymbol{\chi}} + \varphi(\tilde{\mathbf{y}}) - \tilde{\boldsymbol{\varphi}}, \cdot \big)_{L^2_{\mathbb{P}}(\Xi; L^2(\Omega))} \Big\|_{\hat{\boldsymbol{A}}^{-1}_{\mathrm{ref}}},\end{aligned}$$

which yields the result together with (7.33). □

Theorem 7.12 provides a nice split of the error into different contributions. The last term in (7.32) is exactly the algebraic error due to the inexact solution of the discrete system by, e. g., a low-rank tensor solver. The first term consists of

- a term related to the FE discretization error, which can itself be split into error contributions for each element,
- the error contribution due to the discretization of the stochastic space, which consists of terms for each stochastic parameter, and
- the interpolation error coming from the approximation of the nonlinear terms in the equation.

**Remark 7.13.** It remains to discuss the condition $\boldsymbol{r} \in L^2_{\mathbb{P}}(\Xi, Y^*)$ for the state and adjoint equation from Section 3.2, cf. Remark 7.4, provided $Q(\xi) \equiv \iota : H^1_0(\Omega) \hookrightarrow L^2(\Omega)$.
If $\tilde{\mathbf{y}} \in L^{r_f}_{\mathbb{P}}(\Xi; Y)$ $(r_f \geq p)$ is an inexact solution of the state equation, the residual $\boldsymbol{r}$ belongs to $L^{r_f/(p-1)}_{\mathbb{P}}(\Xi; Y^*)$ due to the growth condition (3.7) on the nonlinearity $\varphi$. Thus, we need $r_f \geq 2p-2$ to have $\boldsymbol{r} \in L^2_{\mathbb{P}}(\Xi, Y^*)$. Under the regularity assumption in the "Example" column of Table 3.1, i. e., $r_Q = \infty$ and $r_{\hat{q}} \geq r_f$, it is realistic that $\tilde{\mathbf{z}} \in L^{r_f}_{\mathbb{P}}(\Xi; Y)$ holds, because the exact adjoint state $\boldsymbol{z}$ has exactly this regularity. The residual of the adjoint equation then also belongs to $L^{r_f/(p-1)}_{\mathbb{P}}(\Xi; Y^*)$ which is a subset of $L^2_{\mathbb{P}}(\Xi; Y^*)$ by $r_f \geq 2p-2$.

The error estimate given in Theorem 7.12 uses a deterministic reference operator as the one discussed in [22, 23]. The results in these papers rely on a saturation assumption, which we do not make here, but derive an error estimate similar to the one presented in [38] and used in [39, 40]. All mentioned papers present error estimates for linear equations whereas we consider a class of semilinear equations with a monotone nonlinearity, but have used a linear reference equation to derive the error estimate. In the linear case, it makes sense to discuss the efficiency of the estimator using a lower bound as indicated in Remark 7.2. Furthermore, the convergence of the adaptive solution technique should be investigated. Since it is already discussed in the linear case for a very similar estimator in [39], we skip this topic at this point.

For the case $\hat{\varphi} \equiv 0$ and $\chi \equiv 0$, i. e., we have a linear equation and no interpolation error, we briefly compare our Theorem 7.12 to [38, Thm. 6.2]: There, an additional discrete error term (analog of $\|\mathbf{w}\|_{\hat{\boldsymbol{A}}_{\mathrm{ref}}}$) appears as an additional summand in the "large bracket" in (7.32). We do not get this term because we make use of the *exact* solution of the reference equation (7.19), which is realistic when we work with low-rank tensors, where the discrete rank-1-operator $\mathbf{A}^{-1}_{\mathrm{ref}} = \mathsf{A}^{-1}_{\mathrm{ref}} \otimes \mathsf{I} \cdots \otimes \mathsf{I}$ can be applied by an $i$-mode matrix product so that it can be used as a typically very good preconditioner, see below.

## 7.4. Realization with Low-Rank Tensors

In order to implement an adaptive solver for (7.17) based on the estimate (7.32) with low-rank tensors, we have to apply an iterative low-rank tensor solver such as truncated PCG [75], ALS [62], or AMEn [36] to the discretized system

$$\mathbf{A}\mathbf{y} + \mathbf{N}(\mathbf{y}) - \mathbf{b} = 0, \tag{7.34}$$

cf. (6.16) and (6.21). In addition, we have to implement the evaluation of the error indicators $\boldsymbol{\eta}_T(\tilde{\mathbf{y}})$, $\boldsymbol{\eta}_E(\tilde{\mathbf{y}})$, and $\boldsymbol{\zeta}_i(\tilde{\mathbf{y}})$ from the low-rank solution in an efficient manner. Based on the value of these error indicators, the discrete subspace with the largest error contribution is refined, provided the algebraic error $\|\mathbf{w}\|_{\hat{\boldsymbol{A}}_{\mathrm{ref}}}$ does not dominate. In the latter case, additional iterations of the low-rank tensor solver have to be performed.

### Solution of the Discrete System

Let us set $\mathbf{N} \equiv 0$ for a moment. As pointed out in Subsection 2.1.3, iterative solvers working with the low-rank tensor representation such as ALS or AMEn aim for solving the linear

system $\mathbf{Ay} = \mathbf{b}$ by minimizing the squared Frobenius norm $\|\mathbf{Ay} - \mathbf{b}\|_\mathsf{F}^2$ of the residual. This problem can be badly conditioned and as indicated by (7.32), we are in fact interested in minimizing

$$\|\mathbf{w}\|_{\hat{\boldsymbol{A}}_{\text{ref}}}^2 = \langle \boldsymbol{r}, \mathbf{w} \rangle_{\hat{\boldsymbol{Y}}^*, \hat{\boldsymbol{Y}}} = \langle \mathbf{r}, \hat{\mathbf{A}}_{\text{ref}}^{-1} \mathbf{r} \rangle =: \|\mathbf{r}\|_{\hat{\mathbf{A}}_{\text{ref}}^{-1}}^2,$$

where (7.19) translates to $\hat{\mathbf{A}}_{\text{ref}} \mathbf{w} = \mathbf{r} = \mathbf{Ay} - \mathbf{b}$ in the discrete tensor space. Let the symmetric, positive definite operator $\hat{\mathbf{A}}_{\text{ref}}$ now be decomposed as $\hat{\mathbf{A}}_{\text{ref}} = \hat{\mathbf{R}}_{\text{ref}}^* \hat{\mathbf{R}}_{\text{ref}}$, where $\hat{\mathbf{R}}_{\text{ref}}^*$ is the adjoint operator of $\hat{\mathbf{R}}_{\text{ref}} : \mathbb{R}^{n_0 \times \cdots \times n_m} \to \mathbb{R}^{n_0 \times \cdots \times n_m}$ w. r. t. the Frobenius inner product. Then,

$$\begin{aligned} \arg\min_{\mathbf{y}} ||\mathbf{Ay} - \mathbf{b}||_{\hat{\mathbf{A}}_{\text{ref}}^{-1}}^2 &= \arg\min_{\mathbf{y}} \langle \mathbf{Ay} - \mathbf{b}, \hat{\mathbf{A}}_{\text{ref}}^{-1} (\mathbf{Ay} - \mathbf{b}) \rangle \\ &= \arg\min_{\mathbf{y}} ||\hat{\mathbf{R}}_{\text{ref}}^{-*} (\mathbf{Ay} - \mathbf{b})||_\mathsf{F}^2 \\ &= \hat{\mathbf{R}}_{\text{ref}}^{-1} \arg\min_{\widehat{\mathbf{y}}} ||\hat{\mathbf{R}}_{\text{ref}}^{-*} \mathbf{A} \hat{\mathbf{R}}_{\text{ref}}^{-1} \widehat{\mathbf{y}} - \mathbf{R}_{\text{ref}}^{-*} \mathbf{b}||_\mathsf{F}^2. \end{aligned} \tag{7.35}$$

This means that the preconditioned, symmetric system $\hat{\mathbf{R}}_{\text{ref}}^{-*} \mathbf{A} \hat{\mathbf{R}}_{\text{ref}}^{-1} \widehat{\mathbf{y}} = \hat{\mathbf{R}}_{\text{ref}}^{-*} \mathbf{b}$ can be solved by a standard implementation of, e. g., AMEn and its (approximate) solution $\widehat{\mathbf{y}}$ can be transformed to the (approximate) solution $\mathbf{y} = \hat{\mathbf{R}}_{\text{ref}}^{-1} \widehat{\mathbf{y}}$ of the original system $\mathbf{Ay} = \mathbf{b}$. In particular, we can use the tensor product form $\hat{\mathbf{A}}_{\text{ref}} = \hat{\mathsf{A}}_{\text{ref}} \otimes \mathsf{I} \otimes \cdots \otimes \mathsf{I}$ with the positive definite, quadratic matrix $\hat{\mathsf{A}}_{\text{ref}}$, and compute, e. g., a sparse Cholesky decomposition $\hat{\mathsf{A}}_{\text{ref}} = \hat{\mathsf{R}}_{\text{ref}}^\top \hat{\mathsf{R}}_{\text{ref}}$, where $\hat{\mathsf{R}}_{\text{ref}}$ is the product of a triangular and a permutation matrix. Then, with $\hat{\mathbf{R}}_{\text{ref}} := \hat{\mathsf{R}}_{\text{ref}} \otimes \mathsf{I} \otimes \cdots \otimes \mathsf{I}$ we obtain a decomposition of $\hat{\mathbf{A}}_{\text{ref}}$ as desired because $\hat{\mathbf{R}}_{\text{ref}}^* \hat{\mathbf{R}}_{\text{ref}} = (\hat{\mathsf{R}}_{\text{ref}}^\top \hat{\mathsf{R}}_{\text{ref}}) \otimes \mathsf{I} \otimes \cdots \otimes \mathsf{I}$. It is then algorithmically relevant that $\hat{\mathbf{R}}_{\text{ref}}^{-1} = \hat{\mathsf{R}}_{\text{ref}}^{-1} \otimes \mathsf{I} \otimes \cdots \otimes \mathsf{I}$, i. e., the inverse can be applied very efficiently to low-rank tensors. For our concrete implementation, we have extended the AMEn algorithm to be able to handle preconditioners of canonical rank 1 given in decomposed form. Since, e. g., the inverse Cholesky factor $\hat{\mathsf{R}}_{\text{ref}}^{-1}$ should never be formed explicitly, but can be applied to vectors efficiently by backward substitution and permutation, we have extended the {d, R}-format [35], which is essentially a collection of sparse matrices, each of which acts on one mode of the tensor, to function handles. We want to remark that it would be even more desirable to have a preconditioned version of AMEn which does only require the application of the preconditioner $\hat{\mathbf{A}}_{\text{ref}}^{-1}$ itself and not a decomposition of it (similar to the PCG method). But investigating and implementing such an algorithm is out of the scope of this thesis. Additionally, the operator $\hat{\mathbf{A}}_{\text{ref}}^{-1}$ is needed more frequently in the algorithm, e. g., to evaluate the norm of the residual of the nonlinear equation. Since $\hat{\mathbf{A}}_{\text{ref}}$ is symmetric, positive definite, and sparse, the application of its inverse is computed by the mentioned sparse Cholesky decomposition in MATLAB so that this can be precomputed and used in AMEn.

A similar idea can be used to truncate a given tensor $\mathbf{y}$ to smaller rank w. r. t. the $\mathbf{M}$-norm, where $\mathbf{M} : \mathbb{R}^{n_0 \times \cdots \times n_m} \to \mathbb{R}^{n_0 \times \cdots \times n_m}$ induces some inner product and is decomposed as $\mathbf{M} = \tilde{\mathbf{R}}^* \tilde{\mathbf{R}}$ w. r. t. the Frobenius inner product. In fact, SVD-based truncation would quasi-minimize the function $\|\mathbf{z} - \mathbf{y}\|_\mathsf{F}^2$ w. r. t. $\mathbf{z}$, see Subsection 2.1.2. Instead, we can use the transformation (7.35) and approximate the tensor $\tilde{\mathbf{R}}\mathbf{y}$ by standard truncation w. r. t. the Frobenius norm to obtain the tensor $\widehat{\mathbf{z}}$. Then, $\tilde{\mathbf{R}}^{-1} \widehat{\mathbf{z}}$ is the respective approximation of $\mathbf{y}$ w. r. t. the $\mathbf{M}$-norm.

If $\mathbf{N} \not\equiv 0$ is nonlinear, we solve the system (7.34) by Newton's method, where each Newton step is computed approximately by the preconditioned AMEn solver. Especially in the last iterations of the trust-region algorithm, the state and the adjoint state do not change very much. Then it is often enough to take one inexact Newton step, which is computed by one AMEn sweep, to decrease the residual of the nonlinear equation sufficiently.

### Evaluation of the Error Indicators

The discretization error indicators can be implemented as follows: If, e. g.,

$$\tilde{\mathbf{y}}(x,\xi) = \sum_{k \le d} \tilde{\mathrm{y}}_k(x)\theta_k(\xi)$$

is represented by the tensor $\tilde{\mathbf{y}}$ analogously to (6.8),

$$\tilde{\mathrm{y}}_k(x) = \sum_{k_0=1}^{d_0} \tilde{\mathbf{y}}(k_0,k)\phi_{k_0}(x),$$

holds, i. e., the function $\tilde{\mathrm{y}}_k$ is represented by the vector $\tilde{\mathbf{y}}(\bullet,k)$.

Knowing that, we establish the evaluation of the triangle error contribution (7.27). Let $x^1, x^2, x^3 \in \Omega$ be the three vertices of the triangle $T$ and let the function $\tilde{\mathrm{f}}$ be affine on $T$. Then,

$$\begin{aligned}\|\tilde{\mathrm{f}}\|^2_{L^2(T)} = \int_T \tilde{\mathrm{f}}(x)^2 \,\mathrm{d}x &= \tfrac{a_T}{6}\big(\tilde{\mathrm{f}}(x^1)^2 + \tilde{\mathrm{f}}(x^2)^2 + \tilde{\mathrm{f}}(x^3)^2 + \tilde{\mathrm{f}}(x^1)\tilde{\mathrm{f}}(x^2) + \tilde{\mathrm{f}}(x^1)\tilde{\mathrm{f}}(x^3) + \tilde{\mathrm{f}}(x^2)\tilde{\mathrm{f}}(x^3)\big)\\ &= \tfrac{a_T}{6}\sum_{j=1}^{3}\sum_{l=j}^{3}\tilde{\mathrm{f}}(x^j)\tilde{\mathrm{f}}(x^l),\end{aligned}$$

where $a_T$ is the area of $T$. We approximate each $\tilde{\chi}_k + \tilde{\varphi}_k - f_k - \chi_{\mathrm{ref}}\mathrm{w}_k \approx \tilde{\mathrm{f}}_k$ by interpolation, where $\tilde{\mathrm{f}}_k$ are linear finite element functions for all $k$. These functions are represented altogether by a single tensor $\tilde{\mathbf{f}} \in \mathbb{R}^{\tilde{d}\times d_1\times\cdots\times d_m}$ of values at the FE nodes. Note that $\tilde{d} > d_0$ is the number of all FE nodes whereas $d_0$ counts only the interior nodes. Let $k_0^{(1)}, k_0^{(2)}, k_0^{(3)} \in [\tilde{d}]$ be the indices corresponding to the vertices $x^1$, $x^2$, $x^3$, respectively, so that $\tilde{\mathbf{f}}(k_0^{(j)},k) = \tilde{\mathrm{f}}_k(x^j)$ for all $j$ and all $k$. We obtain

$$\begin{aligned}\boldsymbol{\eta}_T(\tilde{\mathbf{y}})^2 \approx h_T^2\sum_{k\le d}\|\tilde{\mathrm{f}}_k\|^2_{L^2(T)} &= h_T^2\tfrac{a_T}{6}\sum_{k\le d}\Big(\sum_{j=1}^{3}\sum_{l=j}^{3}\tilde{\mathbf{f}}(k_0^{(j)},k)\tilde{\mathbf{f}}(k_0^{(l)},k)\Big)\\ &= h_T^2\tfrac{a_T}{6}\sum_{j=1}^{3}\sum_{l=j}^{3}\langle \mathbb{1}, \tilde{\mathbf{f}}(k_0^{(j)},\bullet)\odot\tilde{\mathbf{f}}(k_0^{(l)},\bullet)\rangle.\end{aligned}$$

The advantage of this formulation is that it can be vectorized using the tensors of evaluations at all first, second, and third triangle vertices, respectively, to evaluate the error indicator for all triangles $T$ simultaneously. Then, only componentwise multiplication and contraction with the rank-1-tensor $\mathbb{1}$ are needed.

To compute the edge error contribution (7.28), one first computes the values of the partial derivatives $\partial_{x_1}\tilde{\mathrm{y}}_k$ and $\partial_{x_2}\tilde{\mathrm{y}}_k$ on the elements. These values are contained in the tensors $\mathsf{G}_1 \circ_1 \tilde{\mathbf{y}}$ and $\mathsf{G}_2 \circ_1 \tilde{\mathbf{y}}$, where $\mathsf{G}_1, \mathsf{G}_2 \in \mathbb{R}^{|\mathcal{T}|\times d_0}$ are sparse matrices mapping a vector of function values on the FE nodes to a vector of derivative values on the triangles numbered from 1 to $|\mathcal{T}|$. Analogously, we obtain the tensors $\mathsf{G}_1 \circ_1 \mathbf{w}$ and $\mathsf{G}_2 \circ_1 \mathbf{w}$ containing the values of $\partial_{x_1}\mathrm{w}_k$ and $\partial_{x_2}\mathrm{w}_k$. Furthermore, we create the tensors $\boldsymbol{\kappa}_{\mathrm{ref}} = \kappa_{\mathrm{ref}} \otimes \mathbb{1} \in \mathbb{R}^{|\mathcal{T}|\times d_1 \times \cdots \times d_m}$, where $\kappa_{\mathrm{ref}} \in \mathbb{R}^{|\mathcal{T}|}$ contains the values of $\kappa_{\mathrm{ref}}$ on the triangles, and $\boldsymbol{\kappa} \in \mathbb{R}^{|\mathcal{T}|\times d_1 \times \cdots \times d_m}$ such that $\boldsymbol{\kappa}(\bullet, k)$ contains the respective values of $\kappa(\cdot, \mathsf{a}_k)$. Let now $\boldsymbol{\nu}^{j,1} \in \mathbb{R}^{|\mathcal{T}|}$ contain the first components of the outwards-pointing, normal vectors corresponding to the $j$-th edge of the triangles $T$, and let $\boldsymbol{\nu}^{j,2} \in \mathbb{R}^{|\mathcal{T}|}$ contain the second components for $j \in \{1,2,3\}$. This means that if $T$ is the $\ell$-th triangle, the outer normal vector $\nu_{T,j}$ corresponding to its $j$-th edge is given by $\nu_{T,j} = \begin{pmatrix} \boldsymbol{\nu}_\ell^{j,1} & \boldsymbol{\nu}_\ell^{j,2} \end{pmatrix}^\top$. Then, the values of

$$\kappa(\cdot, \mathsf{a}_k)\nabla\tilde{\mathrm{y}}_k \cdot \nu_{T,j}$$

for all $T$ are given by the single tensor

$$\boldsymbol{\kappa} \odot \big(\mathrm{diag}(\boldsymbol{\nu}^{j,1}) \circ_1 \mathsf{G}_1 \circ_1 \tilde{\mathbf{y}} + \mathrm{diag}(\boldsymbol{\nu}^{j,2}) \circ_1 \mathsf{G}_2 \circ_1 \tilde{\mathbf{y}}\big) = \boldsymbol{\kappa} \odot \big((\mathrm{diag}(\boldsymbol{\nu}^{j,1})\mathsf{G}_1 + \mathrm{diag}(\boldsymbol{\nu}^{j,2})\mathsf{G}_2) \circ_1 \tilde{\mathbf{y}}\big).$$

With an analogous consideration we get that the values of

$$(\kappa(\cdot, \mathsf{a}_k)\nabla\tilde{\mathrm{y}}_k - \kappa_{\mathrm{ref}}\nabla\mathrm{w}_k) \cdot \nu_{T,j}$$

are contained in the tensor

$$\tilde{\mathbf{g}}_j := \boldsymbol{\kappa} \odot \big((\mathrm{diag}(\boldsymbol{\nu}^{j,1})\mathsf{G}_1 + \mathrm{diag}(\boldsymbol{\nu}^{j,2})\mathsf{G}_2) \circ_1 \tilde{\mathbf{y}}\big) - \boldsymbol{\kappa}_{\mathrm{ref}} \odot \big((\mathrm{diag}(\boldsymbol{\nu}^{j,1})\mathsf{G}_1 + \mathrm{diag}(\boldsymbol{\nu}^{j,2})\mathsf{G}_2) \circ_1 \mathbf{w}\big),$$

which can be computed using standard low-rank tensor arithmetics, namely $i$-mode matrix products as well as componentwise multiplication and subtraction. In a next step, we number the interior edges $E \in \mathcal{E}_0$ from 1 to $|\mathcal{E}_0|$ and create sparse matrices $\mathsf{H}_j \in \{0,1\}^{|\mathcal{E}_0|\times|\mathcal{T}|}$ ($j \in \{1,2,3\}$) mapping vectors of values on the triangles to vectors of values on the respective $j$-th triangle edges in the correct order, provided they are interior edges. Then the tensor of jumps over the interior edges is given by

$$\tilde{\mathbf{h}} := \sum_{j=1}^{3} \mathsf{H}_j \circ_1 \tilde{\mathbf{g}}_j \in \mathbb{R}^{|\mathcal{E}_0|\times d_1 \times \cdots \times d_m}. \tag{7.36}$$

It is correct to sum here because the outer unit normals of two neighboring triangle point in the opposite direction. Using the presented definitions, the edge error indicator is given by

$$\boldsymbol{\eta}_E(\tilde{\mathbf{y}})^2 = h_E \sum_{k\le d} \|[\![(\kappa(\cdot, \mathsf{a}_k)\nabla\tilde{\mathrm{y}}_k - \kappa_{\mathrm{ref}}\nabla\mathrm{w}_k) \cdot \nu_E]\!]_E\|^2_{L^2(E)} = h_E^2\langle \mathbb{1}, \tilde{\mathbf{h}}(l, \bullet).^2\rangle,$$

where $l \in \{1, \ldots, |\mathcal{E}_0|\}$ is the number of the edge $E$. As before, this procedure can be vectorized to compute the error contributions of all interior edges simultaneously. Finally, we assign half of the error to each of the two neighboring triangles. Based on that, we mark all

triangles with the largest error contributions which constitute a certain amount $\vartheta_{\boldsymbol{\eta}} \in (0,1)$ of the total error, see [38, Sec. 7.1], a so-called Dörfler strategy [37]. These triangles are refined regularly, i. e., divided into four triangles of the same shape. To avoid hanging nodes, additional triangles have to be divided into two new ones possibly.

To evaluate $\boldsymbol{\zeta}_i(\tilde{\mathbf{y}})$ (see (7.31)) based on the tensor $\tilde{\mathbf{y}}$ representing the function $\tilde{\mathrm{y}}$, one first computes the values $\mathsf{c}_{l_i}^{(i)} = \int_{\Xi_i} \xi_i \vartheta_{l_i}^{(i)}(\xi_i) \vartheta_{d_i+1}^{(i)}(\xi_i) \,\mathrm{d}\mathbb{P}_i$ by a quadrature formula of high enough order, e. g., by Gaussian quadrature with $d_i + 1$ nodes if $\vartheta_{l_i}^{(i)}$ has degree $d_i - 1$ and $\vartheta_{d_i+1}^{(i)}$ has degree $d_i$, or analytically and writes them as one vector $\mathsf{c}^{(i)} \in \mathbb{R}^{d_i}$. Furthermore, the matrix $\bar{\mathsf{A}}_i \in \mathbb{R}^{d_0 \times d_0}$ given by

$$(\bar{\mathsf{A}}_i)_{k_0 l_0} := (|\kappa_i| \nabla \phi_{l_0}, \nabla \phi_{k_0})_{L^2(\Omega)^n},$$

cf. Example 6.5, is assembled. Then we have

$$\int_\Omega |\kappa_i| \Big( \sum_{k_i \le d_i} \mathsf{c}_{k_i}^{(i)} \nabla \tilde{\mathrm{y}}_k^* \Big) \cdot \Big( \sum_{k_i \le d_i} \mathsf{c}_{k_i}^{(i)} \nabla \tilde{\mathrm{y}}_k^* \Big) \,\mathrm{d}x =$$
$$((\mathsf{c}^{(i)^\top} \circ_{i+1} \tilde{\mathbf{y}})(\bullet, k_1, \dots, k_{i-1}, k_{i+1}, \dots, k_m))^\top \bar{\mathsf{A}}_i ((\mathsf{c}^{(i)^\top} \circ_{i+1} \tilde{\mathbf{y}})(\bullet, k_1, \dots, k_{i-1}, k_{i+1}, \dots, k_m)).$$

Summing these values over all $k_j \in \{1, \dots, d_j\}$ $(j \neq i)$ can be implemented as an inner product of tensors giving

$$\boldsymbol{\zeta}_i(\tilde{\mathbf{y}})^2 = \| \tfrac{\kappa_i}{\kappa_{\mathrm{ref}}} \|_{L^\infty(\Omega)} \langle \mathsf{c}^{(i)^\top} \circ_{i+1} \tilde{\mathbf{y}}, \bar{\mathsf{A}}_i \circ_1 \mathsf{c}^{(i)^\top} \circ_{i+1} \tilde{\mathbf{y}} \rangle.$$

If this is the largest contribution to the total error (7.32), we increase the respective polynomial degree $d_i - 1$ by 1.

### Interpolation to a Finer Subspace

Lifting the current solution tensor $\tilde{\mathbf{y}}$ representing the function $\tilde{\mathrm{y}} \in \mathbf{Y}$ to a tensor representing the same function as an element of a finer subspace is easy: For the refinement of the linear FE space Y, a sparse interpolation matrix can be constructed. Its upper block is the identity matrix because the coefficients belonging to the retained FE nodes are not changed. The coefficients of new nodes inserted on an edge are computed as a convex combination of the coefficients of the edge endpoints. Lifting the tensor $\tilde{\mathbf{y}}$ to a finer FE space then consists of applying the sparse interpolation matrix to the first tensor mode. The interpolation matrices of the polynomial spaces $\mathcal{P}_{d_i-1}(\Xi_i)$ with the weighted Lagrangian bases can be constructed by, e. g., evaluating the Lagrange polynomials of the coarse space at the Gaussian quadrature nodes of the finer space. Analogously, these dense but typically small interpolation matrices can be applied to the $(i+1)$-st mode of $\tilde{\mathbf{y}}$.

# 8. Implementation and Numerical Results

The inexact trust-region method (Algorithm 1) is implemented in MATLAB to solve different instances of the model problem presented in Chapter 3. In order to solve the arising PDEs with uncertain inputs, the adaptive solution technique described in Chapter 7 is used. The objective function evaluation and gradient error are estimated as derived in Chapter 5.

For this purpose, we consider the following concrete setup for the model problem (3.3) with state equation (3.8) and objective function (3.4):

- We choose uniformly distributed parameters, meaning that $\Xi_i = (-1, 1)$ and $\mathbb{P}_i = \frac{1}{2}\lambda$ for all $i \in [m]$, where $\lambda$ is the Lebesgue measure on $(-1, 1)$. Assumption 6.3 is satisfied by this choice.

- The domain $\Omega := (-1, 1)^2 \setminus (-1, 0]^2 \subset \mathbb{R}^2$ is the polygonal (cf. Assumptions 6.1 and 7.5) L-shaped domain, frequently used to test adaptive FE codes. It is divided into $m = 6$ subdomains $\Omega_1, \dots, \Omega_6$ as depicted in Figure 8.1. The initial FE mesh, which contains 113 nodes, and thus all refined meshes respect this partition of the domain, meaning that each element $T \in \mathcal{T}$ is contained in exactly one subdomain $\Omega_i$, see also Figure 8.1.

- The coefficient function is chosen to be $\kappa(x, \xi) = \kappa_0(x)\big(1 + \sum_{i=1}^{m} \xi_i \eta_i(x)\big)$ (cf. Example 3.17) with $\kappa_0 \equiv 1$ and $\eta_i(x) = \sigma_i \, 1_{\Omega_i(x)}$, where the vector $\sigma \in [0, 1)^m$ describing the amounts of influence of the uncertain parameters is chosen differently for different tests. The reference coefficient is $\kappa_{\text{ref}} \equiv 1$. Thus, Assumption 7.5 and the first part of Assumption 7.9 with $\kappa_i = \kappa_0 \eta_i$ hold true. Furthermore, $\|\cdot\|_{A_{\text{ref}}} \equiv \|\cdot\|_{H_0^1(\Omega)}$ and the constant $c_{A_{\text{ref}}}$ introduced in (7.13) comes from the Poincaré inequality $\|\cdot\|_{H^1(\Omega)} \leq c_{A_{\text{ref}}} \|\cdot\|_{H_0^1(\Omega)}$.

- We use $U = L^2(\Omega)$, i. e., $\Omega_u = \Omega$, $D \equiv I : L^2(\Omega) \to L^2(\Omega)$, giving that the second part of Assumption 6.1 is satisfied, $f(\cdot) \equiv 0$, and $\varphi(t) := t^3$. The reason for the latter choice is that this nonlinearity and its derivative can be evaluated simply with low-rank tensors because only componentwise multiplication is needed.

Overall, Assumption 3.3 is fulfilled with $\underline{\kappa} = 1 - \max_{i \in [m]} \sigma_i$, $\overline{\kappa} = 1 + \max_{i \in [m]} \sigma_i$, $a''_{\varphi''} = 0$, $c''_{\varphi''} = 6$, $p = 4$, $r_f = \infty$. Additionally, (7.3) holds with $\lambda = \underline{\kappa}$ and $\Lambda = \overline{\kappa}$.

- The observation space is $H = L^2(\Omega)$ and $Q(\cdot) \equiv \iota : H_0^1(\Omega) \hookrightarrow\hookrightarrow L^2(\Omega)$ is constant.

- The desired state $\hat{q}(\cdot) \equiv \hat{q} \in L^2(\Omega)$ is constant and chosen differently for different tests.

Hence, we have the integrability exponents $r_Q = \infty$ and $r_{\hat{q}} = \infty$ and thus $r_{Qy} = \infty$, $\hat{r} = \infty$, and $r_z = \infty$ (see Table 3.1). Assumption 5.6 holds true and the error estimation scheme

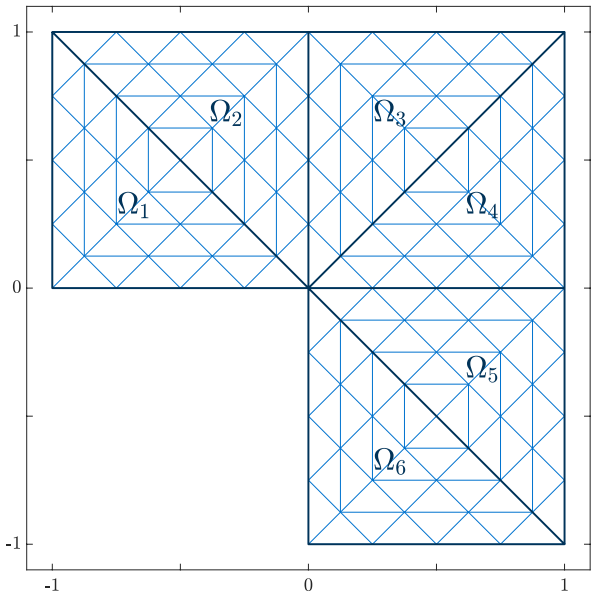



Figure 8.1.: The used physical domain $\Omega$, its partition into subdomains $\Omega_i$, and the intial finite element mesh.

given in Figure 5.2 can be applied. Assumption 6.1 is satisfied for arbitrary degree vectors $d \in \mathbb{N}^m$. Additionally, the right-hand sides of the state and adjoint equation are polynomials of coordinate degree at most $d - \mathbb{1}$ if the current approximate state $\tilde{\boldsymbol{y}}$ is such a polynomial, see Assumption 7.9 and also Remark 7.4.

- We take $\gamma = 10^{-3}$ and the set of admissible controls shall be $U_{\text{ad}} := \{u \in L^2(\Omega) : u(x) \leq 14 \text{ for a. e. } x \in \Omega\}$, where the upper bound is chosen such that it becomes active, but the optimal state is in the order of magnitude of the desired state. Therefore, the prerequisites of Theorem 3.19 as well as Assumption 4.2 are fulfilled and we can apply the exact and the inexact projection formula given in Section 6.4. The constant upper bound makes it possible to evaluate the error due to the "node-wise" projection for any FE mesh (Remark 5.4). In all tests, we initialize the algorithm with $u^0 = 0$.

The constants required for error estimation are set as follows: In Theorem 5.8 and in Lemma 5.10, we choose the Sobolev constant $c_{\tilde{r}} = c_p = c_4 = 0.5$ because we estimated $c_2 \approx 0.3$ numerically. The norm $\|\boldsymbol{y}\|_{L^\infty_{\mathbb{P}}(\Xi;Y)}$ in (5.20) is estimated by (3.9) with the Poincaré constant set to $C_\Omega = 1$. Additionally, we assume that $\|\tilde{\boldsymbol{y}}\|_{L^\infty_{\mathbb{P}}(\Xi;Y)}$ is approximately of the same size as $\|\boldsymbol{y}\|_{L^\infty_{\mathbb{P}}(\Xi;Y)}$ and neglect this term in (5.20). For a more rigorous implementation it could be estimated by a method for tensor maximum estimation as described in Subsection 2.1.3, but we skip this step for efficiency reasons. We also ignore the interpolation error in (7.12) and set $c_{\mathcal{T}} c_{A_{\text{ref}}} = 10^{-3}$ in this estimate to have all error contributions in the same order of magnitude. Even if the real, unknown constants are underestimated by the described choices, the algorithm still works because we need to know the error only up to a fixed, but possibly unknown multiplicative constant. In fact, an unrealistic choice of constants can be compensated by the choice of the error functions in the trust-region algorithm.

## 8.1. Implementation Details

The implementation is done in MATLAB R2017b because the use of various toolboxes facilitates it. The Partial Differential Equation Toolbox in MATLAB is used for basic FE tasks, such as mesh generation and refinement, or assembling matrices. Some finite element related code, such as the FE error estimator or the exact $L^2$-projection onto a box, is not available within this toolbox and has to be written additionally. Since the FE error estimator has to be applied to low-rank tensors, we need to create suitable sparse matrices, see (7.36) for example. The exact projection with the suitable grid refinement has to be vectorized in order to have an efficient MATLAB code. This is not a trivial task because many decisions such as which triangles have to be divided into how many new ones and how these new triangles are chosen have to be encoded as vectors. We use the `htucker` toolbox 1.2 [76] for computations with hierarchical Tucker tensors and choose a "TT-like" dimension tree, the `TT-Toolbox` 2.2 [89] for Tensor Train tensors and an own implementation for the conversion between the two formats, see Subsection 2.1.3. This has the advantage that we can benefit from algorithms in both toolboxes. AMEn [36] including operators in the {d, R}-format from the `tamen`-package [35] is used for the rank-adaptive solution of linear systems. Componentwise multiplication operators and a preconditioner are implemented as described in Subsection 2.1.3. These extensions of AMEn are needed due to the nonlinear term in the PDE and for minimizing the $\hat{\mathbf{A}}_{\mathrm{ref}}^{-1}$-norm of the residual, see Theorem 7.12.

Algorithm 1 is implemented such that it can work with "Hilbert space objects", i. e., MATLAB objects for which the usual vector space operations are defined and an inner product is given. Additionally, we pay attention to the fact that the algorithm shall be used for optimal control applications. Therefore, the objective function and gradient evaluations accept initializations for the state and the adjoint state, a refined version of which is returned and used for further computations. All relevant functions, such as the control, the state, and the adjoint state are kept on matched grids to facilitate the computation. We keep all grid refinements stemming from the gradient computation and the projected linesearch for further iterations. In the concrete application it is found that sometimes it is necessary to evaluate the objective function up to high accuracy so that the state needs to be computed on a very fine FE mesh and/or with a high polynomial degree. Keeping this fine FE and polynomial spaces for further iterations would make the algorithm slow so that it turns out that it is better to discard them. We only reuse grids and polynomial degrees in the stochastic case for the objective function evaluation at the possible next iterate. To find a possibly better step than the generalized Cauchy step found by the projected Armijo linesearch with suitably refined projection, our implementation of Algorithm 1 requires an "advanced solver" for the trust-region subproblem. Here we apply a semismooth Newton method as outlined in Section 6.4 with the generalized Cauchy point as initial iterate. We perform at most 5 semismooth Newton steps and stop the iteration earlier if the discrete $L^2$-norm of the residual (6.29) is smaller than $10^{-6}$. For computing the semismooth Newton direction, at most 20 CG iterations are performed until the relative CG-residual is smaller than 0.1. If CG performs 20 iterations without reaching this tolerance and the relative CG-residual is in fact larger than 0.5, the algorithm stops and returns the iterate with the smallest semismooth Newton residual. Otherwise it checks whether the step computed by CG is capable of reducing the semismooth Newton residual. If this is not the case, the semismooth Newton iteration stops.

Note that for the performance of the overall algorithm, the semismooth Newton solver is not a bottleneck because it works on a fixed, discrete subspace of $U$. Compared to it, grid refinement and the solution of the stochastic PDE with low-rank tensors is costly. That is the reason why we use more than one semismooth Newton iteration. With this approach we can expect to obtain a larger model decrease $\mathrm{pred}_k$. This allows then to evaluate the objective function less exactly, see Lemma 4.25, so that typically more computing time can be saved than by taking less semismooth Newton iterations.

To track the mesh refinement, we write MATLAB classes for FE functions carrying, e. g., the mesh, the respective assembled matrices, the coefficients, and functions for grid refinement. Vector space operations are overloaded such that the trust-region algorithm can work with these objects. For the state and the adjoint state we need an additional MATLAB class of functions of the form (6.8) with one FE space mode and $m \in \mathbb{N}$ polynomial parameter modes. The coefficients are represented by an HT tensor and we implement functions for FE and polynomial refinement, pointwise evaluations and computing the expectation etc.

To make sure that all required error bounds in Algorithm 1 can be satisfied, we use the error functions proposed in Section 4.3. The parameter settings for the trust-region algorithm as well as the inexact projection and the projected linesearch are listed in Table 8.1.

Algorithm 1, see Sections 4.1 and 4.3:

| | |
|---|---|
| $\tau$ | $\tau = \frac{1}{\gamma} = 10^3$ |
| $\varrho_\mathrm{c}$ | $\varrho_\mathrm{c}(t) = \mathfrak{c}_\mathrm{c} t$ with $\mathfrak{c}_\mathrm{c} = c_\mathrm{g}\tau\tilde{\mathfrak{c}}_\mathrm{c}$, where $\tilde{\mathfrak{c}}_\mathrm{c} > 0$ is chosen problem-specific |
| $\varrho_\mathrm{g}$ | $\varrho_\mathrm{g}(t) = \mathfrak{c}_\mathrm{g} t$ with $\mathfrak{c}_\mathrm{g} = c_\mathrm{g}\tilde{\mathfrak{c}}_\mathrm{g}$, where $\tilde{\mathfrak{c}}_\mathrm{g} > 0$ is chosen problem-specific |
| $\varrho_{\mathrm{t}1}$ | $\varrho_{\mathrm{t}1}(t) = \mathfrak{c}_{\mathrm{t}1} t$ with $\mathfrak{c}_{\mathrm{t}1} = \frac{\mathfrak{c}_\mathrm{i}\mathfrak{c}_{\mathrm{l}1}^2\mathfrak{c}_{\mathrm{l}2}\mathfrak{c}_\mathrm{e}\mathfrak{c}_\mathrm{f}}{\tau}$, see Lemma 4.20 |
| $\varrho_{\mathrm{t}2}$ | $\varrho_{\mathrm{t}1}(t) = \mathfrak{c}_{\mathrm{t}2} t$ with $\mathfrak{c}_{\mathrm{t}2} = \frac{1}{\mathfrak{c}_{\mathrm{l}2}\tau} \cdot \min\{\frac{2(1-\mathfrak{c}_\mathrm{e})\mathfrak{c}_\mathrm{i}}{\mathfrak{c}_m}, \frac{\mathfrak{c}_\mathrm{a}}{\mathfrak{c}_\mathrm{f}}\}$, see Lemma 4.20 |
| $\varrho_\mathrm{r}$ | $\varrho_\mathrm{r}(t) = 2c_\mathrm{o}\tilde{\mathfrak{c}}_\mathrm{o} t^{\mathfrak{c}_\mathrm{o}}$ with $\tilde{\mathfrak{c}}_\mathrm{o} > 0$ chosen problem-specific and $\mathfrak{c}_\mathrm{o} = 1.1$ |
| $(\mathfrak{r}_k)_{k\in\mathbb{N}_0}$ | $\mathfrak{r}_k = \frac{1000}{k+1}$ |
| $\Delta$ | $\Delta_\mathrm{max} = 10^4$, $\Delta_0 = 1$ |
| $\eta_i$ | $\eta_1 = 0.3$, $\eta_2 = 0.7$, $\eta_3 = 0.2$ |
| $\nu_i$ | $\nu_1 = 0.5$, $\nu_2 = 1.0$, $\nu_3 = 2.0$ |
| $c_\chi\varepsilon_\mathrm{tol}$ | $c_\chi\varepsilon_\mathrm{tol} = 10^{-4}$ (deterministic problems), $c_\chi\varepsilon_\mathrm{tol} = 10^{-3}$ (stochastic problems) |

Inexact projection and projected linesearch, see Section 4.3:

| | |
|---|---|
| $\mathfrak{c}_\mathrm{s}$ | $\mathfrak{c}_\mathrm{s} = 0$ (exact projection used for computing the inexact criticality measure) |
| $\mathfrak{c}_\mathrm{i}$ | $\mathfrak{c}_\mathrm{i} = 0.5$ |
| $\mathfrak{c}_\mathrm{f}$, $\mathfrak{c}_\mathrm{e}$, $\mathfrak{c}_\mathrm{a}$ | $\mathfrak{c}_\mathrm{f} = 0.5$, $\mathfrak{c}_\mathrm{e} = 10^{-2}$, $\mathfrak{c}_\mathrm{a} = 10^{-3}$ |
| $\mathfrak{c}_{\mathrm{a},k}$ | $\mathfrak{c}_{\mathrm{a},k} = \max\{\mathfrak{c}_\mathrm{a}, \min\{\tau, \frac{\mathfrak{c}_{\mathrm{l}2}\Delta_k}{\|\nabla m_k(0)\|_U}\}\}$, see (4.40) |
| $\mathfrak{c}_{\mathrm{l}1}$ | $\mathfrak{c}_{\mathrm{l}1} = 0.7$ |
| $\mathfrak{c}_{\mathrm{l}2}$ | $\mathfrak{c}_{\mathrm{l}2} = \frac{2\mathfrak{c}_{\mathrm{l}1}-1}{\mathfrak{c}_{\mathrm{l}1}} = \frac{4}{7}$, see Remark 4.21 |
| $\mathfrak{c}_\mathrm{d}$ | $\mathfrak{c}_\mathrm{d} = 10^{-2}$ |

Semismooth Newton, see Sections 4.4 and 6.4:

| | |
|---|---|
| $\mathfrak{c}_{\mathrm{n},k}$ | $\mathfrak{c}_{\mathrm{n},k} = \mathfrak{c}_\mathrm{n}$ (constant over all iterations), problem-specific |

Table 8.1.: Parameters used in Algorithm 1

In particular, we stop the algorithm if $\chi_k(0) < 10^{-4}$ or $\chi_k(0) < 10^{-3}$ holds in the deterministic or in the stochastic case, respectively. Note that in the unconstrained case this would correspond to having $\|\nabla m_k(0)\| < 10^{-7}$ or $\|\nabla m_k(0)\| < 10^{-6}$ because $\tau = 10^3$ is chosen, cf. (4.9). Almost all parameters have the same value over all tests, but the error bound parameters $\tilde{\mathfrak{c}}_\mathrm{c} > 0$, $\tilde{\mathfrak{c}}_\mathrm{g} > 0$, and $\tilde{\mathfrak{c}}_\mathrm{o} > 0$ for the criticality measure, the gradient, and the objective function evaluation, respectively, are chosen problem-specific because they reflect how good the applied error estimators are or compensate unknown constants. This makes some experimentation necessary. On the one hand, one should not choose the parameters $\tilde{\mathfrak{c}}_\mathrm{c}$ and $\tilde{\mathfrak{c}}_\mathrm{g}$ too small because grid refinement happens early then and makes the algorithm slow. On the other hand, if $\tilde{\mathfrak{c}}_\mathrm{c}$ and $\tilde{\mathfrak{c}}_\mathrm{g}$ are too large, the trust-region model is not accurate enough to compute a direction of objective function decrease. Then, unsuccessful iterations happen and the trust-region radius is decreased until the gradient is accurate enough. This can take some time and should be avoided. A similar situation appears when choosing $\tilde{\mathfrak{c}}_\mathrm{o}$. If it is too large, a very inexact objective function evaluation can cause an unsuccessful iteration. Conversely, if it is too small, too much time is spent on the objective function evaluations. Additionally, we sometimes adapt the regularization parameter $\mathfrak{c}_{\mathrm{n},k}$ appearing in the semismooth Newton problem (4.42), but choose it constant over all iterations. For example, it can be increased to have a better condition number and typically better convergence of the semismooth Newton iteration, but should not be too large because then the computed steps are too small.

The adaptive solution of the deterministic PDE (Section 7.1) is implemented as follows: The discretized PDE is always solved by Newton's method until the residual norm is below $10^{-6}$. Often, we only need one Newton step to fulfill this criterion. This tolerance is chosen such that the FE error dominates in Theorem 7.8 during the whole computation. The triangles contributing 30% of the error are refined. The procedure is repeated until the error estimate given in Theorem 7.8 is small enough. To avoid an infeasible number of FE nodes and unknowns, the refinement is stopped if $2 \cdot 10^5$ FE nodes would be exceeded.

The stochastic PDE is solved adaptively based on Theorem 7.12. We start with the coarse FE mesh shown in Figure 8.1 and with the polynomial degrees $d = \mathbb{1}$, and refine iteratively. Depending on the highest error contribution in (7.32), one of the following tasks is pursued:

- If the algebraic error dominates, one inexact Newton step is computed by performing one AMEn sweep on the Newton equation. To avoid rank growth, the updated tensor is truncated to rank at most 100 w. r. t. to the $\hat{\mathbf{A}}_\mathrm{ref}$-norm as described in Section 7.4.
- If the polynomial discretization error is the largest one, the polynomial degree of one parameter, namely the one with the largest error contribution, is increased by 1.
- If the FE error dominates, the triangles contributing 30% of the error are refined. To make the evaluation of the FE error efficient, we have to round the componentwisely multiplied tensors to avoid too large tensor ranks.

The refinement is programmed such that the last iteration is always an AMEn sweep to have a meaningful solution on the current discrete subspace. Again, the number of FE nodes is bounded by $2 \cdot 10^5$ in order to avoid complexity issues so that the algorithm adapts to the semi-discrete solution using the finest FE mesh at the end. In some cases, AMEn stagnates and does not manage to compute a solution with small enough algebraic error in the last iterations of Algorithm 1. Then the adaptive solution procedure is stopped.

The projection onto the box $U_{\text{ad}}$ is computed adaptively by computing the exact projection and the corresponding refined grid first. Then the exact error made by the node-wise projection can be evaluated on each element. If the node-wise projection does not meet one of the exactness requirements, the triangles contributing 30% of the projection error are refined. We use uniform refinement to not destroy the quality of the mesh, which is important to uniformly bound the constants $c_{\mathcal{T}}$ in (7.16) and (7.32), which depend on the smallest angle in the triangulations. The refinement is repeated until the inexact projection is accurate enough.

All computations are run on a Linux cluster with 1 TB RAM and 4 Intel Xeon E7-8857 processors, each of which has 12 kernels and a base frequency of 3.00 GHz. However, only serial but vectorized implementations are used in MATLAB because current toolboxes do not offer, e. g., parallel computations with low-rank tensors or parallel FE codes.

## 8.2. Results for the Deterministic Problem

First, we implement the deterministic problem, i. e., the optimal control problem with the objective function $J[\xi]$ (3.4) and the state equation (3.8), where the reference coefficient $\xi = \bar{\xi} = 0$ is inserted. Equivalently, we could solve the stochastic problem with $\sigma = 0$. We do this in order to test the error estimates from Sections 5.1 and 7.1, to visualize the adaptivity of the FE grid and to have reference solutions for the problem under uncertainty.

Since the discretized PDE is often solved to good accuracy by only one Newton step, there is no need to balance the errors in Theorem 7.8. Therefore, we choose $c_{\mathcal{T}} c_{A_{\text{ref}}} = 1$ in (7.13) and adapt the error estimation parameters adequately.

The first setup uses the desired state $\hat{q} \equiv 1$. Here we choose the parameters $\tilde{\mathfrak{c}}_{\text{c}} = 200$, $\tilde{\mathfrak{c}}_{\text{g}} = 200$, $\tilde{\mathfrak{c}}_{\text{o}} = 10^8$, and $\mathfrak{c}_{\text{n}} = 0.1$. The high value for $\tilde{\mathfrak{c}}_{\text{o}}$ is needed to avoid too severe grid refinement due to inexact objective function evaluation. This value comes also from the fact that we overestimate the error given in Proposition 5.5, because we measure the state error in the $H_0^1(\Omega)$-norm although the $L^2(\Omega)$-error would be sufficient. The computed optimal state and and the control are depicted in Figure 8.2. It can be observed that the optimal state is in the order of magnitude of the desired state. Furthermore, the three active sets of the control and the symmetry of the problem can be recognized. The FE mesh obtained in the final iteration of the algorithm is shown in Figure 8.3. Around the corner $x = 0$, the mesh is refined locally due to the adaptive solution of the PDE. Furthermore, the boundary of the active set of the control is resolved by the mesh due to the refined projection.

The convergence is shown in Figure 8.4. This is the typical convergence behavior so that we do not show convergence plots for the other deterministic setups. The number of FE nodes increases until it reaches the allowed upper bound. The criticality measure decreases in general until the desired tolerance is reached. But by refining the mesh the inexactness of the computed criticality measure is recognized sometimes. Then the computed criticality measure on the refined mesh is larger than the one on the coarser grid. We have to mention that it is not worth counting iterations in this setting. Typically, the first iterations run very fast within some seconds. For instance, the first 16 iterations needed to decrease the computed optimality measure from 23.2 to $3.09 \cdot 10^{-2}$ take about 19 seconds in total. Only the last iterations required to obtain the desired high accuracy last increasingly longer so that

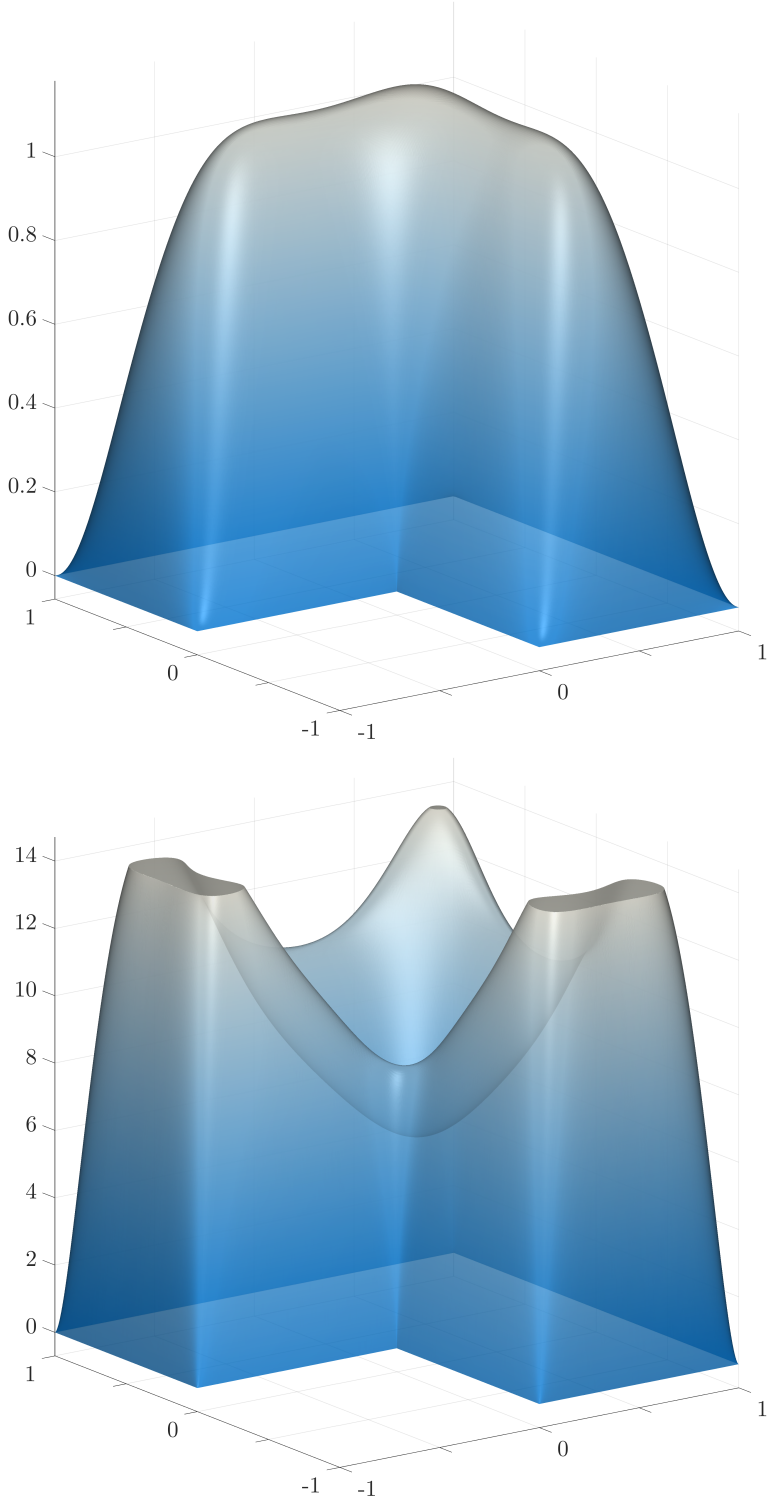



Figure 8.2.: Optimal state (top) and optimal control (bottom) for the deterministic problem with $\hat{q} \equiv 1$.

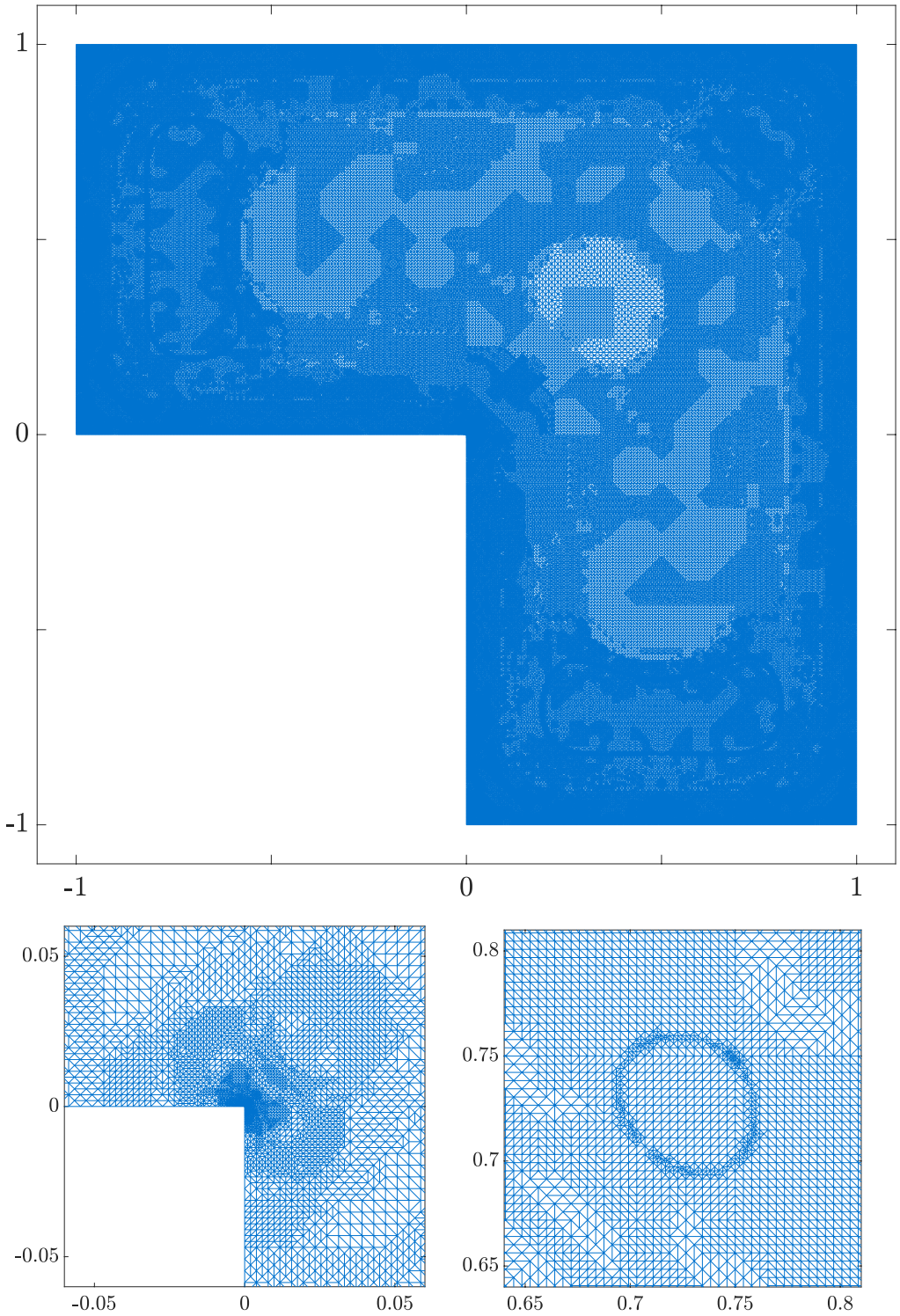



Figure 8.3.: Final mesh (with details at the bottom) for the deterministic problem with $\hat{q} \equiv 1$.

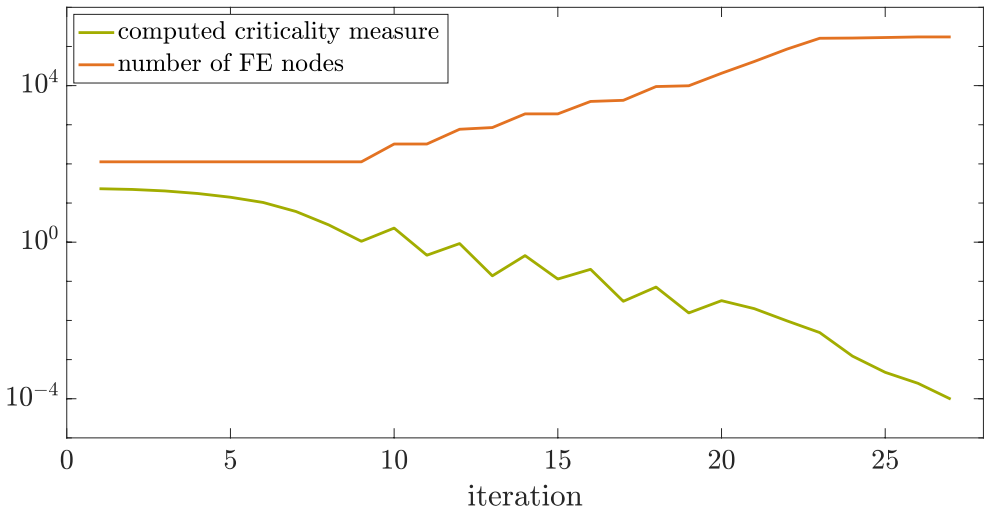



Figure 8.4.: Convergence and refinement plot for the deterministic problem with $\hat{q} \equiv 1$.

the total computing time is around 22 minutes. The last three iterations, during which the computed criticality measure is decreased from $1.23 \cdot 10^{-3}$ to $0.98 \cdot 10^{-4}$, take approximately 13 minutes. Often, the grid refinement due to the projection errors is done to more extent during the last iterations. Figure 8.5 shows the grid 5 iterations before the last one. It has approximately half the number of FE nodes compared to the one shown in Figure 8.3 and the resolution of the active sets is still missing.

As a second setup, we change the desired state to the polynomial $\hat{q}(x) = 9(x_1^3 - x_1)(x_2^3 - x_2)$. This is a smooth function fulfilling the zero boundary conditions on $\Omega$. Additionally, it is very smooth around the corner point $x = 0$. We do not change the parameters compared to the first setup. The computed optimal state and control are depicted in Figure 8.6. The active set of the control is also nicely resolved in the final mesh (Figure 8.7). Around the corner $x = 0$, the mesh is not locally refined as before because of the smoothness of the optimal state around this point.

In the third setup, the desired state is $\hat{q} \equiv 1_{\{x \in \Omega : x_2 > 0.5\}}$. Here, it is necessary to choose the parameters $\tilde{\mathfrak{c}}_c = 20$, $\tilde{\mathfrak{c}}_g = 20$, $\tilde{\mathfrak{c}}_o = 10^8$, and $\mathfrak{c}_n = 1.0$. The gradient is computed more exactly by this choice so that that the algorithm does not run into unsuccessful iterations. The larger regularization parameter for semismooth Newton yields a more robust iteration. In this case, the optimal state is almost zero for $x_2 < 0$, see Figure 8.8, and also the optimal control has very small values in this area. By the first-order optimality condition, the adjoint state is also small there. This yields that the final mesh is quite coarse for $x_2 < 0$ except for the area around the corner $x = 0$, see Figure 8.9. This can be explained by taking a look at the error estimates (7.14) and (7.15). They become small on the triangles and their respective edges if the solution $\tilde{y}$ and the jump in its gradient, the residual w and the jump in its gradient, and the right-hand side $\hat{f}$ are small. This is true for both the state and the adjoint equation. Again, the active set of the control is resolved nicely, cf. Figures 8.8 and 8.9.

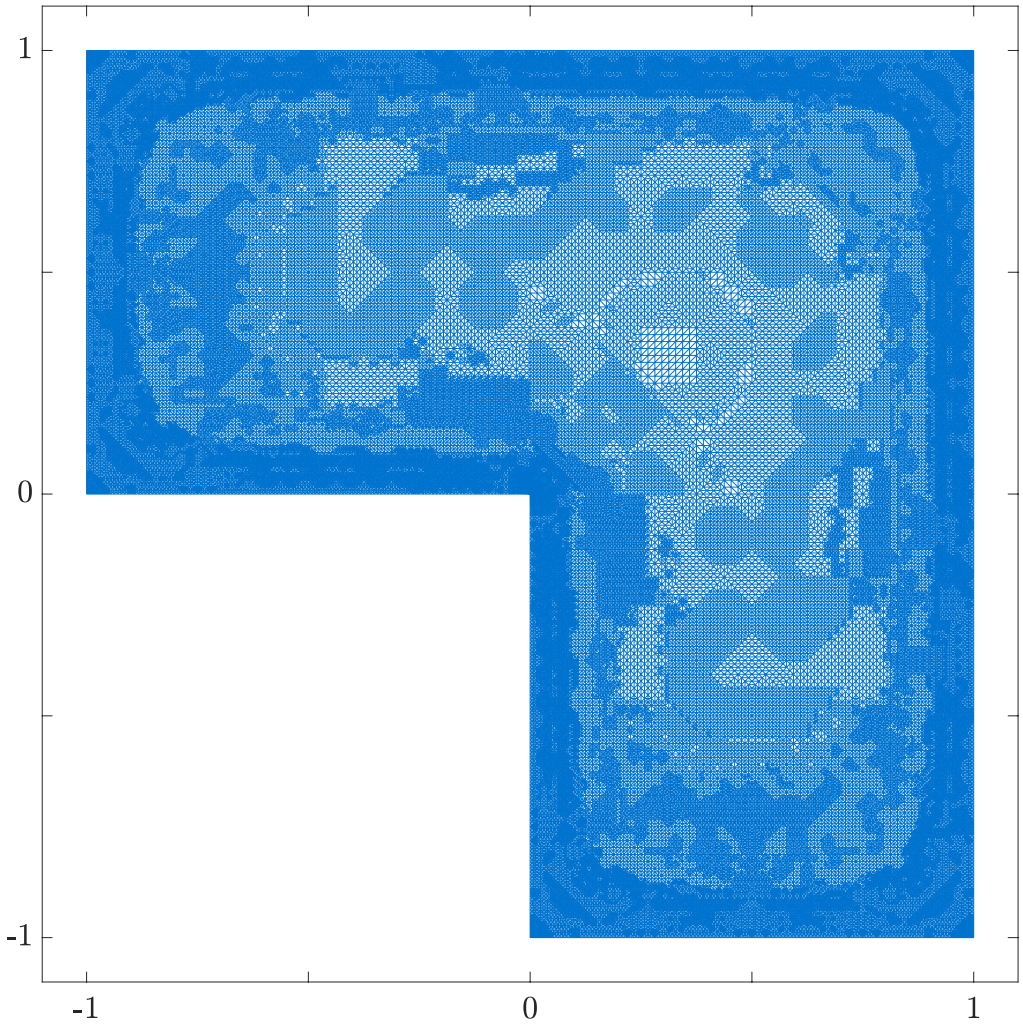



Figure 8.5.: Mesh from iteration 22 (of 27) for the deterministic problem with $\hat{q} \equiv 1$.

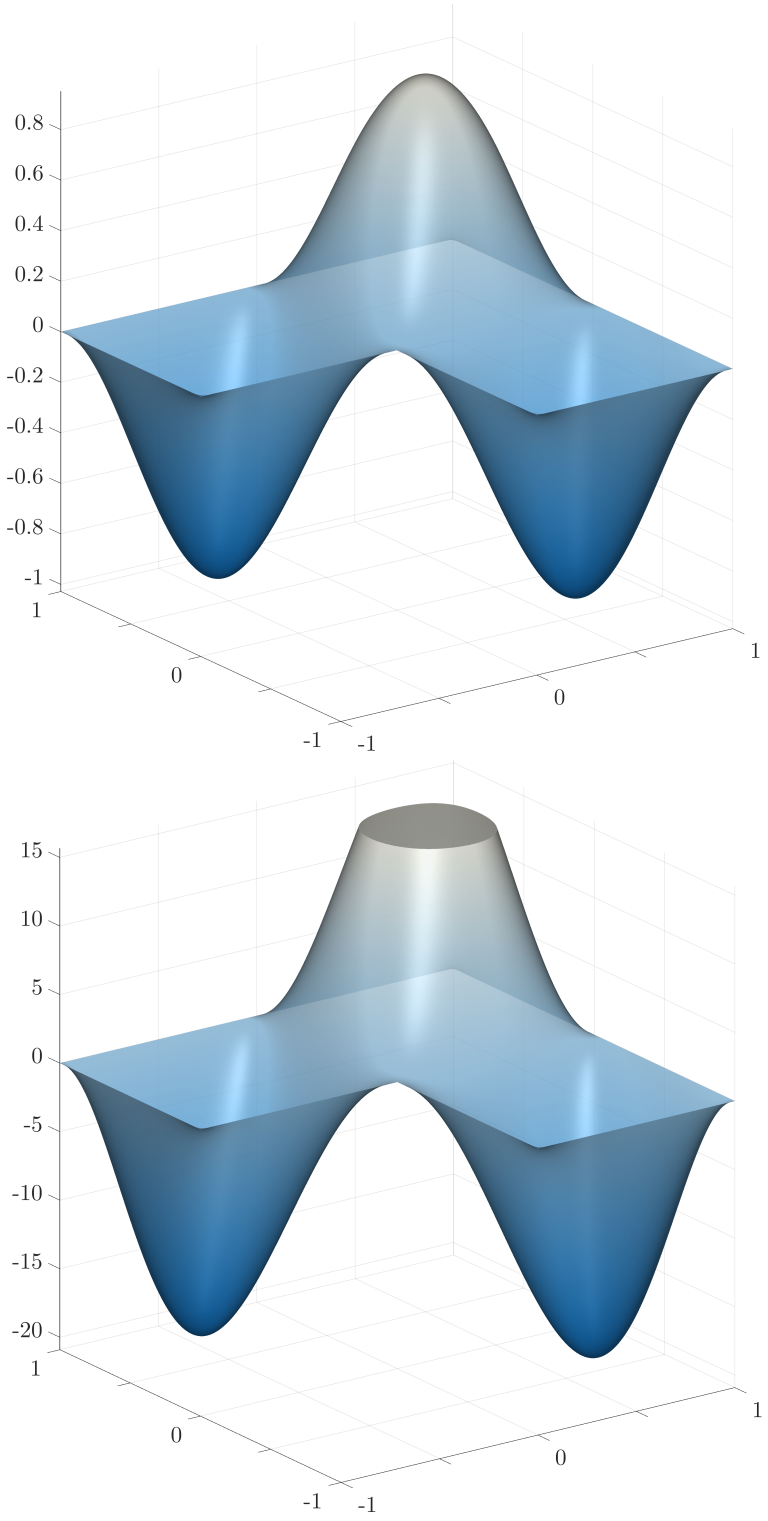



Figure 8.6.: Optimal state (top) and optimal control (bottom) for the deterministic problem with $\hat{q}(x) = 9(x_1^3 - x_1)(x_2^3 - x_2)$.

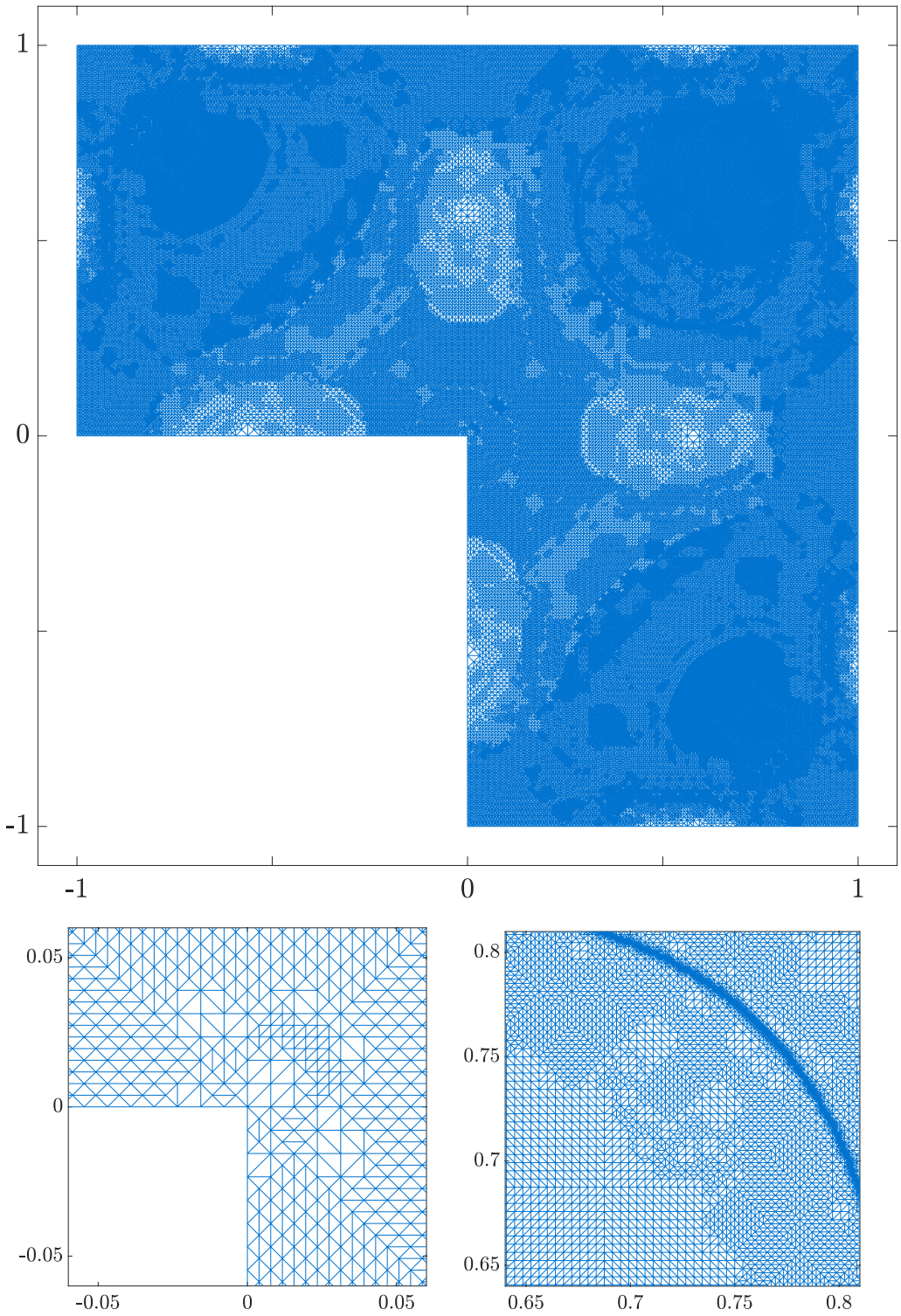



Figure 8.7.: Final mesh (with details at the bottom) for the deterministic problem with $\hat{q}(x) = 9(x_1^3 - x_1)(x_2^3 - x_2)$.

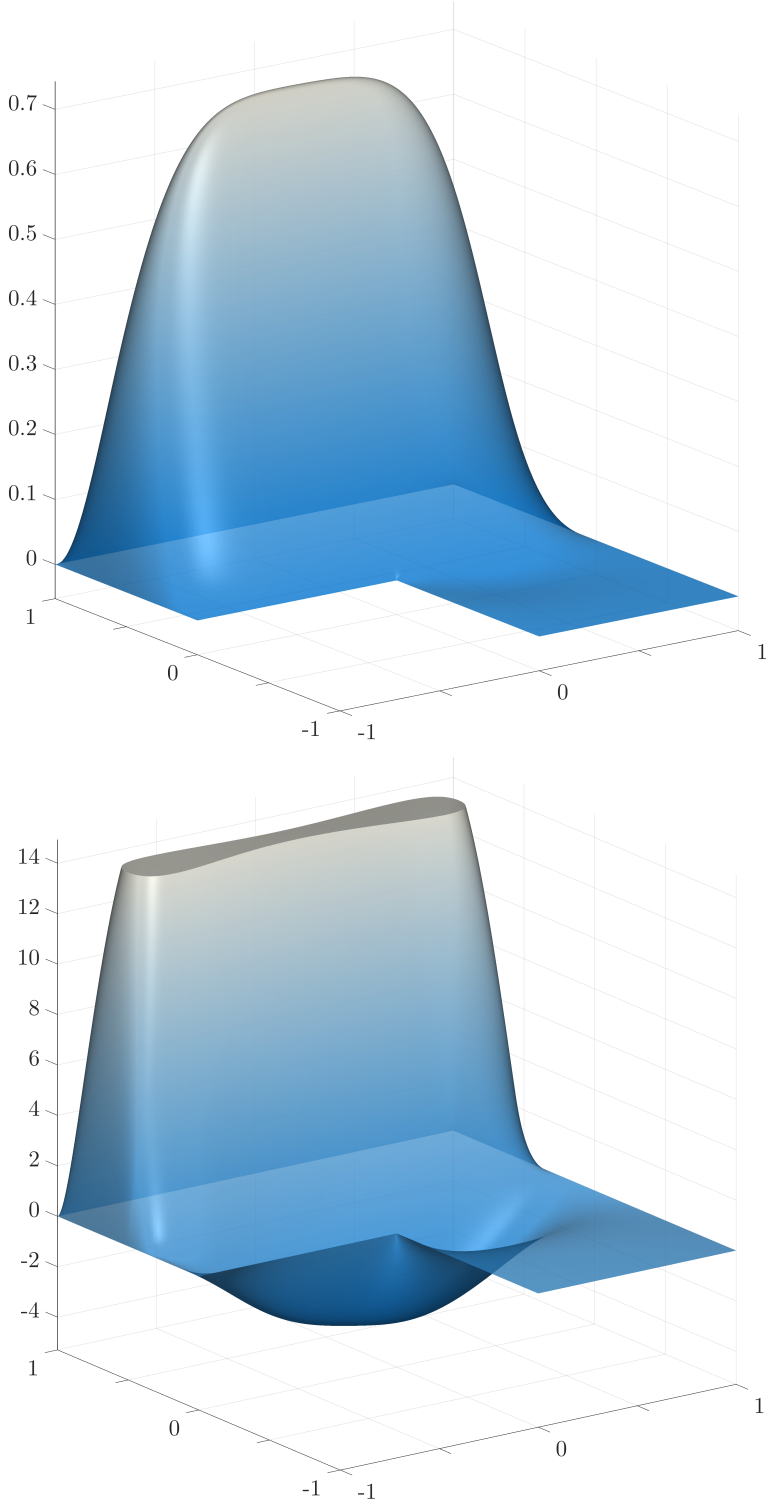



Figure 8.8.: Optimal state (top) and optimal control (bottom) for the deterministic problem with $\hat{q} \equiv 1_{\{x \in \Omega:\, x_2 > 0.5\}}$.

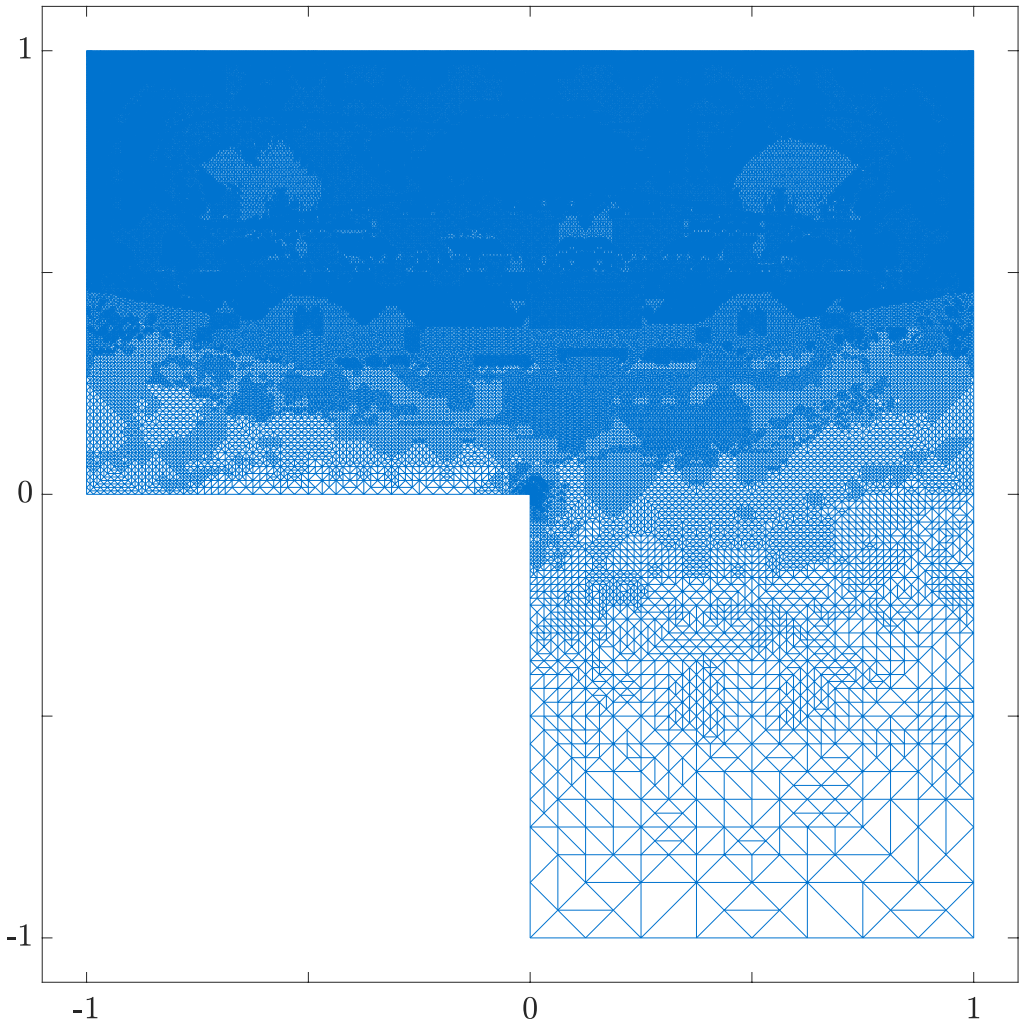



Figure 8.9.: Final mesh for the deterministic problem with $\hat{q} \equiv 1_{\{x \in \Omega:\, x_2 > 0.5\}}$.

## 8.3. Results for the Problem with Uncertainties

Now we consider the stochastic problem (3.3) with the state equation defined in (3.11) and vary the problem data to see how the algorithm adapts to the situation. As already noted, we choose $c_{\mathcal{T}} c_{A_{\mathrm{ref}}} = 10^{-3}$ to balance the error contributions coming from the FE and the polynomial chaos discretization. This makes different choices of the error control constants necessary.

Again, we start with the desired state $\hat{q} \equiv 1$ and with the vector of influence amounts $\sigma = (0.25, 0.25, 0.25, 0.25, 0.25, 0.25)^\top$, i. e., the coefficient function varies $\pm 25\%$ equally on each subdomain independently and each parameter has the same influence on it. We choose $\tilde{\mathfrak{c}}_{\mathrm{c}} = 0.05$, $\tilde{\mathfrak{c}}_{\mathrm{g}} = 0.05$, $\tilde{\mathfrak{c}}_{\mathrm{o}} = 10^3$, and $\mathfrak{c}_{\mathrm{n}} = 0.01$. The obtained optimal control and the expected optimal state are quite similar to the ones from the deterministic setup, see Figure 8.10, but the active sets are a bit smaller so that the smallest one almost disappears. The final mesh (Figure 8.11) is very similar to the one obtained before. Grid refinements due to the active sets can be recognized, but are not as pronounced as in the deterministic case because the iteration is already stopped when $\chi_k(0) < 10^{-3}$ instead of $\chi_k(0) < 10^{-4}$. The reason for this is that the high fidelity PDE solves in the last iterations become quite expensive. The algorithm takes about 42 hours to compute this solution, where 37 hours are spent for the last two iterations decreasing the computed criticality measure from $1.39 \cdot 10^{-2}$ to $5.82 \cdot 10^{-4}$. In practice, one would probably refrain from performing the last two iterations because less accurate solutions are sufficient if, e. g., the problem data or the parameter distribution is not known exactly. The convergence and refinement behavior is shown in Figure 8.12. We see that the computed criticality measure decreases monotonically and the algorithm stops in iteration 13, whereas a criticality measure smaller than $10^{-3}$ is obtained in iteration 25 in the deterministic case, see Figure 8.4. This is because the error control constants $\tilde{\mathfrak{c}}_{\mathrm{c}}$ and $\tilde{\mathfrak{c}}_{\mathrm{g}} = 0.05$ are chosen a bit smaller even if we take account of the smaller constant $c_{\mathcal{T}} c_{A_{\mathrm{ref}}}$ so that better steps are computed. The number of FE nodes grows until it reaches its upper bound. Furthermore, the polynomial degrees for each parameter increase almost equally during all iterations because the parameters have roughly the same influence on the optimal uncertain state. Figure 8.14 shows the difference between the deterministic and the robust control. We see the symmetry of the problem and can recognize the boundaries of the parameter influence sets $\Omega_i$. The difference between the controls is zero where their active sets coincide.

The situation changes if we keep the desired state $\hat{q} \equiv 1$, but choose the influence amounts $\sigma = (0.05, 0.10, 0.20, 0.30, 0.40, 0.45)^\top$. The average influence of each uncertain parameter is still 25%, but now it is increasing clockwise, i. e., parameter 1 has the smallest influence of 5% on the coefficient function, whereas parameter 6 has the largest influence of 45%. The convergence plot (Figure 8.13) is changed by this choice in the sense that higher polynomial degrees are needed for the parameters with more influence on the system. This becomes clear if we review the stochastic error indicator (7.31), which becomes smaller if the coefficient function $\kappa_i$ is smaller. Additionally, it can be seen in Figure 8.15 that the difference between the deterministic and the robust control is larger in areas where the variation of the coefficient function is larger. Clearly, the problem is not symmetric anymore and one can recognize the size change of the active sets, especially the one on the bottom right.

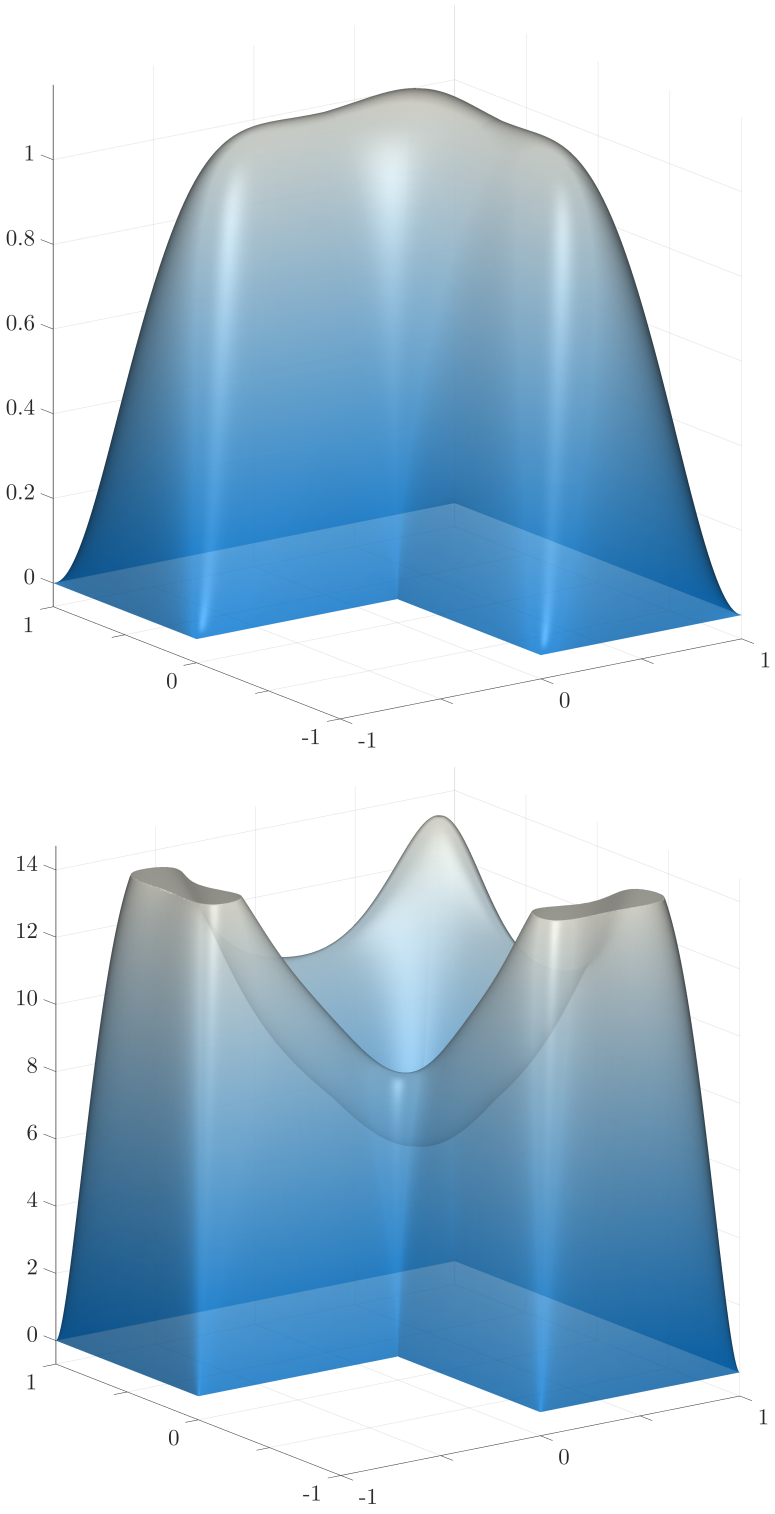



Figure 8.10.: Optimal control (bottom) and expected optimal state (top) for the stochastic problem with $\hat{q} \equiv 1$ and $\sigma = (0.25, 0.25, 0.25, 0.25, 0.25, 0.25)^\top$.

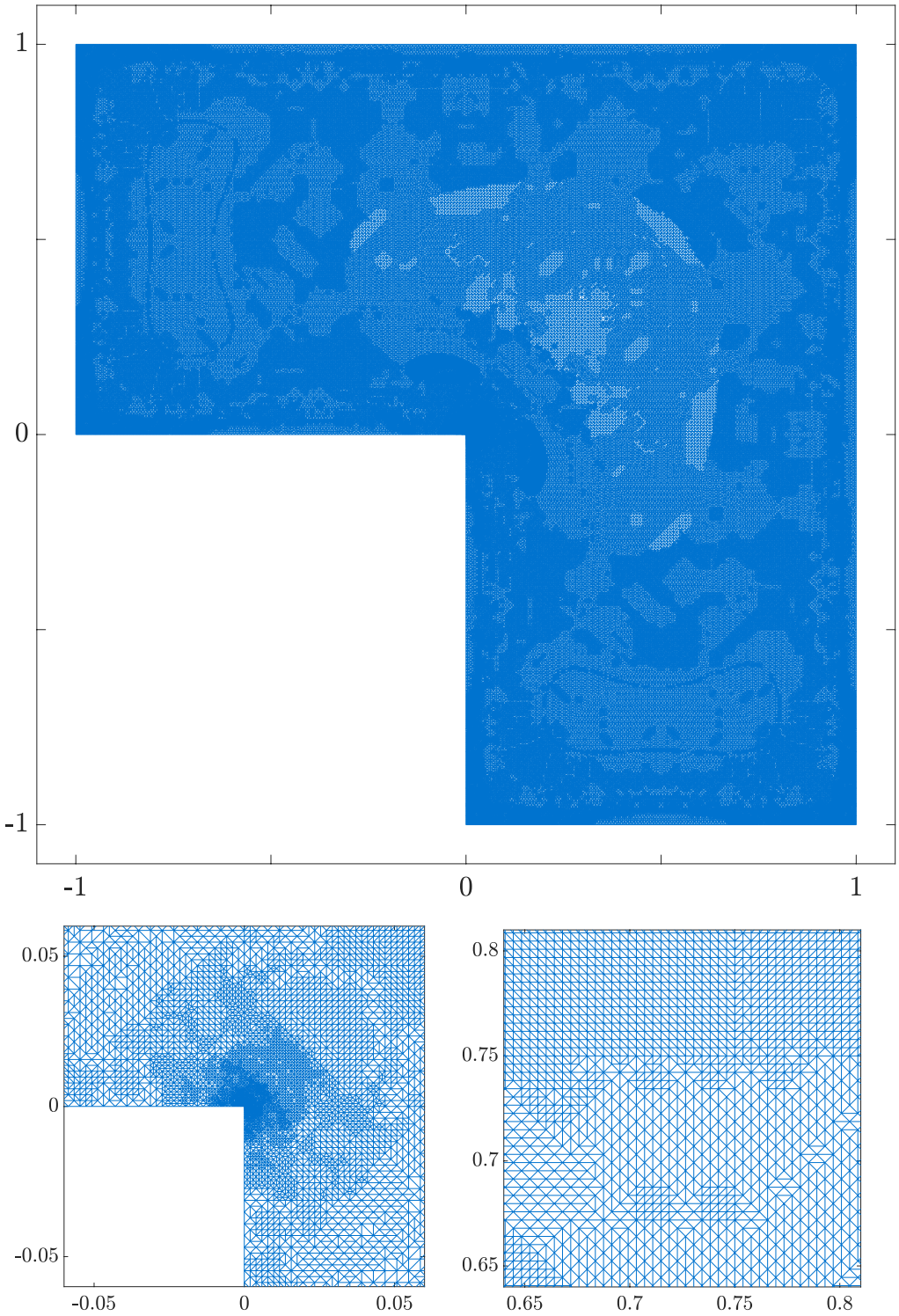



Figure 8.11.: Final mesh (with details at the bottom) for the stochastic problem with $\hat{q} \equiv 1$ and $\sigma = (0.25, 0.25, 0.25, 0.25, 0.25, 0.25)^\top$.

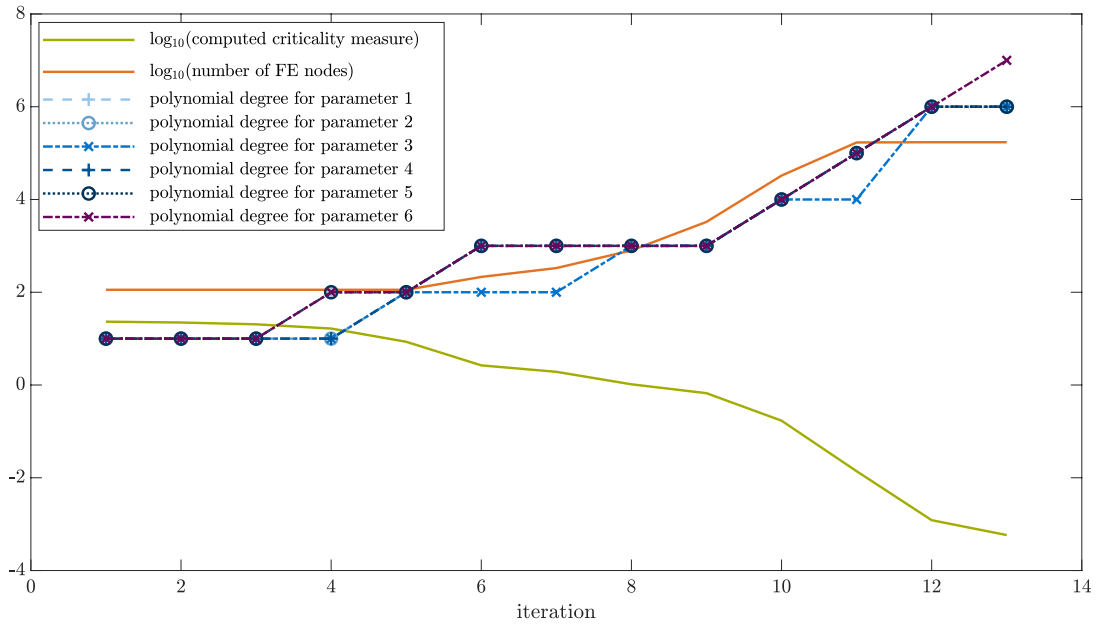



Figure 8.12.: Convergence and refinement plot for the stochastic problem with $\hat{q} \equiv 1$ and $\sigma = (0.25, 0.25, 0.25, 0.25, 0.25, 0.25)^\top$.

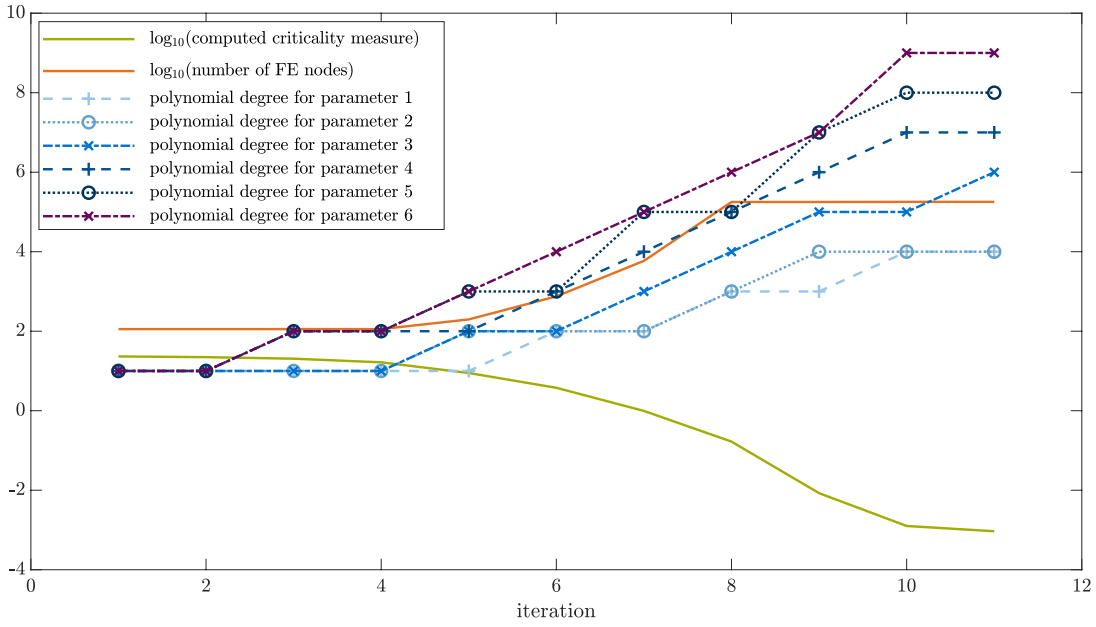



Figure 8.13.: Convergence plot and refinement for the stochastic problem with $\hat{q} \equiv 1$ and $\sigma = (0.05, 0.10, 0.20, 0.30, 0.40, 0.45)^\top$.

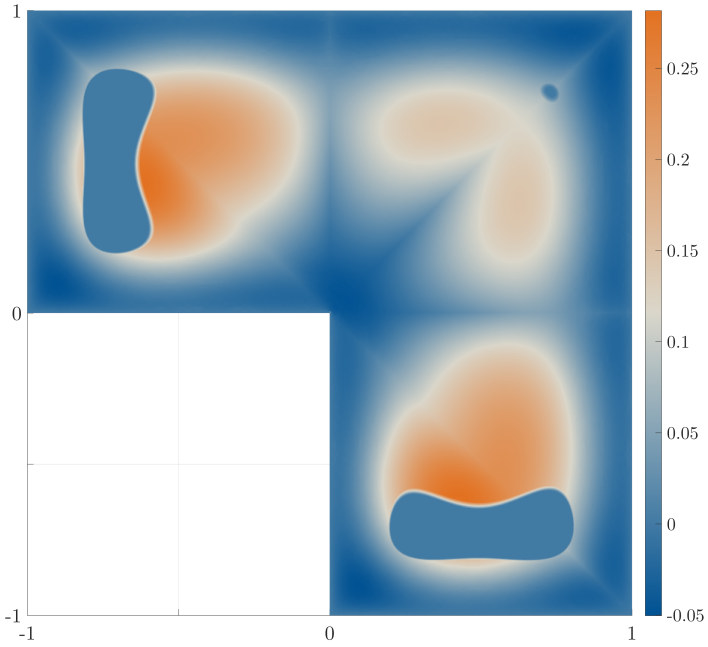



Figure 8.14.: Difference between the deterministic and the robust control for $\hat{q} \equiv 1$ and $\sigma = (0.25, 0.25, 0.25, 0.25, 0.25, 0.25)^\top$.

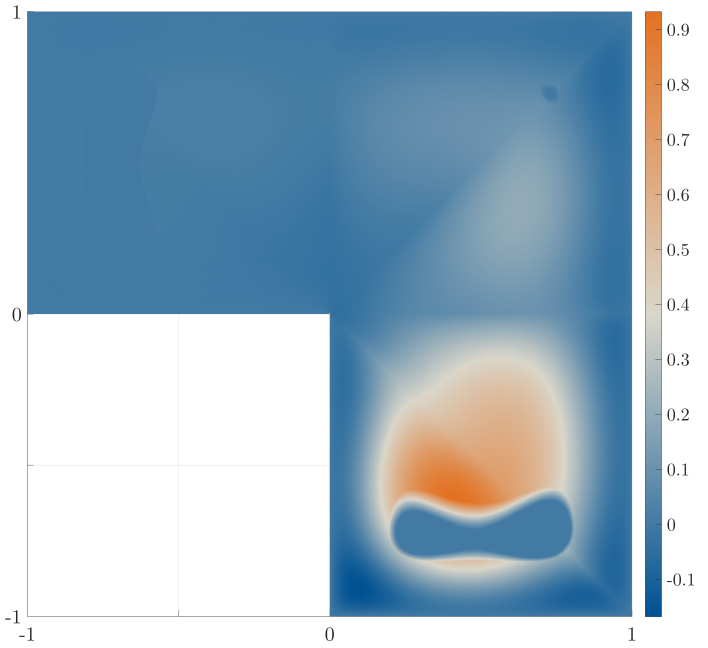



Figure 8.15.: Difference between the deterministic and the robust control for $\hat{q} \equiv 1$ and $\sigma = (0.05, 0.10, 0.20, 0.30, 0.40, 0.45)^\top$.

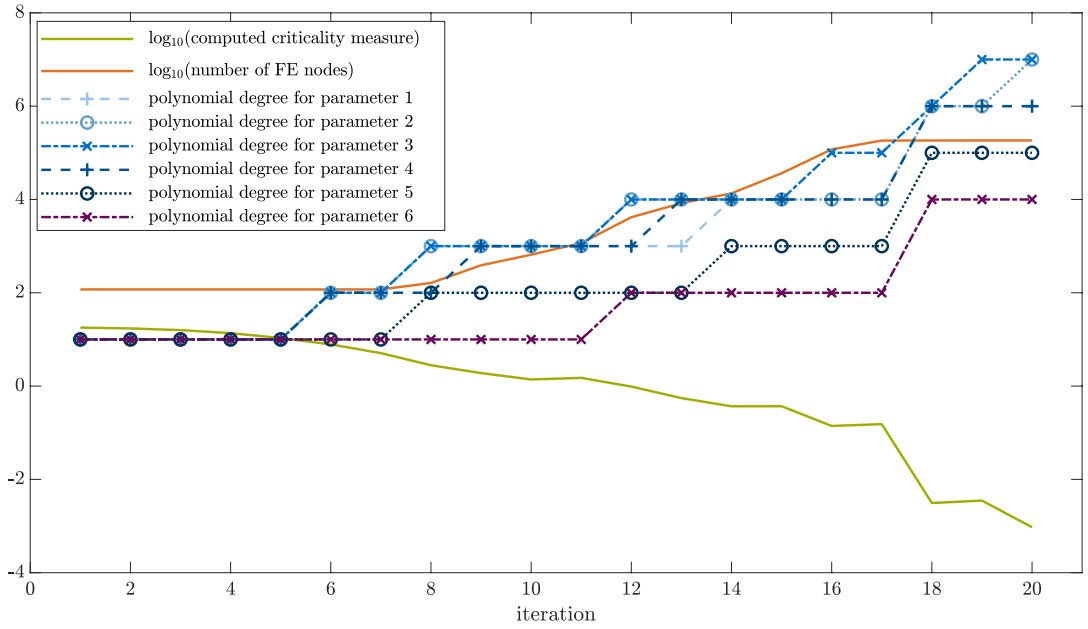



Figure 8.16.: Convergence and refinement plot for the stochastic problem with $\hat{q} \equiv 1_{\{x \in \Omega:\, x_2 > 0.5\}}$ and $\sigma = (0.25, 0.25, 0.25, 0.25, 0.25, 0.25)^\top$.

As a last setup, we consider the desired state $\hat{q} \equiv 1_{\{x \in \Omega:\, x_2 > 0.5\}}$ with equal influence $\sigma = (0.25, 0.25, 0.25, 0.25, 0.25, 0.25)^\top$. We choose the parameters $\tilde{\mathfrak{c}}_\mathrm{c} = 0.01$, $\tilde{\mathfrak{c}}_\mathrm{g} = 0.01$, $\tilde{\mathfrak{c}}_\mathrm{o} = 10^3$, and $\mathfrak{c}_\mathrm{n} = 0.1$ in analogy to the choice in the deterministic setup. Again, we obtain a similar optimal control and expected state. The convergence and refinement plot in Figure 8.16 reveals that smaller polynomial degrees are chosen especially for parameters 5 and 6, which act in the area where the optimal state and its gradient are almost zero. Because of this fact, the stochastic error indicator (7.31) becomes smaller for these parameters since it depends on the size of the state gradient in the area where the coefficient $\kappa_i$ is not small. The difference to the deterministic control is larger in the subdomains $\Omega_1, \ldots, \Omega_4$, see Figure 8.17, which is another indicator for the fact that parameters 5 and 6 do not influence the solution as much as parameters 1 to 4.

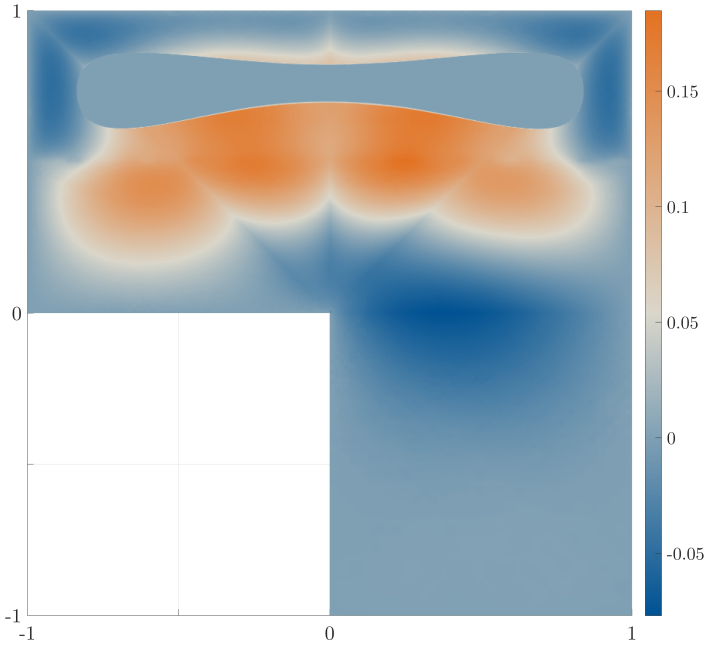



Figure 8.17.: Difference between the deterministic and the robust control for $\hat{q} \equiv 1_{\{x \in \Omega:\, x_2 > 0.5\}}$ and $\sigma = (0.25, 0.25, 0.25, 0.25, 0.25, 0.25)^\top$.

# 9. Alternative Risk Measures

Until now, all considerations in this thesis have referred to the expectation as risk measure, i. e., $\mathcal{R} \equiv \mathbb{E}$ in (1.2). This means that we want the controlled system to perform well *on average*. On the one hand, this is a viable approach if, e. g., the objective function stands for a cost one has to pay and this cost shall be small averaged over a long time horizon during which the uncertain parameters may change their value according to their distribution. On the other hand, in certain engineering applications, such as aeronautics or nuclear reactors, a high cost function value can correspond to a system fail, meaning that the plane crashes or the reactor core melts for example. Such situations should clearly be avoided with high probability. This is not taken into account if $\mathcal{R} \equiv \mathbb{E}$ is used as a risk measure because those scenarios are typically unlikely to occur so that they do not influence the expected value very much. Analogously, the expectation might not be the risk measure of choice if the randomness is only observed once and the probability of having a high cost shall be small.

In this chapter, we give an introduction to risk measures, which are functionals rating the probability distribution of a random variable, and their properties. Furthermore, we provide two examples which we discuss in the context of optimal control. Both lead to *risk-averse* choices of the optimal control instead of the *risk-neutral* controls computed with $\mathcal{R} \equiv \mathbb{E}$. The mean-variance risk measure $\mathcal{R} \equiv \mathbb{E} + \lambda \operatorname{Var}$ with $\lambda > 0$ (Section 9.2) is smooth so that the trust-region algorithm (Algorithm 1) can be applied to solve optimal control problems involving it. We discuss this in theory and show how the corresponding required error estimates (cf. Chapter 5) change in this case. The conditional value-at-risk (CVaR, Section 9.3) is a risk measure with better theoretical properties than the mean-variance risk measure and is used frequently in financial mathematics [94, 95, 77, 103]. For instance, in the example control problem from the previous chapters the mean-variance risk measure would penalize upward and downward deviations of the cost term from its mean equally although downwards deviations lead to a probably desired smaller cost. In contrast, the CVaR is computed based on large deviations from the desired state only since it is the expectation of the upper tail of the considered random variable. Since CVaR is nonsmooth, different solution techniques have to be applied for solving optimal control problems involving it. We discuss the application of a primal interior-point method and its implementation with low-rank tensors.

## 9.1. Definition and Properties of Risk Measures

**Definition 9.1.** Let $(\Xi, \mathcal{F}, \mathbb{P})$ be a complete probability space. A *risk measure* $\mathcal{R}$ is a functional $\mathcal{R} : L^p_{\mathbb{P}}(\Xi) \to \mathbb{R} \cup \{\infty\}$, where $p \in [1, \infty]$.

**Remark 9.2.** Instead of $L^p_{\mathbb{P}}(\Xi)$, we could also consider more general spaces or sets of random variables. Furthermore, the value $-\infty$ could be incorporated, see [103, Sec. 6.3].

**Example 9.3.** Prominent examples for risk measures are (see also [96, Sec. 2])

- the expectation $\mathcal{R} \equiv \mathbb{E}$, which is well-defined and always finite on $L^1_{\mathbb{P}}(\Xi)$,
- the *mean-variance risk measure* $\mathcal{R}[X] = \mathbb{E}[X] + \lambda \operatorname{Var}[X]$ for some $\lambda > 0$, which is well-defined on $L^2_{\mathbb{P}}(\Xi)$,
- the risk measure $\mathcal{R}[X] = \mathbb{E}[X] + \lambda\, \sigma[X]$ with $\sigma[X] = \sqrt{\operatorname{Var}[X]}$, which is well-defined on $L^2_{\mathbb{P}}(\Xi)$, and
- the *$\beta$-quantile* or *value-at-risk* $\mathcal{R}[X] = \operatorname{VaR}_\beta[X]$ for some confidence level $\beta \in (0,1)$, which is defined by
$$\operatorname{VaR}_\beta[X] := \min\{t \in \mathbb{R} : F_X(t) \geq \beta\}$$
[95, Def. 1] with the distribution function $F_X(t) := \mathbb{P}(X \leq t)$. Note that $\operatorname{VaR}_\beta$ is well-defined since $F_X$ is non-decreasing and right-continuous. Lastly, we have
- the *conditional value-at-risk* $\mathcal{R}[X] = \operatorname{CVaR}_\beta[X]$ defined by
$$\operatorname{CVaR}_\beta[X] := \text{“mean of the } \beta\text{-tail distribution of } X\text{”},$$
i. e., the mean of a random variable with distribution function $F^\beta_X(t) := \frac{1}{1-\beta}(F_X(t) - \beta)\, 1_{\{t \geq \operatorname{VaR}_\beta[X]\}}(t)$ [95, Def. 3]. We will see how the CVaR can be computed.

**Definition 9.4.** Let $\mathcal{R}$ be a risk measure. It may have the following properties (see [96]):

P1. *Convexity*: $\mathcal{R}[(1-\tau)X + \tau Y] \leq (1-\tau)\mathcal{R}[X] + \tau\mathcal{R}[Y]$ for all $\tau \in [0,1]$ and all $X, Y \in L^p_{\mathbb{P}}(\Xi)$.

P2. *Positive homogeneity*: $\mathcal{R}[0] = 0$ and $\mathcal{R}[\lambda X] = \lambda\mathcal{R}[X]$ for all $\lambda \in (0,\infty)$ and all $X \in L^p_{\mathbb{P}}(\Xi)$.

P3. *Subadditivity*: $\mathcal{R}[X+Y] \leq \mathcal{R}[X] + \mathcal{R}[Y]$ for all $X, Y \in L^p_{\mathbb{P}}(\Xi)$.

P4. *Monotonicity*: $\mathcal{R}[X] \leq \mathcal{R}[Y]$ for all $X, Y \in L^p_{\mathbb{P}}(\Xi)$ such that $X \leq Y$ almost surely (a. s.)

P5. *Translation equivariance*: $\mathcal{R}[X + c] = \mathcal{R}[X] + c$ for all $X \in L^p_{\mathbb{P}}(\Xi)$ and all $c \in \mathbb{R}$.

P6. *Closedness*: The set $\{X \in L^p_{\mathbb{P}}(\Xi) : \mathcal{R}[X] \leq c\} \subset L^p_{\mathbb{P}}(\Xi)$ is closed for all $c \in \mathbb{R}$.

P7. *Aversity to risk*: $\mathcal{R}[X] > \mathbb{E}[X]$ for all non-constant $X \in L^p_{\mathbb{P}}(\Xi)$.

Some of these properties are used in the definition of other classes of risk measures:

**Definition 9.5.** A proper, convex, positively homogeneous, monotonic, and translation-equivariant risk measure $\mathcal{R} : L^p_{\mathbb{P}}(\Xi) \to \mathbb{R} \cup \{\infty\}$ is called *coherent* [103, Sec. 6.3].

**Remark 9.6.** By [17, Prop. 10.3], a coherent risk measure is automatically *subadditive* because it is convex and positively homogeneous. Moreover, in the original definition [6, Def. 2.4], it is not explicitly mentioned that the risk measure has to be proper because only *finite* risk measures are considered. As noted in [96], a finite, convex and monotonic risk measure is automatically continuous and subdifferentiable on $L^p_{\mathbb{P}}(\Xi)$ by [99, Prop. 3.1].

**Definition 9.7.** A risk measure $\mathcal{R} : L^p_{\mathbb{P}}(\Xi) \to \mathbb{R} \cup \{\infty\}$ is called *regular* [96] if it is closed convex, risk averse, and fulfills $\mathcal{R}[C] = C$ for all $C \in \mathbb{R}$.

**Example 9.8.**

- $\mathcal{R} \equiv \mathbb{E} : L^p_{\mathbb{P}}(\Xi) \to \mathbb{R}$ $(p \geq 1)$ fulfills properties P1, P2, P3, P4, P5, and P6, but not P7.
- $\mathcal{R} \equiv \mathbb{E} + \lambda \operatorname{Var} : L^p_{\mathbb{P}}(\Xi) \to \mathbb{R}$ $(\lambda > 0,\, p \geq 2)$ is regular and fulfills property P5, but not P2, P3, and P4.
- $\mathcal{R} \equiv \mathbb{E} + \lambda \sigma : L^p_{\mathbb{P}}(\Xi) \to \mathbb{R}$ $(\lambda > 0,\, p \geq 2)$ is regular and fulfills properties P2, P3, and P5, but not P4.

For the last two risk measures, convexity and closedness can be observed by $\operatorname{Var}[X] = \|X - \mathbb{E}[X]\|^2_{L^2_{\mathbb{P}}(\Xi)}$ and $\sigma[X] = \|X - \mathbb{E}[X]\|_{L^2_{\mathbb{P}}(\Xi)}$, i. e., they both are a composition of a closed, convex and a linear, bounded function. In addition, $\mathbb{E} + \lambda \sigma$ is subadditive, because it is convex and positively homogeneous, see Remark 9.6. But $\mathbb{E} + \lambda \operatorname{Var}$ is not subadditive: Consider a random variable $X$ with positive variance. Then, $\mathbb{E}[X + X] + \lambda \operatorname{Var}[X + X] = 2\mathbb{E}[X] + 4\lambda \operatorname{Var}[X] > 2\,(\mathbb{E}[X] + \lambda \operatorname{Var}[X])$.

The lacking monotonicity for both $\mathbb{E} + \lambda \operatorname{Var}$ and $\mathbb{E} + \lambda \sigma$ is one of the less obvious properties: Consider, e. g., a random variable $X$ such that $X = d < 0$ with probability $p \in (0, 1)$ and $X = 0$ with probability $1 - p \in (0, 1)$. Furthermore, consider $Y \equiv 0$. Then, $X \leq Y$ a. s.,

$$\mathbb{E}[X] = pd, \quad \operatorname{Var}[X] = pd^2 - (pd)^2 = p(1-p)d^2,$$

$\mathbb{E}[Y] = 0$, and $\operatorname{Var}[Y] = 0$. Thus,

$$0 = \mathbb{E}[Y] + \lambda \operatorname{Var}[Y] < \mathbb{E}[X] + \lambda \operatorname{Var}[X] = pd + \lambda\, p(1-p)d^2 = pd(1 + \lambda(1-p)d)$$

if and only if $d \in (-\frac{1}{(1-p)\lambda}, 0)$. Analogously, $0 < \mathbb{E}[X] + \lambda \sigma[X] = pd - \lambda \sqrt{p(1-p)}d$ if and only if $p \in (0, \frac{\lambda^2}{1+\lambda^2})$.

Furthermore, $\operatorname{VaR}_\beta$ is *not* subadditive in general [6], but $\operatorname{CVaR}_\beta$ fulfills all properties given in Definition 9.4 as shown later.

## 9.2. Mean-Variance Risk Measure

We consider the reduced form (1.3) of an optimal control problem with the risk measure $\mathbb{E} + \lambda \operatorname{Var}$ for some $\lambda > 0$ as done in, e. g., [3]. This gives the problem

$$\min_{u \in U} \hat{\boldsymbol{J}}(u) := \mathbb{E}[\hat{J}_1(u, \cdot)] + \lambda \operatorname{Var}[\hat{J}_1(u, \cdot)] + J_2(u) \quad \text{s. t.} \quad u \in U_{\text{ad}}. \tag{9.1}$$

Compared to the risk-neutral case, this approach shall additionally reduce the variability (measured by the variance) of the tracking term $\hat{J}_1(u, \cdot)$. But it should be mentioned that the objective function $\hat{\boldsymbol{J}}$ may be non-convex even though the functions $\hat{J}_1(\cdot, \xi)$ and $\hat{J}_2(\cdot)$ are convex for almost every $\xi \in \Xi$ and the variance itself is a convex functional. Hence, it is possible that only a local minimizer is computed by a local optimization algorithm, such as

the proposed trust-region method (Algorithm 1). The basic assumptions from Chapter 1 are not enough to discuss this problem. We require:

**Assumption 9.9.**

- $(\Xi, \mathcal{F}, \mathbb{P})$ is a complete probability space.
- $U$ is a Hilbert space.
- $U_{\mathrm{ad}} \subset U$ is nonempty, closed and convex.
- $J_2 : U \to \mathbb{R}$ is twice continuously differentiable.
- $\hat{J}_1 : U \times \Xi \to \mathbb{R}$ is such that $\hat{J}_1(u, \cdot) =: \hat{J}_1[\cdot](u) \in L^{p_{\mathrm{fun}}}_{\mathbb{P}}(\Xi)$ for all $u \in U$ with some $p_{\mathrm{fun}} \in [1, \infty]$.
- The function $\hat{J}_1[\cdot] : U \to L^{p_{\mathrm{fun}}}_{\mathbb{P}}(\Xi)$ is continuously differentiable. Its first derivative $\hat{J}_1[\cdot]'(u) \in \mathcal{L}(U, L^{p_{\mathrm{fun}}}(\Xi))$ can be identified with the partial gradient $\nabla_u \hat{J}_1(u, \cdot) \in L^{p_{\mathrm{fun}}}_{\mathbb{P}}(\Xi; U)$ via
$$\hat{J}_1[\cdot]'(u)w = (\nabla_u \hat{J}_1(u, \cdot), w)_U.$$
Here, the $U$-inner product is taken separately for almost every $\xi \in \Xi$.
- The function $\hat{J}_1[\cdot] : U \to L^{p_{\mathrm{fun}}}_{\mathbb{P}}(\Xi)$ is twice continuously differentiable. Its second derivative $\hat{J}_1[\cdot]''(u) \in \mathcal{L}(U, \mathcal{L}(U, L^{p_{\mathrm{fun}}}(\Xi)))$ can be identified with the partial Hessian $\nabla^2_{uu} \hat{J}_1(u, \cdot) \in L^{p_{\mathrm{fun}}}_{\mathbb{P}}(\Xi; \mathcal{L}(U, U))$ via
$$[\hat{J}_1[\cdot]''(u)s]w = (\nabla^2_{uu} \hat{J}_1(u, \cdot)s, w)_U.$$

In particular, we need $p_{\mathrm{fun}} \in [2, \infty]$ for the mean-variance risk measure to be finite. Assumption 9.9 allows to apply the chain rule to compute the derivative of the objective function $\hat{\boldsymbol{J}}$. We note that due to

$$\mathbb{E} : L^{p_{\mathrm{fun}}}_{\mathbb{P}}(\Xi) \to \mathbb{R}, \quad \mathbb{E}[X] = \int_\Xi \mathbb{1} \cdot X \, \mathrm{d}\mathbb{P}$$

we can identify the derivative $\mathbb{E}'[X] = \mathbb{1} \in L^{p^*_{\mathrm{fun}}}_{\mathbb{P}}(\Xi)$ for all $X \in L^{p_{\mathrm{fun}}}_{\mathbb{P}}(\Xi)$, where $p_{\mathrm{fun}} \in [1, \infty]$ and $\frac{1}{p_{\mathrm{fun}}} + \frac{1}{p^*_{\mathrm{fun}}} = 1$. Furthermore, it holds that

$$\mathrm{Var} : L^{p_{\mathrm{fun}}}_{\mathbb{P}}(\Xi) \to \mathbb{R}, \quad \mathrm{Var}[X] = \|X - \mathbb{E}[X] \cdot \mathbb{1}\|^2_{L^2_{\mathbb{P}}(\Xi)} = \|X\|^2_{L^2_{\mathbb{P}}(\Xi)} - \mathbb{E}[X]^2$$

and therefore $\mathrm{Var}'[X] = 2X - 2\,\mathbb{E}[X] \cdot \mathbb{1} \in L^{p^*_{\mathrm{fun}}}_{\mathbb{P}}(\Xi)$ with $p_{\mathrm{fun}} \in [2, \infty]$. Note that we embed $L^{p_{\mathrm{fun}}}_{\mathbb{P}}(\Xi) \hookrightarrow L^2_{\mathbb{P}}(\Xi)$ and use also the adjoint embedding $L^2_{\mathbb{P}}(\Xi) \hookrightarrow L^{p^*_{\mathrm{fun}}}_{\mathbb{P}}(\Xi)$ to compute this Fréchet derivative involving the derivative of the squared $L^2_{\mathbb{P}}(\Xi)$-norm. In general, $L^\infty_{\mathbb{P}}(\Xi)^*$ contains also functionals which cannot be identified with $L^1_{\mathbb{P}}(\Xi)$ functions, but all respective objects are indeed in $L^1_{\mathbb{P}}(\Xi)$ for $p_{\mathrm{fun}} = \infty$. The second derivative of Var is $\mathrm{Var}''[X]Y = 2Y - 2\,\mathbb{E}[Y] \cdot \mathbb{1} \in L^{p^*_{\mathrm{fun}}}_{\mathbb{P}}(\Xi)$ for $Y \in L^{p_{\mathrm{fun}}}_{\mathbb{P}}(\Xi)$.

By the chain rule, we obtain the gradient of $\hat{\boldsymbol{J}}$ as

$$\begin{aligned}\nabla\hat{\boldsymbol{J}}(u) &= \mathbb{E}[\nabla_u\hat{J}_1(u,\cdot)] + 2\lambda\int_\Xi \big(\hat{J}_1(u,\cdot) - \mathbb{E}[\hat{J}_1(u,\cdot)]\cdot\mathbb{1}\big)\nabla_u\hat{J}_1(u,\cdot)\,\mathrm{d}\mathbb{P} + \nabla J_2(u)\\ &= (1-2\lambda\mathbb{E}[\hat{J}_1(u,\cdot)])\mathbb{E}[\nabla_u\hat{J}_1(u,\cdot)] + 2\lambda\int_\Xi \hat{J}_1(u,\cdot)\,\nabla_u\hat{J}_1(u,\cdot)\,\mathrm{d}\mathbb{P} + \nabla J_2(u)\\ &= \mathbb{E}[\nabla_u\hat{J}_1(u,\cdot)] + 2\lambda\,\mathrm{Cov}[\hat{J}_1(u,\cdot),\nabla_u\hat{J}_1(u,\cdot)] + \nabla J_2(u)\end{aligned} \tag{9.2}$$

and the Hessian

$$\begin{aligned}\nabla^2\hat{\boldsymbol{J}}(u)s &= \mathbb{E}[\nabla^2_{uu}\hat{J}_1(u,\cdot)s] + \lambda\left\langle \mathrm{Var}''[\hat{J}_1(u,\cdot)](\nabla_u J_1(u,\cdot),s)_U, \nabla_u J_1(u,\cdot)\right\rangle_{L_{\mathbb{P}}^{p^*_{\mathrm{fun}}}(\Xi),L_{\mathbb{P}}^{p_{\mathrm{fun}}}(\Xi)}\\ &\qquad + \lambda\left\langle \mathrm{Var}'[\hat{J}_1(u,\cdot)], \nabla^2_{uu}J_1(u,\cdot)s\right\rangle_{L_{\mathbb{P}}^{p^*_{\mathrm{fun}}}(\Xi),L_{\mathbb{P}}^{p_{\mathrm{fun}}}(\Xi)} + \nabla^2 J_2(u)s\\ &= \mathbb{E}[\nabla^2_{uu}\hat{J}_1(u,\cdot)s] + 2\lambda\int_\Xi \big((\nabla_u J_1(u,\cdot),s)_U - \mathbb{E}[(\nabla_u J_1(u,\cdot),s)_U]\cdot\mathbb{1}\big)\,\nabla_u\hat{J}_1(u,\cdot)\,\mathrm{d}\mathbb{P}\\ &\qquad + 2\lambda\int_\Xi \big(\hat{J}_1(u,\cdot) - \mathbb{E}[\hat{J}_1(u,\cdot)]\cdot\mathbb{1}\big)\nabla^2_{uu}\hat{J}_1(u,\cdot)s\,\mathrm{d}\mathbb{P} + \nabla^2 J_2(u)s\\ &= \mathbb{E}[\nabla^2_{uu}\hat{J}_1(u,\cdot)s] + 2\lambda\,\mathrm{Cov}[(\nabla_u J_1(u,\cdot),s)_U,\nabla_u\hat{J}_1(u,\cdot)]\\ &\qquad + 2\lambda\,\mathrm{Cov}[\big(\hat{J}_1(u,\cdot),\nabla^2_{uu}\hat{J}_1(u,\cdot)s] + \nabla^2 J_2(u)s.\end{aligned}$$

The advantage of these derivative formulations is that they use the partial derivatives w. r. t. $u$ of the parameter-dependent reduced objective function. Often, the computation of these derivatives is already implemented by the adjoint approach and can be reused for an implementation of the mean-variance risk measure. Overall, we do not have to define an alternative adjoint state or adjoint equation for the objective function involving the variance term.

### 9.2.1. Discussion of the Example

We discuss under which conditions Assumption 9.9 is satisfied for the example from Chapter 3, i. e., we have

$$\hat{J}[\xi](u) = J_1[\xi](S[\xi](u),u,\xi) + J_2(u) = \tfrac{1}{2}\|Q(\xi)S[\xi](u) - \hat{q}(\xi)\|^2_H + \tfrac{\gamma}{2}\|u\|^2_U,$$

see (3.5). According to (3.17), the derivatives are given by

$$\hat{J}_1[\xi]'(u) = -B(\xi)^*T[\xi](u) \quad\text{and}\quad \nabla J_2(u) = \gamma u,$$

where $\hat{J}_1[\xi](u) = J_1[\xi](S[\xi](u),u,\xi)$. Due to Corollary 3.15 it holds that $S[\cdot](u)\in L^{r_f}(\Xi;Y)$ and Lemma 3.23 yields $T[\cdot](u)\in L^{r_z}_{\mathbb{P}}(\Xi;Y)$ with $r_z$ depending on the integrability exponents $r_f$, $r_Q$, and $r_{\hat{q}}$ of the state, the state-to-observation map $Q(\cdot)$, and the desired state $\hat{q}(\cdot)$, respectively. Especially, the functions $\xi\mapsto\hat{J}_1[\xi](u)$ and $\xi\mapsto\hat{J}_1[\xi]'(u)$ are measurable and one obtains the integrability exponents

$$\hat{J}_1[\cdot](u)\in L^{\hat{r}/2}_{\mathbb{P}}(\Xi) \quad\text{and}\quad \hat{J}_1[\cdot]'(u)\in L^{r_z}_{\mathbb{P}}(\Xi;U^*)$$

with $\hat{r}$ from Table 3.1 and because $B(\cdot) \in L^\infty_{\mathbb{P}}(\Xi; \mathcal{L}(U, Y^*))$. If $\hat{r} \geq 2p_{\text{fun}}$ and $r_z \geq p_{\text{fun}}$, the objective function and the gradient are both $p_{\text{fun}}$-integrable with $p_{\text{fun}} \geq 2$. In the "Example" column of Table 3.1 this would correspond to having $r_f \geq \max\{p, 4\}$, $r_Q = \infty$, and $r_{\hat{q}} \geq r_f$.

The differentiability assumption on $\hat{J}_1[\cdot]$ can be proven using a priori estimates on the adjoint state and the same techniques as in the proof of Theorem 3.26. For this purpose, let the regularity exponents $r_Q, r_f, r_{\hat{q}} \in [1, \infty]$ be such that $\frac{2}{r_Q} + \frac{1}{r_f} \leq \frac{1}{p_{\text{fun}}}$ and $\frac{1}{r_Q} + \frac{1}{r_{\hat{q}}} \leq \frac{1}{p_{\text{fun}}}$ hold. One computes for $w \in U \setminus \{0\}$:

$$
\begin{aligned}
0 &\leq \lim_{\|w\|_U \to 0} \frac{\big\| \hat{J}_1(u+w,\cdot) - \hat{J}_1(u,\cdot) - (\nabla_u \hat{J}_1(u,\cdot), w)_U \big\|_{L^{p_{\text{fun}}}_{\mathbb{P}}(\Xi)}}{\|w\|_U} \\
&= \lim_{\|w\|_U \to 0} \frac{\big\| \int_0^1 (\nabla_u \hat{J}_1(u+\tau w,\cdot) - \nabla_u \hat{J}_1(u,\cdot), w)_U \,\mathrm{d}\tau \big\|_{L^{p_{\text{fun}}}_{\mathbb{P}}(\Xi)}}{\|w\|_U} \\
&\leq \lim_{\|w\|_U \to 0} \Big\| \int_0^1 \|\nabla_u \hat{J}_1(u+\tau w,\cdot) - \nabla_u \hat{J}_1(u,\cdot)\|_U \,\mathrm{d}\tau \Big\|_{L^{p_{\text{fun}}}_{\mathbb{P}}(\Xi)} \\
&\leq \lim_{\|w\|_U \to 0} \Big\| \sup_{\tau \in [0,1]} \|\nabla_u \hat{J}_1(u+\tau w,\cdot) - \nabla_u \hat{J}_1(u,\cdot)\|_U \Big\|_{L^{p_{\text{fun}}}_{\mathbb{P}}(\Xi)} = 0.
\end{aligned}
$$

In the following we argue why this limit is indeed zero.

For $p_{\text{fun}} \in [1, \infty)$, this follows from the dominated convergence theorem because

$$
\sup_{\tau \in [0,1]} \|\nabla_u \hat{J}_1(u+\tau w, \xi) - \nabla_u \hat{J}_1(u, \xi)\|_U \longrightarrow 0 \quad \text{as} \quad \|w\|_U \to 0
$$

for almost every $\xi \in \Xi$ due to the continuity of $u \mapsto \nabla_u \hat{J}_1(u, \xi)$. Additionally, we have

$$
\begin{aligned}
&\sup_{\tau \in [0,1]} \|\nabla_u \hat{J}_1(u+\tau w, \xi) - \nabla_u \hat{J}_1(u, \xi)\|_U = \sup_{\tau \in [0,1]} \| - B(\xi)^*(T[\xi](u+\tau w) - T[\xi](u))\|_{U^*} \\
&\leq \|B(\xi)\|_{\mathcal{L}(U,Y^*)} \big( \sup_{\tau \in [0,1]} \|T[\xi](u+\tau w)\|_Y + \|T[\xi](u))\|_Y \big) \\
&\leq \tfrac{C_\Omega}{\underline{\kappa}^2} \|B(\xi)\|_{\mathcal{L}(U,Y^*)} \|Q(\xi)\|^2_{\mathcal{L}(Y,H)} \big( \|D\|_{\mathcal{L}(U,L^2(\Omega))} (2\|u\|_U + 1) + \|f(\xi) - \varphi(0)\|_{L^2(\Omega)} \big) \\
&\qquad + \tfrac{1}{\underline{\kappa}} \|B(\xi)\|_{\mathcal{L}(U,Y^*)} \|Q(\xi)\|_{\mathcal{L}(Y,H)} \|\hat{q}(\xi)\|_H
\end{aligned}
$$

for $\|w\|_U \leq 1$. The upper bound is due to the a priori estimate

$$
\begin{aligned}
\|T[\xi](u)\|_Y &\leq \tfrac{1}{\underline{\kappa}} \|Q(\xi)\|_{\mathcal{L}(Y,H)} \|(Q(\xi) S[\xi](u) - \hat{q}(\xi))\|_H \\
&\leq \tfrac{1}{\underline{\kappa}} \|Q(\xi)\|_{\mathcal{L}(Y,H)} \big( \|Q(\xi)\|_{\mathcal{L}(Y,H)} \|S[\xi](u)\|_Y + \|\hat{q}(\xi))\|_H \big) \\
&\leq \tfrac{C_\Omega}{\underline{\kappa}^2} \|Q(\xi)\|^2_{\mathcal{L}(Y,H)} \big( \|D\|_{\mathcal{L}(U,L^2(\Omega))} \|u\|_U + \|f(\xi) - \varphi(0)\|_{L^2(\Omega)} \big) \\
&\qquad + \tfrac{1}{\underline{\kappa}} \|Q(\xi)\|_{\mathcal{L}(Y,H)} \|\hat{q}(\xi)\|_H
\end{aligned}
$$

on the adjoint state $T[\xi](u)$, which is obtained by combining (3.20) and (3.9). It is an $L^{p_{\text{fun}}}_{\mathbb{P}}(\Xi)$-function w. r. t. $\xi$ because $\frac{2}{r_Q} + \frac{1}{r_f} \leq \frac{1}{p_{\text{fun}}}$, $\frac{1}{r_Q} + \frac{1}{r_{\hat{q}}} \leq \frac{1}{p_{\text{fun}}}$, and $B(\cdot)$ is essentially bounded w. r. t. $\xi$.

For $p_{\text{fun}} = \infty$, the uniform Lipschitz continuity of the gradient is used to show the result. In analogy to (3.9), one can estimate

$$\|S[\xi](u+w) - S[\xi](u)\|_Y \leq \tfrac{C_\Omega}{\underline{\kappa}} \|Dw\|_{L^2(\Omega)}$$

independently of $\xi$. Using the derivations in the proofs of Theorem 5.1 and Lemma 5.2 with $y = S[\xi](u)$, $\tilde{y} = S[\xi](u+w)$, $z = T[\xi](u)$, $\hat{z} = T[\xi](u+w)$ as well as (3.9), this yields

$$\begin{aligned}
&\|T[\xi](u+w) - T[\xi](u)\|_Y \\
&\leq \tfrac{1}{\underline{\kappa}} \|Q(\xi)\|^2_{\mathcal{L}(Y,H)} \|S[\xi](u) - S[\xi](u+w)\|_Y \\
&\quad + \tfrac{1}{\underline{\kappa}^2} \|N'(S[\xi](u+w)) - N'(S[\xi](u))\|_{\mathcal{L}(Y,Y^*)} \|Q(\xi)^*(Q(\xi)S[\xi](u) - \hat{q}(\xi))\|_{Y^*} \\
&\leq \tfrac{C_\Omega}{\underline{\kappa}^2} \|Q(\xi)\|^2_{\mathcal{L}(Y,H)} \|Dw\|_{L^2(\Omega)} \\
&\quad + \tfrac{1}{\underline{\kappa}^2} c_p^3 \Big( a''_{\varphi''} \cdot \lambda(\Omega)^{(p-3)/p} + c''_{\varphi''} c_p^{p-3} \cdot \big( \tfrac{C_\Omega}{\underline{\kappa}} (\|Du + f(\xi) - \varphi(0)\|_{L^2(\Omega)} + \|Dw\|_{L^2(\Omega)}) \big)^{p-3} \Big) \\
&\qquad \cdot \tfrac{C_\Omega}{\underline{\kappa}} \|Dw\|_{L^2(\Omega)} \|Q(\xi)\|_{\mathcal{L}(Y,H)} \big( \|Q(\xi)\|_{\mathcal{L}(Y,H)} \|Du + f(\xi) - \varphi(0)\|_{L^2(\Omega)} + \|\hat{q}(\xi)\|_H \big),
\end{aligned}$$

cf. (3.20). Since $r_Q = r_{\hat{q}} = r_f = \infty$, this bound converges to zero in $L^\infty_{\mathbb{P}}(\Xi)$ as $\|w\|_U \to 0$. Hence,

$$\begin{aligned}
0 &\leq \Big\| \sup_{\tau \in [0,1]} \|\nabla_u \hat{J}_1(u + \tau w, \cdot) - \nabla_u \hat{J}_1(u, \cdot)\|_U \Big\|_{L^\infty_{\mathbb{P}}(\Xi)} \\
&\leq \Big\| \|B(\cdot)\|_{\mathcal{L}(U,Y^*)} \Big\|_{L^\infty_{\mathbb{P}}(\Xi)} \Big\| \sup_{\tau \in [0,1]} \|T[\xi](u + \tau w) - T[\xi](u)\|_Y \Big\|_{L^\infty_{\mathbb{P}}(\Xi)} \longrightarrow 0 \quad \text{as} \quad \|w\|_U \to 0
\end{aligned}$$

can be concluded since $B(\cdot)$ is essentially bounded w. r. t. $\xi$.

The considerations show the Fréchet approximation condition for the first derivative. For the second derivative, it can be proven analogously using the a priori bound (3.26). Continuity of the second derivative can also be shown with similar arguments, i. e., using pointwise convergence and the dominated convergence theorem for $p_{\text{fun}} \in [1, \infty)$ or refined bounds in the case $p_{\text{fun}} = \infty$.

### 9.2.2. Possible Error Estimates for Reduced Objective Function and Gradient

In the following, we investigate how the required error estimates from Section 5.2 change in the mean-variance case.

#### Objective Function Evaluation Error

We assume that the function $\hat{J}_1(u, \cdot)$ in (9.1) is evaluated inexactly due to an inexact state computation, giving the inexact version $\tilde{J}_1(u, \cdot)$, whereas $J_2$ is evaluated exactly. Altogether, we obtain the inexact version $\tilde{\boldsymbol{J}}$ of the objective function $\hat{\boldsymbol{J}}$. In the proof of Proposition 5.11, the error in $\tilde{\boldsymbol{J}}$ is derived from the error in $\tilde{J}_1$ by means of

$$|\hat{\boldsymbol{J}}(u) - \tilde{\boldsymbol{J}}(u)| = |\mathbb{E}[\hat{J}_1(u, \cdot) - \tilde{J}_1(u, \cdot)]| \leq \|\hat{J}_1(u, \cdot) - \tilde{J}_1(u, \cdot)\|_{L^1_{\mathbb{P}}(\Xi)},$$

which leads to an error bound based on the $L^2_{\mathbb{P}}(\Xi)$-error $\|\boldsymbol{Q}(\tilde{\boldsymbol{y}} - \boldsymbol{y})\|_{L^2_{\mathbb{P}}(\Xi;H)}$ in the inexact state. In the mean-variance case, we have

$$\begin{aligned}
|\hat{\boldsymbol{J}}(u) - \tilde{\boldsymbol{J}}(u)| &= \Big|\mathbb{E}[\hat{J}_1(u,\cdot) - \tilde{J}_1(u,\cdot)] + \lambda\big(\mathrm{Var}[\hat{J}_1(u,\cdot)] - \mathrm{Var}[\tilde{J}_1(u,\cdot)]\big)\Big| \\
&= \big|\mathbb{E}[\hat{J}_1(u,\cdot) - \tilde{J}_1(u,\cdot)] + \lambda\,\mathbb{E}[(\hat{J}_1(u,\cdot) - \tilde{J}_1(u,\cdot))(\hat{J}_1(u,\cdot) + \tilde{J}_1(u,\cdot))] \\
&\qquad + \lambda\,\mathbb{E}[\tilde{J}_1(u,\cdot) - \hat{J}_1(u,\cdot)]\mathbb{E}[\hat{J}_1(u,\cdot) + \tilde{J}_1(u,\cdot)]\big| \\
&= \big|\mathbb{E}[\mathbb{1} - \lambda\hat{J}_1(u,\cdot) - \lambda\tilde{J}_1(u,\cdot)]\mathbb{E}[\hat{J}_1(u,\cdot) - \tilde{J}_1(u,\cdot)] \\
&\qquad + \lambda\,\mathbb{E}[(\hat{J}_1(u,\cdot) - \tilde{J}_1(u,\cdot))(\hat{J}_1(u,\cdot) + \tilde{J}_1(u,\cdot))]\big| \\
&\leq \|\mathbb{1} - \lambda\hat{J}_1(u,\cdot) - \lambda\tilde{J}_1(u,\cdot)\|_{L^1_{\mathbb{P}}(\Xi)}\|\hat{J}_1(u,\cdot) - \tilde{J}_1(u,\cdot)\|_{L^1_{\mathbb{P}}(\Xi)} \\
&\qquad + \lambda\|\hat{J}_1(u,\cdot) - \tilde{J}_1(u,\cdot)\|_{L^1_{\mathbb{P}}(\Xi)}\|\hat{J}_1(u,\cdot) + \tilde{J}_1(u,\cdot)\|_{L^\infty_{\mathbb{P}}(\Xi)}.
\end{aligned}$$

The estimate requires $p_{\mathrm{fun}} = \infty$ for both the function $\hat{J}_1$ and its inexact version $\tilde{J}_1$. In Table 3.1, we would therefore need $r_f = \infty$, $r_Q = \infty$ and $r_{\hat{q}} = \infty$ to be able to apply it. But this is the only way to obtain an estimate which depends on the $L^1_{\mathbb{P}}(\Xi)$-error in $\hat{J}_1$ so that an $L^2_{\mathbb{P}}(\Xi)$-error estimate for the inexact state is sufficient. For the more natural choice that $\|\hat{J}_1(u,\cdot) - \tilde{J}_1(u,\cdot)\|_{L^2_{\mathbb{P}}(\Xi)}$ has to be controlled, the $L^4_{\mathbb{P}}(\Xi)$-error in the state would have to be estimated, see the proof of Proposition 5.11.

### Model gradient error

The gradient (9.2) is computed inexactly due to an inexact objective function evaluation $\tilde{J}(u,\cdot)$ and an inexact gradient, denoted by $\tilde{\nabla}_u\hat{J}(u,\cdot) \approx \nabla_u\hat{J}(u,\cdot)$. This yields the inexact gradient $\tilde{\nabla}\hat{\boldsymbol{J}}(u)$, the error in which shall be estimated in the following.

$$\begin{aligned}
&\big\|\nabla\hat{\boldsymbol{J}}(u) - \tilde{\nabla}\hat{\boldsymbol{J}}(u)\big\|_U \\
&= \Big\|\mathbb{E}[\nabla_u\hat{J}_1(u,\cdot)] + 2\lambda\int_\Xi \big(\hat{J}_1(u,\cdot) - \mathbb{E}[\hat{J}_1(u,\cdot)]\cdot\mathbb{1}\big)\nabla_u\hat{J}_1(u,\cdot)\,\mathrm{d}\mathbb{P} \\
&\quad - \mathbb{E}[\nabla_u\tilde{J}_1(u,\cdot)] - 2\lambda\int_\Xi \big(\tilde{J}_1(u,\cdot) - \mathbb{E}[\tilde{J}_1(u,\cdot)]\cdot\mathbb{1}\big)\tilde{\nabla}_u\hat{J}_1(u,\cdot)\,\mathrm{d}\mathbb{P}\Big\|_U \\
&\leq \mathbb{E}\big[\|\nabla_u\hat{J}_1(u,\cdot) - \nabla_u\tilde{J}_1(u,\cdot)\|_U\big] \\
&\quad + 2\lambda\Big\|\int_\Xi \big(\hat{J}_1(u,\cdot) - \mathbb{E}[\hat{J}_1(u,\cdot)]\cdot\mathbb{1}\big)\nabla_u\hat{J}_1(u,\cdot) - \big(\tilde{J}_1(u,\cdot) - \mathbb{E}[\tilde{J}_1(u,\cdot)]\cdot\mathbb{1}\big)\tilde{\nabla}_u\hat{J}_1(u,\cdot)\,\mathrm{d}\mathbb{P}\Big\|_U \\
&\leq \|\nabla_u\hat{J}_1(u,\cdot) - \nabla_u\tilde{J}_1(u,\cdot)\|_{L^1_{\mathbb{P}}(\Xi;U)} \\
&\quad + 2\lambda\big\|\hat{J}_1(u,\cdot) - \tilde{J}_1(u,\cdot) + \mathbb{E}[\tilde{J}_1(u,\cdot) - \hat{J}_1(u,\cdot)]\cdot\mathbb{1}\big\|_{L^{\tilde{r}}_{\mathbb{P}}(\Xi)}\big\|\nabla_u\hat{J}_1(u,\cdot) - \tilde{\nabla}_u\hat{J}_1(u,\cdot)\big\|_{L^{\tilde{r}^*}_{\mathbb{P}}(\Xi;U)} \\
&\quad + 2\lambda\big\|\tilde{J}_1(u,\cdot) - \mathbb{E}[\tilde{J}_1(u,\cdot)]\cdot\mathbb{1}\big\|_{L^{\tilde{r}}_{\mathbb{P}}(\Xi)}\big\|\nabla_u\hat{J}_1(u,\cdot) - \tilde{\nabla}_u\hat{J}_1(u,\cdot)\big\|_{L^{\tilde{r}^*}_{\mathbb{P}}(\Xi;U)} \\
&\quad + 2\lambda\big\|\hat{J}_1(u,\cdot) - \tilde{J}_1(u,\cdot) + \mathbb{E}[\tilde{J}_1(u,\cdot) - \hat{J}_1(u,\cdot)]\cdot\mathbb{1}\big\|_{L^{\tilde{r}}_{\mathbb{P}}(\Xi)}\big\|\tilde{\nabla}_u\hat{J}_1(u,\cdot)\big\|_{L^{\tilde{r}^*}_{\mathbb{P}}(\Xi;U)}
\end{aligned}$$

with $\tilde{r}, \tilde{r}^* \in [1,\infty]$, $\frac{1}{\tilde{r}} + \frac{1}{\tilde{r}^*} = 1$ and provided the respective quantities enjoy the required integrability properties. Controlling the objective function evaluation in the $L^{\tilde{r}}_{\mathbb{P}}(\Xi)$-norm requires to estimate the $L^{2\tilde{r}}_{\mathbb{P}}(\Xi)$-error in the state. Even if the state equation is linear, i. e., $\varphi'$

is constant, the $L^{\tilde{r}^*}_{\mathbb{P}}(\Xi)$-error in the state and adjoint state is needed to estimate the gradient error in the $L^{\tilde{r}^*}_{\mathbb{P}}(\Xi;U)$-norm, cf. Theorem 5.8. The "best" choice of $\tilde{r}$, which demands the weakest integrability properties of the state, is therefore $\tilde{r} = \frac{3}{2}$, which yields that the $L^3_{\mathbb{P}}(\Xi)$-error in the state has to be controlled because $2\tilde{r} = 3 = \tilde{r}^*$.

Overall, we see that the risk-neutral implementation can be extended to the mean-variance case by only changing the evaluation of the reduced objective function, its gradient and its Hessian based on the already derived (linearized) state and adjoint state computation. In addition, the error estimation would have to be changed accordingly. Here, the issue arises that it becomes necessary to control the error in the state at least in the $L^3_{\mathbb{P}}(\Xi)$-norm even for $r_f = \infty$ and a linear state equation. This is not a simple task and would require a new error estimation procedure. We leave this for future research. Additionally, we want to mention again that the mean-variance risk measure is *not* monotonic. This means that we could have a control for which the tracking term is larger than the one for a different control almost surely although the corresponding risk measured by $\mathbb{E} + \lambda \operatorname{Var}$ is smaller. We therefore refrain from presenting a numerical implementation and results here and consider a class of risk measures with better theoretical properties.

## 9.3. Convex Combination of Mean and Conditional Value-at-Risk

In this section, we introduce a class of risk measures, which are convex combinations of the expected value and the conditional value-at-risk $\mathrm{CVaR}_\beta$, and derive the respective properties for them. These risk measures are nonsmooth, but minimizing them can be reformulated introducing a pointwisely constrained auxiliary variable. We propose solving this problem by an interior-point method. The introduced barrier term leads to a smoothed version of the original risk measure. It is investigated which properties of it are retained by the smoothing procedure. Optimal control problems with this smoothed risk measure are investigated in theory and numerical results of an implementation with low-rank tensors are presented.

### 9.3.1. Definition and Derivation of the Properties

In [95] it is shown that the conditional value-at-risk of an $L^1_{\mathbb{P}}(\Xi)$ random variable $X$ can be computed as follows:

$$\mathrm{CVaR}_\beta[X] = \inf_{t\in\mathbb{R}} t + \tfrac{1}{1-\beta}\mathbb{E}[(X-t)^+]$$

with $(s)^+ := \max\{0, s\}$. This gives rise to considering risk measures of the form

$$\mathcal{R}_v : L^p_{\mathbb{P}}(\Xi) \to \mathbb{R}\cup\{\infty\}, \quad \mathcal{R}_v[X] = \inf_{t\in\mathbb{R}} t + \mathbb{E}[v(X-t)], \tag{9.3}$$

where $v : \mathbb{R} \to \mathbb{R}$ is a given function. Risk measures of this kind correspond to the expectation quadrangle presented in [96]. In the following, we say that the risk measure $\mathcal{R}_v$ *is induced by* the function $v$. Two examples are the expectation $\mathbb{E}$ for $v(s) = s$ and the conditional value-at-risk $\mathrm{CVaR}_\beta$ for $v(s) = \frac{1}{1-\beta}(s)^+$.

The risk measure $\mathcal{R}_v$ inherits several properties from the function $v$ as shown in the following.

**Lemma 9.10.** *The risk measure $\mathcal{R}_v$ is proper, i. e., $\mathcal{R}_v[X] > -\infty$ for all $X$ and $\mathcal{R}_v[\tilde{X}] < \infty$ for some $\tilde{X}$ if and only if there exists $c \in \mathbb{R}$ such that $v(s) \geq s + c$ holds for all $s \in \mathbb{R}$.*

*Proof.* For the "if" part observe that from $v(s) \geq s+c$ it follows that $t+v(X-t) \geq t+(X-t+c) = X+c$. Therefore, $\mathcal{R}_v[X] = \inf_{t\in\mathbb{R}} \mathbb{E}[t+v(X-t)] \geq \inf_{t\in\mathbb{R}} \mathbb{E}[X+c] = \mathbb{E}[X]+c > -\infty$ holds because the expectation is finite on $L^p_{\mathbb{P}}(\Xi)$ for all $p \in [1,\infty]$. Furthermore, for any constant random variable $\tilde{X} \equiv d \in \mathbb{R}$ and some given $\hat{t} \in \mathbb{R}$ we have that $\mathcal{R}_v[\tilde{X}] = \inf_{t\in\mathbb{R}} t + \mathbb{E}[v(\tilde{X}-t)] = \inf_{t\in\mathbb{R}} t + v(d-t) \leq \hat{t} + v(d-\hat{t}) < \infty$.

Now we consider the "only if" part: Assume that there does not exist any $c \in \mathbb{R}$ such that $v(s) \geq s + c$ holds for all $s \in \mathbb{R}$, i. e., for every $c \in \mathbb{R}$ there exists $\tilde{s} \in \mathbb{R}$ with $v(\tilde{s}) < \tilde{s} + c$. Now let $(c_k) \subset \mathbb{R}$ be a sequence diverging to $-\infty$, e. g., $c_k = -k$, and let $X \equiv d \in \mathbb{R}$. Then there exists $(t_k) \subset \mathbb{R}$ such that $t_k + v(X - t_k) < d + c_k$ holds for all $k \in \mathbb{N}$. It follows that $\inf_{t\in\mathbb{R}} t + \mathbb{E}[v(X-t)] = -\infty$ for all constant $X$. □

**Corollary 9.11.** *If $v$ is convex and there exists $\tilde{s} \in \mathbb{R}$ such that $1 \in \partial v(\tilde{s})$ (convex subdifferential), the risk measure $\mathcal{R}_v$ is proper.*

*Proof.* Lemma 9.10 can be applied because $v(s) \geq v(\tilde{s}) + 1 \cdot (s - \tilde{s}) = s + c$ holds with $c = v(\tilde{s}) - \tilde{s}$. □

**Lemma 9.12.** *If there exist $c, d \in \mathbb{R}$ such that $|v(s)| \leq c|s|^p + d$ holds for all $s \in \mathbb{R}$, we have $\mathcal{R}_v[X] < \infty$ for all $X \in L^p_{\mathbb{P}}(\Xi)$.*

*Proof.* We have $\mathbb{E}[v(X)] \leq \mathbb{E}[|v(X)|] \leq \mathbb{E}[c\,|X|^p] + d = c\,||X||^p_{L^p_{\mathbb{P}}(\Xi)} + d < \infty$. Therefore, the argument in the infimum defining $\mathcal{R}_v$ is finite for $t = 0$. □

**Lemma 9.13.** *The risk measure $\mathcal{R}_v$ is always translation-equivariant.*

*Proof.* The statement follows from

$$\mathcal{R}_v[X+c] = \inf_{t\in\mathbb{R}} t + \mathbb{E}[v(X+c-t)] \overset{\tilde{t}=t-c}{=} \inf_{\tilde{t}\in\mathbb{R}} \tilde{t} + c + \mathbb{E}[v(X-\tilde{t})] = \mathcal{R}_v[X] + c.$$

□

**Lemma 9.14.** *If the function $v$ is convex, the risk measure $\mathcal{R}_v$ is convex.*

For the proof of this statement, we use the following general result:

**Proposition 9.15.** *Let $U$ and $W$ be convex subsets of $\mathbb{R}$-vector spaces and let $f : U \times W \to (-\infty, \infty]$ be a convex function. Then the function $g : U \to [-\infty, \infty]$, $g(u) := \inf_{w\in W} f(u,w)$ is convex.*

*Proof.* This claim follows as in the proof of [17, Prop. 8.26]. □

*Proof of Lemma 9.14.* The result follows by applying Proposition 9.15 with $W = \mathbb{R}$, $U = L^p_{\mathbb{P}}(\Xi)$ and $f(u,w) := w + \mathbb{E}[v(u-w)]$. □

**Lemma 9.16.** *If $v$ is positively homogeneous and $\inf_{t\in\mathbb{R}} v(t) - t = 0$, the risk measure $\mathcal{R}_v$ is positively homogeneous.*

*Proof.* We use that $v(\lambda s) = \lambda v(s)$ holds for all $x \in \mathbb{R}$ and all $\lambda > 0$. Let $X \in L^p_{\mathbb{P}}(\Xi)$ be a random variable and $\lambda > 0$. Then:

$$\begin{aligned}\mathcal{R}_v[\lambda X] = \inf_{\tau\in\mathbb{R}} \tau + \mathbb{E}[v(\lambda X - \tau)] &\overset{v \text{ pos. hom.}}{=} \inf_{\tau\in\mathbb{R}} \tau + \mathbb{E}[\lambda\, v(X - \tfrac{\tau}{\lambda})] \\ &\overset{\tau=\lambda t}{=} \inf_{t\in\mathbb{R}} \lambda t + \lambda\, \mathbb{E}[v(X-t)] \overset{\lambda>0}{=} \lambda\, \mathcal{R}_v[X].\end{aligned}$$

Furthermore, $\mathcal{R}_v[0] = \inf_{\tau\in\mathbb{R}} \tau + \mathbb{E}[v(0-\tau)] \overset{t=-\tau}{=} \inf_{t\in\mathbb{R}} v(t) - t = 0$. □

**Lemma 9.17.** *If $v$ is subadditive, the risk measure $\mathcal{R}_v$ is subadditive.*

*Proof.* We use that $v(s+\tilde{s}) \leq v(s)+v(\tilde{s})$ holds for all $s, \tilde{s} \in \mathbb{R}$. Let $X, \tilde{X} \in L^p_{\mathbb{P}}(\Xi)$ be random variables. Then:

$$\begin{aligned}\mathcal{R}_v[X+\tilde{X}] &= \inf_{\tau\in\mathbb{R}} \tau + \mathbb{E}[v(X+\tilde{X}-\tau)] \overset{\tau=t+\tilde{t}}{=} \inf_{t,\tilde{t}\in\mathbb{R}} t+\tilde{t} + \mathbb{E}[v(X+\tilde{X}-t-\tilde{t})] \\ &\overset{v \text{ subadditive}}{\leq} \inf_{t,\tilde{t}\in\mathbb{R}} t+\tilde{t} + \mathbb{E}[v(X-t)+v(\tilde{X}-\tilde{t})] \\ &= \inf_{t\in\mathbb{R}} t + \mathbb{E}[v(X-t)] + \inf_{\tilde{t}\in\mathbb{R}} \tilde{t} + \mathbb{E}[v(\tilde{X}-\tilde{t})] = \mathcal{R}_v[X] + \mathcal{R}_v[\tilde{X}].\end{aligned}$$

□

**Lemma 9.18.** *If $v$ is increasing, the risk measure $\mathcal{R}_v$ is monotonic.*

*Proof.* Let $X \leq Y$ a. s. Then, $v(X-t) \leq v(Y-t)$ a. s. for all $t \in \mathbb{R}$ because $v$ is increasing. Using the monotonicity of the expectation it follows that

$$\mathcal{R}_v[X] = \inf_{t\in\mathbb{R}} t + \mathbb{E}[v(X-t)] \leq \inf_{t\in\mathbb{R}} t + \mathbb{E}[v(Y-t)] = \mathcal{R}_v[Y].$$

□

**Lemma 9.19.** *Let some constant $d \in \mathbb{R}$ be given. Then the function $\hat{v} : \mathbb{R} \to \mathbb{R}$, $\hat{v}(s) := v(s+d) - d$ induces the risk measure $\mathcal{R}_v$, i. e., $\mathcal{R}_{\hat{v}} \equiv \mathcal{R}_v$.*

*Proof.* For any random variable $X$ we have

$$\mathcal{R}_{\hat{v}}[X] = \inf_{t\in\mathbb{R}} t + \mathbb{E}[v(X-t+d)-d] \overset{\tilde{t}=t-d}{=} \inf_{\tilde{t}\in\mathbb{R}} \tilde{t} + \mathbb{E}[v(X-\tilde{t})] = \mathcal{R}_v[X].$$

□

**Theorem 9.20.** *If $v : \mathbb{R} \to \mathbb{R}$ is closed convex with $v(0) = 0$ and $v(s) > s$ for all $s \neq 0$, we obtain a regular risk measure $\mathcal{R}_v$ by the construction* (9.3)*. It is coherent if and only if it holds $v(s) = \max\{a_1 s, a_2 s\}$ with (w. l. o. g.) $a_1 \in [0, 1)$ and $a_2 \in (1, \infty)$, i. e., $v(s) = a_1 s$ for $s < 0$, $v(0) = 0$ and $v(s) = a_2 s$ for $s > 0$.*

*Proof.* This result is an excerpt of the "Expectation Theorem" in [96]. Note that the function $v$ is convex, monotonically increasing, and positively homogeneous. □

It follows that $\mathrm{CVaR}_\beta$ with $\beta \in (0, 1)$ is a coherent and regular risk measure since it belongs to the mentioned class with $a_1 = 0$ and $a_2 = \frac{1}{1-\beta}$.

### 9.3.2. Smoothing by a Log-Barrier Approach

Motivated by Theorem 9.20, we consider regular, coherent risk measures by choosing $a_1 \in [0,1)$ and $a_2 \in (1,\infty)$ and defining

$$\mathcal{R} : L^p_{\mathbb{P}}(\Xi) \to \mathbb{R},\ \mathcal{R}[X] := \inf_{t\in\mathbb{R}} t + \mathbb{E}[\max\{a_1(X-t), a_2(X-t)\}], \tag{9.4}$$

i. e., we have $\mathcal{R} \equiv \mathcal{R}_v$ with $v(s) = \max\{a_1 s, a_2 s\}$. In fact, as noted in [72, Sec. 2.4.1], this class of risk measures consists exactly of convex combinations of the expectation and the $\mathrm{CVaR}_\beta$ for every quantile parameter $\beta \in (0,1)$ and every combination parameter $\lambda \in (0,1]$:

$$\begin{aligned}
\lambda\,\mathrm{CVaR}_\beta[X] + (1-\lambda)\,\mathbb{E}[X] &= \inf_{t\in\mathbb{R}} \lambda\, t + \lambda\,\mathbb{E}[\tfrac{1}{1-\beta}(X-t)^+] + (1-\lambda)\,\mathbb{E}[X] \\
&= \inf_{t\in\mathbb{R}} t + \mathbb{E}[\tfrac{\lambda}{1-\beta}(X-t)^+ + (1-\lambda)(X-t)] \\
&= \inf_{t\in\mathbb{R}} t + \mathbb{E}[\max\{(1-\lambda)(X-t), (1-\lambda+\tfrac{\lambda}{1-\beta})(X-t)\}],
\end{aligned} \tag{9.5}$$

which fits into the setting (9.4) with $a_1 = 1-\lambda \in [0,1)$ and $a_2 = 1 + \frac{\beta}{1-\beta}\lambda \in (1,\infty)$. Conversely, $\lambda = 1 - a_1$ and $\beta = \frac{a_2-1}{a_2-a_1}$ can be computed. Due to the max-term in (9.4) one has to solve a nonsmooth optimization problem over $t \in \mathbb{R}$ to compute the risk measure $\mathcal{R}$. Furthermore, $\mathcal{R}$ itself is nonsmooth, see, e. g., [71]. Thus, we typically need a smoothing procedure that will allow us to develop derivative-based optimization methods. In [71] it is suggested to smooth CVaR using a suitable, smooth approximation of the $(\cdot)^+$-function so that many properties of the risk measure, such as convexity and monotonicity, can be preserved. Another advantage of this approach is the improved accuracy of quadrature formulas for the evaluation of the expectation. The convergence of sparse grid quadrature, for example, depends strongly on the smoothness of the integrand, see, e. g., [86, 47].

Alternatively, we can reformulate problem (9.4) and solve

$$\inf_{W\in L^p_{\mathbb{P}}(\Xi), t\in\mathbb{R}} t + \mathbb{E}[W] \quad \text{s. t.} \quad W \geq a_1(X-t) \text{ a. s.}, \quad W \geq a_2(X-t) \text{ a. s.}, \tag{9.6}$$

which is a linear optimization problem. Solving such a problem also requires the discretization of the space of random variables $L^p_{\mathbb{P}}(\Xi)$ to be able to optimize over $W$.

In [46], a problem with similar constraints as in (9.6), namely a stochastic obstacle problem, was solved successfully using a stochastic Galerkin discretization with a low-rank tensor representation of the coefficients, cf. Chapter 6, and a primal interior-point method implemented with low-rank tensors. Thus, we want to follow a similar approach here and replace the inequality constraints in (9.6) by a log-barrier term with a barrier parameter $\mu > 0$, see also [85, Sec. 19.6]. We obtain

$$\inf_{W\in L^p_{\mathbb{P}}(\Xi), t\in\mathbb{R}} t + \mathbb{E}[W] - \mu\,\mathbb{E}[\ln(W - a_1(X-t))] - \mu\,\mathbb{E}[\ln(W - a_2(X-t))] + \zeta(\mu) \tag{9.7}$$

or equivalently $\inf_{W\in L^p_{\mathbb{P}}(\Xi), t\in\mathbb{R}} F_\mu(X,t,W)$ with

$$F_\mu(X,t,W) := \mathbb{E}[t + W - \mu\,\ln(W - a_1(X-t)) - \mu\,\ln(W - a_2(X-t)) + \zeta(\mu)],$$

where $\zeta(\mu) := \mu\big(\ln(\frac{a_2-a_1}{a_2-1}\mu) + \ln(\frac{a_2-a_1}{1-a_1}\mu) - 2\big) \in \mathbb{R}$ is a constant shift. This shift is chosen such that problem (9.7) has exactly the minimal value $C \in \mathbb{R}$ whenever $X \equiv C$.

In the numerical experiments in [46] it was observed that the pointwise reciprocals of the functions $W - a_i(X - t)$, which appear in the barrier Newton system, cannot be represented sufficiently well by a low-rank tensor if the functions are too close to zero. Therefore, it would be beneficial to not push the barrier parameter $\mu$ arbitrarily close towards zero during the algorithm but to keep it on a fixed, positive level instead. Then it can be expected that the functions $W - a_i(X - t)$ are far enough away from zero during the iteration. We give a theoretical verification of this claim in (9.17).

Keeping the parameter $\mu$ bounded away from zero means that we solve a perturbed problem and obtain an approximation of the original risk measure. In the following, we want to investigate the properties of this approximate risk measure. Note that the goal of this procedure is not to construct new risk measures, but to provide an efficient solution algorithm based on low-rank tensor methods for, e. g., optimal control problems under uncertainty where risk-averse solutions are desired. Since this algorithm changes the properties of the underlying risk measure, which could have a great effect on the resulting solution, we analyze which properties are preserved by the log-barrier smoothing.

**Proposition 9.21.** *It holds that*

$$\inf_{W \in L^p_{\mathbb{P}}(\Xi), t \in \mathbb{R}} F_\mu(X, t, W) = \inf_{t \in \mathbb{R}} \mathbb{E}[\inf_{w \in \mathbb{R}} f_\mu(X(\cdot), t, w)]$$

*with* $f_\mu : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to (-\infty, +\infty]$,

$$f_\mu(x, t, w) = t + w - \mu \ln(w - a_1(x - t)) - \mu \ln(w - a_2(x - t)) + \zeta(\mu), \tag{9.8}$$

*where we set* $-\mu \ln(s) = +\infty$ *for every* $\mu > 0$ *and every* $s \leq 0$.

*Proof.*[11] Observe that

$$\inf_{W \in L^p_{\mathbb{P}}(\Xi), t \in \mathbb{R}} F_\mu(X, t, W) = \inf_{t \in \mathbb{R}} \inf_{W \in L^p_{\mathbb{P}}(\Xi)} \mathbb{E}[f_\mu(X(\cdot), t, W(\cdot))] \tag{9.9}$$

holds. The space $L^p_{\mathbb{P}}(\Xi)$ is decomposable in the sense of [97, Def. 14.59], the measure $\mathbb{P}$ is $\sigma$-finite, and the integrand $\hat{f}_\mu : \Xi \times \mathbb{R} \to \bar{\mathbb{R}}$, $\hat{f}_\mu(\xi, w) := f_\mu(X(\xi), t, w)$ with fixed $t \in \mathbb{R}$ and fixed $X$ is a normal integrand in the sense of [97, Def. 14.27], which we show in the following: Firstly, the function $(w, x) \mapsto f_\mu(x, t, w)$ is lower semicontinuous and convex for all $\xi \in \Xi$ because it is independent of $\xi$. Secondly, the interior of its domain is nonempty for all $\xi$. Thirdly, the function $\xi \mapsto f_\mu(x, t, w)$ is measurable for all $w \in \mathbb{R}$ since it is constant w. r. t. $\xi$. These properties yield that the function $(\xi, w, x) \mapsto f_\mu(x, t, w)$ is a normal integrand by [97, Prop. 14.39]. The composition rule [97, Prop. 14.45(c)] now shows that the function $\hat{f}_\mu$ is a normal integrand. Using this and having that there exists $\tilde{W} \in L^p_{\mathbb{P}}(\Xi)$ such that $\int_\Xi \hat{f}_\mu(\xi, \tilde{W}(\xi)) \, \mathrm{d}\mathbb{P} < \infty$, we can apply the "interchangeability theorem" [97, Thm. 14.60] to

[11] The essential content of this proof with the reference to [97] was provided by Prof. Dr. Thomas M. Surowiec.

derive

$$\begin{aligned}\inf_{W\in L^p_{\mathbb{P}}(\Xi)} \mathbb{E}[f_\mu(X(\cdot),t,W(\cdot))] &= \inf_{W\in L^p_{\mathbb{P}}(\Xi)} \int_\Xi \hat{f}_\mu(\xi, W(\xi))\,\mathrm{d}\mathbb{P} \\ &= \int_\Xi \inf_{w\in\mathbb{R}} \hat{f}_\mu(\xi,w)\,\mathrm{d}\mathbb{P} = \mathbb{E}[\inf_{w\in\mathbb{R}} f_\mu(X(\cdot),t,w)],\end{aligned}$$

which shows the desired result together with (9.9). □

In light of Proposition 9.21, we consider for each $t \in \mathbb{R}$, $\xi \in \Xi$, and $x = X(\xi)$ the one-dimensional problem

$$\min_{w\in\mathbb{R}} f_\mu(x,t,w),$$

where the unknown $w$ stands for $W(\xi)$.

**Proposition 9.22.** *The function $w \mapsto f_\mu(x,t,w)$ with $f_\mu$ defined in (9.8) has the unique minimizer*

$$\bar{w} = w_\mu(x-t) := \mu + \frac{(a_1+a_2)(x-t) + \sqrt{(a_1-a_2)^2(x-t)^2 + 4\mu^2}}{2} \tag{9.10}$$

*for every $x, t \in \mathbb{R}$.*

*Proof.* The function $f_\mu$ is finite for $w > a_i(x-t)$ ($i \in \{1,2\}$) and convex w. r. t. $w$. Furthermore, it is continuously differentiable w. r. t. $w$ on the set $\{(w,x,t) : w > a_i(x-t) \text{ for } i \in \{1,2\}\}$ and has the partial derivative

$$\tfrac{\partial}{\partial w} f_\mu(x,t,w) = 1 - \tfrac{\mu}{w-a_1(x-t)} - \tfrac{\mu}{w-a_2(x-t)},$$

which is zero only at the given stationary point $\bar{w}$. Therefore, this is the unique minimizer. The concrete computation of $\bar{w}$ can be found in Section A.6 in the appendix. □

By Propositions 9.21 and 9.22 we can define the partial solution of problem (9.7) given $X$ and $t$ as

$$\bar{W}_{X,t}(\xi) := \mu + \tfrac{1}{2}\big((a_1+a_2)(X(\xi)-t) + \sqrt{(a_1-a_2)^2(X(\xi)-t)^2 + 4\mu^2}\big).$$

Inserting this into $F_\mu$ gives the reduced problem

$$\inf_{t\in\mathbb{R}} F_\mu(X,t,\bar{W}_{X,t}) = \inf_{t\in\mathbb{R}} t + \mathbb{E}[v_\mu(X-t)] \tag{9.11}$$

which is similar to the initial problem and therefore defines a risk measure $\mathcal{R}_\mu \equiv \mathcal{R}_{v_\mu}$ induced by the function

$$v_\mu(s) := w_\mu(s) - \mu\,\ln(w_\mu(s) - a_1 s) - \mu\,\ln(w_\mu(s) - a_2 s) + \zeta(\mu)$$

having $w_\mu$ defined in (9.10). It remains to discuss and prove interesting properties of the corresponding risk measure $\mathcal{R}_\mu[X] := \inf_{t\in\mathbb{R}} t + \mathbb{E}[v_\mu(X-t)]$, which we call *the log-barrier risk measure.*

**Theorem 9.23.** *The log-barrier risk measure $\mathcal{R}_\mu : L^1_{\mathbb{P}}(\Xi) \to \mathbb{R}$ is well-defined, translation-equivariant, monotonic, and regular.*

*Proof.* The function $v_\mu$ is twice continuously differentiable with derivatives

$$v'_\mu(s) = w'_\mu(s) - \mu \frac{w'_\mu(s) - a_1}{w_\mu(s) - a_1 s} - \mu \frac{w'_\mu(s) - a_2}{w_\mu(s) - a_2 s},$$
$$v''_\mu(s) = \frac{\mu(a_2 - a_1)^2}{2\mu\sqrt{(a_2 - a_1)^2 s^2 + 4\mu^2} + (a_2 - a_1)^2 s^2 + 4\mu^2},$$

where $w_\mu(s) = \mu + \frac{(a_1+a_2)s + \sqrt{(a_1-a_2)^2 s^2 + 4\mu^2}}{2}$ and $w'_\mu(s) = \frac{a_1+a_2}{2} + \frac{(a_1-a_2)^2 s}{2\sqrt{(a_1-a_2)^2 s^2 + 4\mu^2}}$. For the computation of the derivatives see Section A.6.

- Translation equivariance follows directly from Lemma 9.13.
- We have that $\lim_{s\to-\infty} v'_\mu(s) = a_1$ and $\lim_{s\to+\infty} v'_\mu(s) = a_2$, see Section A.6. Since $v'_\mu$ is strictly increasing ($v''_\mu(s) > 0$), $v'_\mu(\mathbb{R}) = (a_1, a_2) \subset (0, \infty)$ follows. Therefore, the function $v_\mu$ itself is strictly increasing. Lemma 9.18 yields that $\mathcal{R}_\mu$ is monotonic.
- Since $v''_\mu(s) > 0$ holds for all $s \in \mathbb{R}$, the function $v_\mu$ is strictly convex. Now, we choose $d := \frac{2-a_1-a_2}{(1-a_1)(a_2-1)}\mu \in \mathbb{R}$. By Lemma 9.19, it holds that $\mathcal{R}_\mu$ is also induced by the function $\hat{v}_\mu(s) = v_\mu(s + d) - d$, which is also strictly convex. One can compute

  $$\hat{v}_\mu(0) = v_\mu(d) - d = \mu\left(2 - \ln\left(\tfrac{a_2-a_1}{a_2-1}\mu\right) - \ln\left(\tfrac{a_2-a_1}{1-a_1}\mu\right)\right) + \zeta(\mu) = 0 \tag{9.12}$$

  and $\hat{v}'_\mu(0) = v'_\mu(d) = 1$, see Section A.6. This yields—together with strict convexity—that $\hat{v}_\mu(s) > s$ for all $s \in \mathbb{R} \setminus \{0\}$. From Theorem 9.20 we get that $\mathcal{R}_\mu$ is regular. Equation (9.12) again motivates the choice of the shift $\zeta(\mu)$.
- Corollary 9.11 yields that the log-barrier risk measure is proper. Due to Lemma 9.12 it is always finite on $L^1_{\mathbb{P}}(\Xi)$:

  $$|v_\mu(s)| = \left|\int_0^s v'_\mu(s)\,\mathrm{d}s + v_\mu(0)\right| \leq a_2\,|s| + |v_\mu(0)|$$

  holds for all $s \in \mathbb{R}$ due to the computed range of $v'_\mu$.

□

The function $\hat{v}_\mu$ defined in the proof of Theorem Theorem 9.23 is plotted in Figure 9.1 two different values of $a_1 = 1 - \lambda$, $a_2 = 1 + \frac{\beta}{1-\beta}\lambda$, and $\mu$. The properties $\hat{v}_\mu(0) = 0$ and $\hat{v}'_\mu(0) = 1$ are clearly visible and the approximation quality depending on the log-barrier parameter $\mu$ is depicted.

**Remark 9.24.** For $\mu > 0$ and any choice of $a_1 \in [0, 1)$ and $a_2 \in (1, \infty)$, the log-barrier risk measure is not a coherent measure of risk by Theorem 9.20 since $v_\mu$ is clearly *not* the maximum of two linear functions.

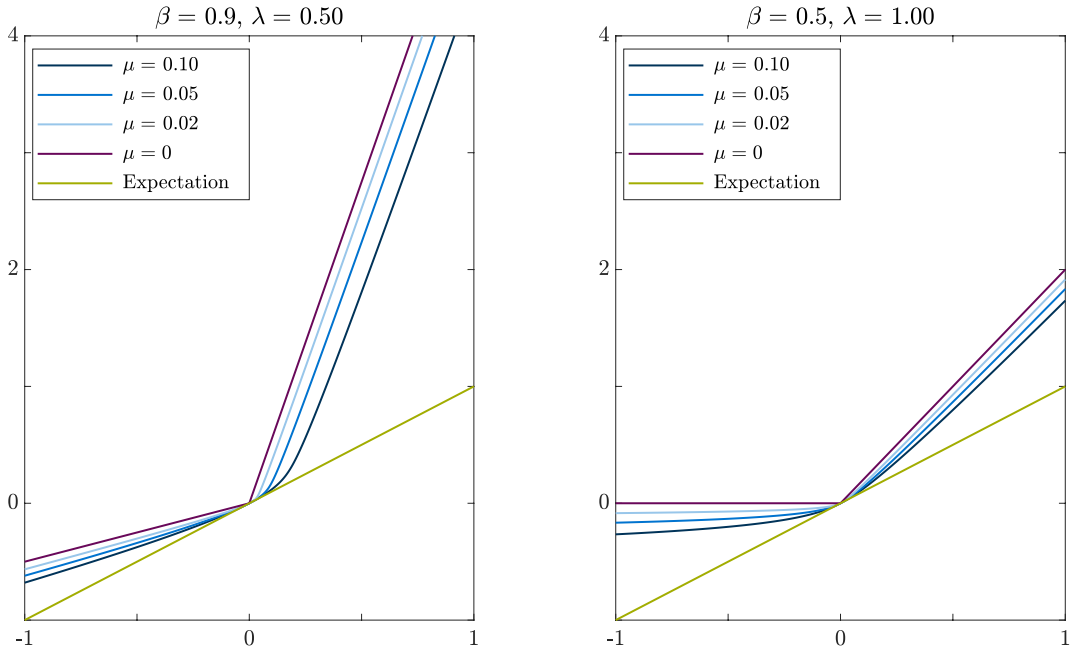



Figure 9.1.: Plots of the function $\hat{v}_\mu$ inducing the risk measure $\mathcal{R}_\mu$ for different values of $a_1 = 1 - \lambda$, $a_2 = 1 + \frac{\beta}{1-\beta}\lambda$, and $\mu$. The green line corresponds to $v(t) = t$, which induces the risk measure $\mathcal{R}_v \equiv \mathbb{E}$.

### 9.3.3. Application to Optimal Control under Uncertainty

Consider now the setting from Section 9.2 and let Assumption 9.9 hold with $p_{\text{fun}} \in [1, \infty]$. Furthermore, we define $\hat{J} : U \times \Xi \to \mathbb{R}, \hat{J}(u, \xi) := \hat{J}_1(u, \xi) + J_2(u)$ and set $U_{\text{ad}} = U$ for simplicity. We have that $\hat{J}(u, \cdot) \in L^{p_{\text{fun}}}_{\mathbb{P}}(\Xi)$ for all $u \in U$. We consider the optimization problem under uncertainty

$$\min_{u \in U} \mathcal{R}_\mu[\hat{J}(u, \cdot)] \tag{9.13}$$

with the derived log-barrier risk measure. Since $\mathcal{R}_\mu$ is translation-equivariant and $J_2$ does not depend on $\xi$, this corresponds to the general setting (1.3) with $U_{\text{ad}} \equiv U$. Clearly, (9.13) is an approximation of an original problem with the nonsmooth risk measure $\mathcal{R}$, e. g., an optimal control problem with the conditional value-at-risk.

As derived before, we can write (9.13) as

$$\min_{u \in U, t \in \mathbb{R}, W \in L^p_{\mathbb{P}}(\Xi)} \hat{F}_\mu(u, t, W) \quad \text{s. t.} \quad W - a_i(\hat{J}(u, \cdot) - t)) \geq 0 \ \text{ a. s. for } i \in \{1, 2\}, \tag{9.14}$$

with

$$\hat{F}_\mu(u, t, W) := \mathbb{E}[t + W - \mu \sum_{i=1}^{2} \ln(W - a_i(\hat{J}(u, \cdot) - t))] + \zeta(\mu), \tag{9.15}$$

$a_1 \in [0, 1)$, $a_2 \in (1, \infty)$, $\mu > 0$, and $\zeta(\mu) = \mu(\ln(\frac{a_2 - a_1}{a_2 - 1}\mu) + \ln(\frac{a_2 - a_1}{1 - a_1}\mu) - 2)$.

**Lemma 9.25.** *The function $\hat{F}_\mu$ defined in (9.15) is convex if the function $u \mapsto \hat{J}(u,\xi)$ is convex for a. e. $\xi \in \Xi$.*

*Proof.* Let $\lambda \in [0,1]$, $u^1, u^2 \in U$, $t_1, t_2 \in \mathbb{R}$, $W_1, W_2 \in L^p_{\mathbb{P}}(\Xi)$. Clearly, the affine part of the function $\hat{F}_\mu$ is convex. Since $\mu > 0$, it remains to show that

$$(u,t,W) \mapsto \mathbb{E}[-\ln(W - a_i\hat{J}(u,\cdot) + a_i t)]$$

is convex for $i \in \{1,2\}$. Due to convexity, $\hat{J}(\lambda u^1 + (1-\lambda)u^2, \xi) \leq \lambda\hat{J}(u^1,\xi) + (1-\lambda)\hat{J}(u^2,\xi)$ holds for a. e. $\xi \in \Xi$. It follows that

$$\begin{aligned} &\lambda W_1(\xi) + (1-\lambda)W_2(\xi) - a_i\hat{J}(\lambda u^1 + (1-\lambda)u^2, \xi) + a_i\lambda t_1 + a_i(1-\lambda)t_2 \\ \geq\ &\lambda(W_1(\xi) - a_i\hat{J}(u^1,\xi) + a_i t_1) + (1-\lambda)(W_2(\xi) - a_i\hat{J}(u^2,\xi) + a_i t_2) \end{aligned}$$

for a. e. $\xi \in \Xi$. Applying the negative logarithm, which is decreasing and convex, on both sides yields

$$\begin{aligned} &-\ln\big(\lambda W_1(\xi) + (1-\lambda)W_2(\xi) - a_i\hat{J}(\lambda u^1 + (1-\lambda)u^2, \xi) + a_i\lambda t_1 + a_i(1-\lambda)t_2\big) \\ \leq\ &-\lambda\ln\big(W_1(\xi) - a_i\hat{J}(u^1,\xi) + a_i t_1\big) + (1-\lambda)\ln\big(W_2(\xi) - a_i\hat{J}(u^2,\xi) + a_i t_2\big) \end{aligned}$$

for a. e. $\xi \in \Xi$. The monotonicity of the expectation establishes the convexity of $\hat{F}_\mu$. □

**Example 9.26.** If $\varphi' \equiv 0$ in the example from Chapter 3, we have a linear state equation and thus a convex reduced objective function $u \mapsto \hat{J}(u,\xi)$ for a. e. $\xi \in \Xi$.

Given $u \in U$, $t \in \mathbb{R}$, we consider the partial problem

$$\min_{W \in L^p_{\mathbb{P}}(\Xi)} \hat{F}_\mu(u,t,W) \quad \text{s. t.} \quad W \geq a_i(\hat{J}(u,\cdot) - t)) \ \text{ a. s. for } i \in \{1,2\}.$$

We have already proven that this problem has the unique global solution $\bar{W}_{u,t} \in L^p_{\mathbb{P}}(\Xi)$ given by

$$\bar{W}_{u,t}(\xi) := \mu + \tfrac{a_1+a_2}{2}(\hat{J}(u,\xi) - t) + \tfrac{1}{2}\sqrt{(a_1-a_2)^2(\hat{J}(u,\xi)-t)^2 + 4\mu^2}. \tag{9.16}$$

This partial solution fulfills

$$\bar{W}_{u,t} - a_i(\hat{J}(u,\cdot) - t)) = \mu \pm \tfrac{a_2-a_1}{2}(\hat{J}(u,\xi) - t) + \tfrac{1}{2}\sqrt{(a_1-a_2)^2(\hat{J}(u,\xi)-t)^2 + 4\mu^2} \geq \mu, \tag{9.17}$$

i. e., if (9.14) has a solution $(\bar{u}, \bar{t}, \bar{W})$, the constraints are uniformly inactive in an $L^\infty$ sense there. Therefore, we can restrict (9.14) to

$$\min_{u \in U, t \in \mathbb{R}, W \in L^p_{\mathbb{P}}(\Xi)} \hat{F}_\mu(u,t,W) \quad \text{s. t.} \quad W - a_i(\hat{J}(u,\cdot) - t) \geq \mu \ \text{ a. s. for } i \in \{1,2\} \tag{9.18}$$

without changing its global solution, provided it exists. We see that the objective function is finite on the feasible set.

### Existence of a Solution

We derive conditions for the existence of a solution of (9.14) or, equivalently, (9.18). For this purpose, we first bound the auxiliary function $f_\mu$:

**Lemma 9.27.** *For any $d \in (a_1, a_2)$ and all $x, w, t \in \mathbb{R}$ it holds that*

$$f_\mu(x,t,w) \geq dx + (1-d)t + \tilde{\zeta}_d(\mu) \tag{9.19}$$

*with the function $f_\mu$ defined in (9.8), and with $\tilde{\zeta}_d(\mu) := \mu\big(\ln(\frac{a_2-d}{a_2-1}) + \ln(\frac{d-a_1}{1-a_1})\big)$.*

*Proof.* We use the fact that $-\mu\,\ln(s) \geq -\mu\,\ln(b) - \mu\,\ln'(b)(s-b) = -\frac{\mu}{b}s + \mu - \mu\,\ln(b)$ holds for all $s, b \in (0,\infty)$ due to convexity of the negative logarithm and $\mu > 0$. Now with $b_1 = c_1\mu$ and $b_2 = c_2\mu$, $c_1, c_2 > 0$, we obtain

$$\begin{aligned} &f_\mu(x,t,w) = t + w + \sum_{i=1}^{2} \big(-\mu\,\ln(w - a_i(x-t))\big) + \zeta(\mu) \geq \\ &t + w + \sum_{i=1}^{2} \big(-\tfrac{1}{c_i}(w - a_i(x-t)) + \mu - \mu\,\ln(c_i\mu)\big) + \zeta(\mu) = \\ &\big(\tfrac{a_1}{c_1} + \tfrac{a_2}{c_2}\big)x + \big(1 - \tfrac{a_1}{c_1} - \tfrac{a_2}{c_2}\big)t + \big(1 - \tfrac{1}{c_1} - \tfrac{1}{c_2}\big)w + \mu(2 - \ln(c_1\mu) - \ln(c_2\mu)) + \zeta(\mu). \end{aligned} \tag{9.20}$$

Choosing $c_1 = \frac{a_2-a_1}{a_2-d} > 0$ and $c_2 = \frac{a_2-a_1}{d-a_1} > 0$ (giving $\frac{1}{c_1} + \frac{1}{c_2} = 1$ and $\frac{a_1}{c_1} + \frac{a_2}{c_2} = d$) in (9.20) yields

$$\begin{aligned} f_\mu(x,t,w) &\geq dx + (1-d)t + \mu(2 - \ln(\tfrac{a_2-a_1}{a_2-d}\mu) - \ln(\tfrac{a_2-a_1}{d-a_1}\mu)) + \zeta(\mu) \\ &= dx + (1-d)t + \tilde{\zeta}_d(\mu). \end{aligned}$$

□

**Corollary 9.28.** *It holds that*

$$\hat{F}_\mu(u,t,W) = \mathbb{E}[f_\mu(\hat{J}(u,\cdot),t,W(\cdot))] \geq \mathbb{E}[\hat{J}(u,\cdot)]$$

*for all $(u,t,W) \in U \times \mathbb{R} \times L^p_{\mathbb{P}}(\Xi)$ with $\hat{F}_\mu$ defined in (9.15).*

*Proof.* Choosing $d = 1 \in (a_1, a_2)$ in (9.19) results in $f_\mu(x,t,w) \geq x$ for all $x, t, w \in \mathbb{R}$, which shows the result. □

Thus, if the function $u \mapsto \mathbb{E}[\hat{J}(u,\cdot)]$ is bounded from below on $U$, the function $\hat{F}_\mu$ is bounded from below on $U \times \mathbb{R} \times L^p_{\mathbb{P}}(\Xi)$. To prove existence of a solution, we restrict problem (9.18) to a bounded, closed, convex feasible set. Since the partial solution $\bar{W}_{u,t}$ is given for any $u \in U$, $t \in \mathbb{R}$, we bound the feasible set for $u$ and $t$.

**Lemma 9.29.** *Let $(\tilde{u}, \tilde{t}, \tilde{W})$ and $(u^*, t^*, W^*)$ be feasible points[12] of (9.18) such that*

$$\hat{F}_\mu(u^*, t^*, W^*) \leq \hat{F}_\mu(\tilde{u}, \tilde{t}, \tilde{W}),$$

*and let $u \mapsto \mathbb{E}[\hat{J}(u,\cdot)]$ be bounded from below on $U$.*

[12] A feasible point can be constructed simply: Given some $\tilde{u} \in U$, $\tilde{t} = 0$, we choose, e. g., $\tilde{W}(\xi) = \max_{i\in\{1,2\}} a_i(\hat{J}(\tilde{u},\xi)) + \mu$.

*Then it holds that*

$$\mathbb{E}[\hat{J}(u^*,\cdot)] \le \hat{F}_\mu(\tilde{u},\tilde{t},\tilde{W}), \text{ and}$$
$$t^* \in [\tilde{t}_2,\tilde{t}_1], \qquad \tilde{t}_i := \tfrac{1}{1-d_i}\big(\hat{F}_\mu(\tilde{u},\tilde{t},\tilde{W}) - d_i \inf_{\hat{u}\in U} \mathbb{E}[\hat{J}(\hat{u},\cdot)] - \tilde{\zeta}_{d_i}(\mu)\big) \quad (i\in\{1,2\})$$

*with any fixed* $d_1 \in (a_1,1)$, $d_2\in(1,a_2)$, *and* $\tilde{\zeta}_d(\mu) = \mu\big(\ln(\frac{a_2-d}{a_2-1}) + \ln(\frac{d-a_1}{1-a_1})\big)$.

*Proof.* With the feasible point $(\tilde{u},\tilde{t},\tilde{W})$ and with $u\mapsto\mathbb{E}[\hat{J}(u,\cdot)]$ being bounded from below on $U$, we can bound $u^*$ and $t^*$ as follows: Observe that $\hat{F}_\mu(u^*,t^*,W^*) \le \hat{F}_\mu(\tilde{u},\tilde{t},\tilde{W})$ implies

$$d\,\mathbb{E}[\hat{J}(u^*,\cdot)] + (1-d)t^* + \tilde{\zeta}_d(\mu) \le \hat{F}_\mu(\tilde{u},\tilde{t},\tilde{W}) \text{ for all } d\in(a_1,a_2)$$

by Lemma 9.27. Choosing $d_0 = 1$ as in the proof of Corollary 9.28, $d_1\in(a_1,1)$ and $d_2\in(1,a_2)$, this gives

$$\begin{aligned}\mathbb{E}[\hat{J}(u^*,\cdot)] &\le \hat{F}_\mu(\tilde{u},\tilde{t},\tilde{W}),\\ t^* &\le \tfrac{1}{1-d_1}\big(\hat{F}_\mu(\tilde{u},\tilde{t},\tilde{W}) - d_1\,\mathbb{E}[\hat{J}(u^*,\cdot)] - \tilde{\zeta}_{d_1}(\mu)\big),\\ t^* &\ge \tfrac{1}{1-d_2}\big(\hat{F}_\mu(\tilde{u},\tilde{t},\tilde{W}) - d_2\,\mathbb{E}[\hat{J}(u^*,\cdot)] - \tilde{\zeta}_{d_2}(\mu)\big)\end{aligned}$$

and thus

$$\begin{aligned}\mathbb{E}[\hat{J}(u^*,\cdot)] &\le \hat{F}_\mu(\tilde{u},\tilde{t},\tilde{W}),\\ t^* &\le \tfrac{1}{1-d_1}\big(\hat{F}_\mu(\tilde{u},\tilde{t},\tilde{W}) - d_1 \inf_{\hat{u}\in U}\mathbb{E}[\hat{J}(\hat{u},\cdot)] - \tilde{\zeta}_{d_1}(\mu)\big),\\ t^* &\ge \tfrac{1}{1-d_2}\big(\hat{F}_\mu(\tilde{u},\tilde{t},\tilde{W}) - d_2 \inf_{\hat{u}\in U}\mathbb{E}[\hat{J}(\hat{u},\cdot)] - \tilde{\zeta}_{d_2}(\mu)\big),\end{aligned}$$

which shows the result. Note that $-\frac{d_1}{1-d_1} < 0$ whereas $-\frac{d_2}{1-d_2} > 0$. □

We see that, if $u\mapsto\mathbb{E}[\hat{J}(u,\cdot)]$ is coercive, i. e., $\lim_{\|u\|_U\to\infty}\mathbb{E}[\hat{J}(u,\cdot)] = \infty$, we can restrict the variable $u$ in problem (9.14) to a nonempty, convex, closed, bounded set $U_{\mathrm{ad}}\subset U$ with

$$\tilde{U}_{\mathrm{ad}} := \{u : \mathbb{E}[\hat{J}(u,\cdot)] \le \hat{F}_\mu(\tilde{u},\tilde{t},\tilde{W})\} \subset U_{\mathrm{ad}}$$

and the variable $t$ to the compact interval given in Lemma 9.29 without changing the global solution of problem (9.18), provided it exists. Note that the set $\tilde{U}_{\mathrm{ad}}$ is bounded.

**Remark 9.30.** Alternatively, the theory would also work if the constraint $u\in U_{\mathrm{ad}}$ with a nonempty, bounded, closed, convex set $U_{\mathrm{ad}}\subset U$ was already posed in (9.14) and if $u\mapsto\mathbb{E}[\hat{J}(u,\cdot)]$ was bounded from below on this set, but not necessarily coercive on $U$.

Moreover, the optimal partial solution $\bar{W}_{u,t}$ given $(u,t)$ can be inserted. This yields the problem

$$\begin{aligned}\min_{u\in U,t\in\mathbb{R},W\in L^p_{\mathbb{P}}(\Xi)}\ & \hat{F}_\mu(u,t,W)\\ \text{s. t.}\ & \mathbb{E}[\hat{J}(u,\cdot)] \le \hat{F}_\mu(\tilde{u},\tilde{t},\tilde{W}), \qquad \tilde{t}_2 \le t \le \tilde{t}_1,\\ & W(\cdot) = \mu + \tfrac{a_1+a_2}{2}(\hat{J}(u,\cdot)-t) + \tfrac{1}{2}\sqrt{(a_1-a_2)^2(\hat{J}(u,\cdot)-t)^2 + 4\mu^2} \text{ a. s.}\end{aligned} \tag{9.21}$$

Based on (9.21), we derive an existence result, which can be applied to convex problems. For a more general framework for risk-averse, PDE-constrained optimization including existence and optimality conditions we refer to [72].

**Proposition 9.31.** *Let the function $(u,t) \mapsto \hat{F}_\mu(u,t,\bar{W}_{u,t}) \in \mathbb{R}$ with $\bar{W}_{u,t}$ defined in* (9.16) *be sequentially weakly lower semicontinuous. Furthermore, let the function $u \mapsto \mathbb{E}[\hat{J}(u,\cdot)]$ be coercive on $U$. Then, problem* (9.14) *has a solution.*

*Proof.* Since $u \mapsto \mathbb{E}[\hat{J}(u,\cdot)]$ is coercive on $U$, it is bounded from below and problem (9.14) is equivalent to problem (9.21). As derived, this is equivalent to

$$\min_{u\in U, t\in\mathbb{R}} \hat{F}_\mu(u,t,\bar{W}_{u,t}) \quad \text{s.t.} \quad u \in U_{\mathrm{ad}},\ t \in [\tilde{t}_2, \tilde{t}_1].$$

Now, a feasible, infimizing sequence for this problem is bounded and, therefore, it has a weakly convergent subsequence with limit $(\bar{u},\bar{t}) \in U \times \mathbb{R}$ because $U$ is reflexive and $U_{\mathrm{ad}}$ is closed, convex, bounded. Sequential lower semicontinuity of the objective function w. r. t. the weak topology yields that $(\bar{u},\bar{t})$ solves the problem. Moreover, $(\bar{u},\bar{t},\bar{W}_{\bar{u},\bar{t}})$ solves (9.14). □

**Lemma 9.32.** *Let the function $u \mapsto \hat{J}(u,\xi) \in \mathbb{R}$ be convex and strongly continuous for a. e. $\xi \in \Xi$ and let the function $u \mapsto \hat{J}(u,\cdot) \in L^1_{\mathbb{P}}(\Xi)$ be such that for every $u \in U$ there exist $\varepsilon_u > 0$ and a function $g_u \in L^1_{\mathbb{P}}(\Xi)$ such that $|\hat{J}(\tilde{u},\xi)| \le g_u(\xi)$ holds for every $\tilde{u}$ with $\|\tilde{u}-u\|_U < \varepsilon_u$ and a. e. $\xi \in \Xi$. Then, the function $(u,t) \mapsto \hat{F}_\mu(u,t,\bar{W}_{u,t}) \in \mathbb{R}$ is weakly lower semicontinuous.*

*Proof.* We observe that $\hat{F}_\mu(u,t,\bar{W}_{u,t}) = t + \mathbb{E}[v_\mu(\hat{J}(u,\cdot)-t)]$ holds, see (9.11). The convexity and the monotonicity of the function $v_\mu$ as well as the monotonicity of the expectation yield that $(u,t) \mapsto \hat{F}_\mu(u,t,\bar{W}_{u,t})$ is convex, cf. Lemma 9.25. Now let $(u^k,t_k)_{k\in\mathbb{N}} \subset U \times \mathbb{R}$ be a sequence converging strongly to some $(u,t) \in U \times \mathbb{R}$. W. l. o. g. we assume that $\|u^k - u\|_U < \varepsilon_u$ holds for every $k$. Then we have that $\hat{J}(u^k,\xi) - t_k \to \hat{J}(u,\xi) - t$ and thus $v_\mu(\hat{J}(u^k,\xi) - t_k) \to v_\mu(\hat{J}(u,\xi) - t)$ for a. e. $\xi \in \Xi$. Furthermore, $|v_\mu(\hat{J}(u^k,\xi) - t_k)| \le a_2(g_u(\xi) + \sup_{k\in\mathbb{N}} |t_k|) + |v_\mu(0)|$ holds for every $k$ and a. e. $\xi \in \Xi$, cf. the proof of Theorem 9.23. As $\sup_{k\in\mathbb{N}} |t_k| < \infty$, this bound is an $L^1_{\mathbb{P}}(\Xi)$-function and independent of $k$. Hence, the dominated convergence theorem yields that $t_k + \mathbb{E}[v_\mu(\hat{J}(u^k,\xi) - t_k)] \to t + \mathbb{E}[v_\mu(\hat{J}(u,\cdot) - t)]$ as $k \to \infty$. This proves continuity of $(u,t) \mapsto \hat{F}_\mu(u,t,\bar{W}_{u,t})$ w. r. t. strong convergence. Together with convexity, this yields lower semicontinuity w. r. t. the weak topology. □

**Example 9.33.** In the example from Chapter 3, the function $u \mapsto \mathbb{E}[\hat{J}(u,\cdot)]$ is coercive because it consists of a non-negative tracking term and the coercive regularization term $\frac{\gamma}{2}\|u\|_U^2$. The function $u \mapsto \hat{J}(u,\xi)$ is strongly continuous for a. e. $\xi \in \Xi$ and we have

$$\begin{aligned}
&|\hat{J}(\tilde{u},\xi)| \le \tfrac{1}{2}\big(\|Q(\xi)\|_{\mathcal{L}(Y,H)}\|S[\xi](\tilde{u})\|_Y + \|\hat{q}(\xi)\|_H\big)^2 + \tfrac{\gamma}{2}\|\tilde{u}\|_U^2 \le \\
&\tfrac{1}{2}\Big(\tfrac{C_\Omega}{\underline{\kappa}}\|Q(\xi)\|_{\mathcal{L}(Y,H)}(\|D\|_{\mathcal{L}(U,L^2(\Omega))}\|\tilde{u}\|_U + \|f(\xi)-\varphi(0)\|_{L^2(\Omega)}) + \|\hat{q}(\xi)\|_H\Big)^2 + \tfrac{\gamma}{2}\|\tilde{u}\|_U^2 \le \\
&\tfrac{1}{2}\Big(\tfrac{C_\Omega}{\underline{\kappa}}\|Q(\xi)\|_{\mathcal{L}(Y,H)}(\|D\|_{\mathcal{L}(U,L^2(\Omega))}(\|u\|_U + \varepsilon_u) + \|f(\xi)-\varphi(0)\|_{L^2(\Omega)}) + \|\hat{q}(\xi)\|_H\Big)^2 \\
&\qquad + \tfrac{\gamma}{2}(\|u\|_U + \varepsilon_u)^2,
\end{aligned}$$

for all $u, \tilde{u} \in U$ and $\varepsilon_u > 0$ with $\|\tilde{u} - u\|_U < \varepsilon_u$, cf. the discussion in Subsection 9.2.1. The derived bound is an $L^{\hat{r}}_{\mathbb{P}}(\Xi)$-function w. r. t. $\xi$ with $\hat{r}$ from Table 3.1 and hence an $L^1_{\mathbb{P}}(\Xi)$-function. Furthermore, if $\varphi' \equiv 0$, the function $u \mapsto \hat{J}(u, \xi)$ is convex for a. e. $\xi \in \Xi$, cf. Example 9.26. We therefore can apply Lemma 9.32 and Proposition 9.31 to derive the existence of a solution.

### Differentiability Properties of the Function $\hat{F}_\mu$

For deriving first-order optimality conditions, we restrict the discussion to the case $p = \infty$. Then, the feasible set of (9.18) has interior points and Fréchet differentiability can be proven. In contrast to that, the case $p < \infty$ requires more sophisticated concepts, see [112].

We consider the function

$$\begin{aligned}
&\tilde{F}_\mu : \{(X, t, W) \in L^\infty_{\mathbb{P}}(\Xi) \times \mathbb{R} \times L^\infty_{\mathbb{P}}(\Xi) : W \geq a_i(X - t) + c_\mu \mu \text{ a. s. for } i \in \{1, 2\}\} \to \mathbb{R}, \\
&\tilde{F}_\mu(X, t, W) := \mathbb{E}[t + W - \mu \ln(W - a_1(X - t)) - \mu \ln(W - a_2(X - t)) + \zeta(\mu)] \qquad (9.22)
\end{aligned}$$

with some $c_\mu \in (0, 1)$. This function is well-defined on the given feasible set. We have enlarged the latter compared to the feasible set of (9.18) to prove differentiability properties on an $L^\infty$-neighborhood.

**Proposition 9.34.** *The function $\tilde{F}_\mu$ defined in* (9.22) *is twice continuously differentiable on the given feasible set with the F-derivatives*

$$\begin{aligned}
\nabla_X \tilde{F}_\mu(X, t, W) &= \mu \sum_{i=1}^{2} \tfrac{a_i}{W - a_i(X-t)} \in L^1_{\mathbb{P}}(\Xi), \\
\nabla_t \tilde{F}_\mu(X, t, W) &= 1 - \mu \sum_{i=1}^{2} \mathbb{E}\big[\tfrac{a_i}{W - a_i(X-t)}\big] \in \mathbb{R}, \\
\nabla_W \tilde{F}_\mu(X, t, W) &= \mathbb{1} - \mu \sum_{i=1}^{2} \tfrac{1}{W - a_i(X-t)} \in L^1_{\mathbb{P}}(\Xi), \\
\nabla^2_{XX} \tilde{F}_\mu(X, t, W)S &= \mu \sum_{i=1}^{2} \tfrac{a_i^2 S}{(W - a_i(X-t))^2} \in L^1_{\mathbb{P}}(\Xi), \\
\nabla^2_{tX} \tilde{F}_\mu(X, t, W)\tau &= -\mu \sum_{i=1}^{2} \tfrac{a_i^2}{(W - a_i(X-t))^2}\tau \in L^1_{\mathbb{P}}(\Xi), \\
\nabla^2_{WX} \tilde{F}_\mu(X, t, W)S &= -\mu \sum_{i=1}^{2} \tfrac{a_i S}{(W - a_i(X-t))^2} \in L^1_{\mathbb{P}}(\Xi), \\
\nabla^2_{Xt} \tilde{F}_\mu(X, t, W)S &= -\mu \sum_{i=1}^{2} \tfrac{a_i^2 S}{(W - a_i(X-t))^2} \in L^1_{\mathbb{P}}(\Xi), \\
\nabla^2_{tt} \tilde{F}_\mu(X, t, W)\tau &= \mu \sum_{i=1}^{2} \mathbb{E}\big[\tfrac{a_i^2}{(W - a_i(X-t))^2}\big]\tau \in \mathbb{R},
\end{aligned}$$

$$\nabla^2_{Wt}\tilde{F}_\mu(X,t,W)S = \mu\sum_{i=1}^{2}\tfrac{a_iS}{(W-a_i(X-t))^2} \in L^1_{\mathbb{P}}(\Xi),$$
$$\nabla^2_{XW}\tilde{F}_\mu(X,t,W)S = -\mu\sum_{i=1}^{2}\tfrac{a_iS}{(W-a_i(X-t))^2} \in L^1_{\mathbb{P}}(\Xi),$$
$$\nabla^2_{tW}\tilde{F}_\mu(X,t,W)\tau = \mu\sum_{i=1}^{2}\tfrac{a_i}{(W-a_i(X-t))^2}\tau \in L^1_{\mathbb{P}}(\Xi),$$
$$\nabla^2_{WW}\tilde{F}_\mu(X,t,W)S = \mu\sum_{i=1}^{2}\tfrac{S}{(W-a_i(X-t))^2} \in L^1_{\mathbb{P}}(\Xi),$$

*where $S \in L^\infty_{\mathbb{P}}(\Xi)$ and $\tau \in \mathbb{R}$. We write $\nabla_X\tilde{F}_\mu(X,t,W) \in L^1_{\mathbb{P}}(\Xi)$ for the representative of the partial derivative $(\tilde{F}_\mu)_X(X,t,W) \in L^\infty_{\mathbb{P}}(\Xi)^*$ etc.*

*Proof.* Clearly, the linear part $(X,t,W) \mapsto t+\mathbb{E}[W]+\zeta(\mu)$ is bounded and twice continuously differentiable. Therefore, we consider only $(X,t,W) \mapsto \mathbb{E}[\ln(W-a_i(X-t))]$. Let $(X,t,W) \in L^\infty_{\mathbb{P}}(\Xi)\times\mathbb{R}\times L^\infty_{\mathbb{P}}(\Xi)$ be such that $W \geq a_i(X-t)+c_\mu\mu$ holds a. s. Consider a sequence $(S_k)_{k\in\mathbb{N}} \subset L^\infty_{\mathbb{P}}(\Xi)$ such that $0 < \|S_k\|_{L^\infty_{\mathbb{P}}(\Xi)} \leq \frac{c_\mu\mu}{2}$ for all $k \in \mathbb{N}$ and $\|S_k\|_{L^\infty_{\mathbb{P}}(\Xi)} \to 0$ as $k \to \infty$. We have

$$\begin{aligned}
&\frac{1}{\|S_k\|_{L^\infty_{\mathbb{P}}(\Xi)}}\Big|\mathbb{E}\big[\ln\big(W+S_k-a_i(X-t)\big)-\ln\big(W-a_i(X-t)\big)-\tfrac{S_k}{W-a_i(X-t)}\big]\Big| \\
&\leq \mathbb{E}\Big[\int_0^1 \big|\tfrac{1}{W+\sigma S_k-a_i(X-t)}-\tfrac{1}{W-a_i(X-t)}\big|\,\mathrm{d}\sigma\Big] \\
&= \mathbb{E}\Big[\int_0^1 \tfrac{\sigma|S_k|}{|(W+\sigma S_k-a_i(X-t))(W-a_i(X-t))|}\,\mathrm{d}\sigma\Big] \\
&\leq \mathbb{E}\Big[\int_0^1 \tfrac{\sigma}{|(W+\sigma S_k-a_i(X-t))(W-a_i(X-t))|}\,\mathrm{d}\sigma\Big]\,\|S_k\|_{L^\infty_{\mathbb{P}}(\Xi)} \leq \tfrac{2}{c_\mu^2\mu^2}\|S_k\|_{L^\infty_{\mathbb{P}}(\Xi)}.
\end{aligned}$$

Note that all appearing functions are $L^\infty$-functions due to

$$W+\sigma S_k-a_i(X-t) \in \big[\tfrac{(2-\sigma)c_\mu\mu}{2}, \|W\|_{L^\infty_{\mathbb{P}}(\Xi)}+\tfrac{\sigma c_\mu\mu}{2}+a_i\|X-t\|_{L^\infty_{\mathbb{P}}(\Xi)}\big] \text{ a. s.}$$

for all $\sigma \in [0,1]$ etc. The estimation proves F-differentiability w. r. t. $W$ with the given derivative. Analogously, the function $\tilde{F}_\mu$ is F-differentiable w. r. t. $X$ and $t$.

Similarly to the estimation above, we have

$$\begin{aligned}
&\frac{1}{\|S_k\|_{L^\infty_{\mathbb{P}}(\Xi)}}\Big\|\tfrac{1}{W+S_k-a_i(X-t)}-\tfrac{1}{W-a_i(X-t)}+\tfrac{S_k}{(W-a_i(X-t))^2}\Big\|_{L^1_{\mathbb{P}}(\Xi)} \\
&\leq \Big\|\int_0^1 \big|-\tfrac{1}{(W+\sigma S_k-a_i(X-t))^2}+\tfrac{1}{(W-a_i(X-t))^2}\big|\,\mathrm{d}\sigma\Big\|_{L^1_{\mathbb{P}}(\Xi)} \\
&\leq \Big\|\int_0^1 \tfrac{|2\sigma W+\sigma^2 S_k-2a_i\sigma(X-t)|}{(W+\sigma S_k-a_i(X-t))^2(W-a_i(X-t))^2}\,\mathrm{d}\sigma\Big\|_{L^1_{\mathbb{P}}(\Xi)}\|S_k\|_{L^\infty_{\mathbb{P}}(\Xi)} \\
&\leq \tfrac{4}{c_\mu^4\mu^4}\big(2\|W\|_{L^\infty_{\mathbb{P}}(\Xi)}+\tfrac{c_\mu\mu}{2}+2a_i\|X-t\|_{L^\infty_{\mathbb{P}}(\Xi)}\big)\|S_k\|_{L^\infty_{\mathbb{P}}(\Xi)}
\end{aligned}$$

with $0 < \|S_k\|_{L^\infty_\mathbb{P}(\Xi)} \leq \frac{c_\mu \mu}{2}$, which shows that the first derivative $\nabla_W \tilde{F}_\mu$ is F-differentiable and thus continuous w. r. t. $W$ at $(X, t, W)$. Again, the remaining second derivatives are shown to be F-derivatives in an analogous way. Continuity of them follows very similarly. □

### First-Order Necessary Optimality Conditions

If the function $U \ni u \mapsto \hat{J}(u, \cdot) \in L^\infty_\mathbb{P}(\Xi)$ is F-differentiable, cf. Assumption 9.9, we can apply the chain rule to compute the first F-derivative of the function $\hat{F}_\mu$ given in (9.15) and obtain

$$\begin{aligned}
\nabla_u \hat{F}_\mu(u,t,W) &= \mu \sum_{i=1}^{2} a_i \, \mathbb{E}\big[\tfrac{\nabla_u \hat{J}(u,\cdot)}{W - a_i(\hat{J}(u,\cdot) - t)}\big], \\
\nabla_t \hat{F}_\mu(u,t,W) &= 1 - \mu \sum_{i=1}^{2} a_i \, \mathbb{E}\big[\tfrac{1}{W - a_i(\hat{J}(u,\cdot) - t)}\big], \\
\nabla_W \hat{F}_\mu(u,t,W) &= \mathbb{1} - \mu \sum_{i=1}^{2} \tfrac{1}{W - a_i(\hat{J}(u,\cdot) - t)} \in L^1_\mathbb{P}(\Xi).
\end{aligned} \tag{9.23}$$

Again, we write $\nabla_W \hat{F}_\mu(u,t,W) \in L^1_\mathbb{P}(\Xi)$ for the representative of $(\hat{F}_\mu)_W(u,t,W) \in L^\infty_\mathbb{P}(\Xi)^*$. Having computed this function as well as $\nabla_u \hat{J}(u, \cdot)$, the rest of the evaluation of $\nabla \hat{F}_\mu$ consists of pointwise multiplications, computing expectations and vector space operations.

The first order necessary conditions for problem (9.18) are given by

$$\nabla_u \hat{F}_\mu(\bar{u}, \bar{t}, \bar{W}) = 0, \quad \nabla_t \hat{F}_\mu(\bar{u}, \bar{t}, \bar{W}) = 0, \quad \nabla_W \hat{F}_\mu(\bar{u}, \bar{t}, \bar{W}) = 0. \tag{9.24}$$

By Lemma 9.25 they are even sufficient if the function $u \mapsto \hat{J}(u, \xi)$ is convex for a. e. $\xi$.

### Barrier-Newton System

To solve (9.24), we apply Newton's method. This is a suitable approach if the function $U \ni u \mapsto \hat{J}(u, \cdot) \in L^\infty_\mathbb{P}(\Xi)$ is twice continuously differentiable, cf. Assumption 9.9, because this property carries over to the function $\hat{F}_\mu$ by Proposition 9.34. Let $(u, t, W) \in U \times \mathbb{R} \times L^\infty_\mathbb{P}(\Xi)$ be the current iterates fulfilling

$$W - a_i(\hat{J}(u, \cdot) - t) \geq c_\mu \tilde{\mu} \quad \text{a. s. for } i \in \{1, 2\} \tag{9.25}$$

with $c_\mu \in (0, 1)$ as in (9.22) and $\tilde{\mu} = \mu$. We say that a triple $(u, t, W)$ satisfying (9.25) is *approximately feasible* for (9.18) w. r. t. $c_\mu$ and $\tilde{\mu} \in (0, \infty)$. Then, the Newton directions $(s, \tau, S) \in U \times \mathbb{R} \times L^\infty_\mathbb{P}(\Xi)$ are given as a solution of the following barrier-Newton system:

$$\begin{pmatrix} \nabla^2_{uu} \hat{F}_\mu & \nabla^2_{tu} \hat{F}_\mu & \nabla^2_{Wu} \hat{F}_\mu \\ \nabla^2_{ut} \hat{F}_\mu & \nabla^2_{tt} \hat{F}_\mu & \nabla^2_{Wt} \hat{F}_\mu \\ \nabla^2_{uW} \hat{F}_\mu & \nabla^2_{tW} \hat{F}_\mu & \nabla^2_{WW} \hat{F}_\mu \end{pmatrix} \begin{pmatrix} s \\ \tau \\ S \end{pmatrix} = \begin{pmatrix} -\nabla_u \hat{F}_\mu \\ -\nabla_t \hat{F}_\mu \\ -\nabla_W \hat{F}_\mu \end{pmatrix}, \tag{9.26}$$

where we skip the arguments $(u, t, W)$ for the sake of brevity.

We compute the derivatives by the chain rule:

$$\nabla^2_{uu}\hat{F}_\mu(u,t,W)s = \mu \sum_{i=1}^{2} \Big( a_i \,\mathbb{E}\big[\tfrac{\nabla^2_{uu}\hat{J}(u,\cdot)s}{(W-a_i(\hat{J}(u,\cdot)-t))}\big] + a_i^2\,\mathbb{E}\big[\tfrac{(\nabla_u\hat{J}(u,\cdot),s)_U\nabla_u\hat{J}(u,\cdot)}{(W-a_i(\hat{J}(u,\cdot)-t))^2}\big]\Big)$$

$$\nabla^2_{tu}\hat{F}_\mu(u,t,W)\tau = -\mu \sum_{i=1}^{2} a_i^2\,\mathbb{E}\big[\tfrac{\nabla_u\hat{J}(u,\cdot)}{(W-a_i(\hat{J}(u,\cdot)-t))^2}\big]\tau$$

$$\nabla^2_{Wu}\hat{F}_\mu(u,t,W)S = -\mu \sum_{i=1}^{2} a_i\,\mathbb{E}\big[\tfrac{\nabla_u\hat{J}(u,\cdot)S}{(W-a_i(\hat{J}(u,\cdot)-t))^2}\big]$$

$$\nabla^2_{ut}\hat{F}_\mu(u,t,W)s = -\mu \sum_{i=1}^{2} a_i^2\,\mathbb{E}\big[\tfrac{(\nabla_u\hat{J}(u,\cdot),s)_U}{(W-a_i(\hat{J}(u,\cdot)-t))^2}\big]$$

$$\nabla^2_{tt}\hat{F}_\mu(u,t,W)\tau = \mu \sum_{i=1}^{2} a_i^2\,\mathbb{E}\big[\tfrac{1}{(W-a_i(\hat{J}(u,\cdot)-t))^2}\big]\tau$$

$$\nabla^2_{Wt}\hat{F}_\mu(u,t,W)S = \mu \sum_{i=1}^{2} a_i\,\mathbb{E}\big[\tfrac{S}{(W-a_i(\hat{J}(u,\cdot)-t))^2}\big]$$

$$\nabla^2_{uW}\hat{F}_\mu(u,t,W)s = -\mu \sum_{i=1}^{2} a_i\,\tfrac{(\nabla_u\hat{J}(u,\cdot),s)_U}{(W-a_i(\hat{J}(u,\cdot)-t))^2}$$

$$\nabla^2_{tW}\hat{F}_\mu(u,t,W)\tau = \mu \sum_{i=1}^{2} a_i\,\tfrac{1}{(W-a_i(\hat{J}(u,\cdot)-t))^2}\tau$$

$$\nabla^2_{WW}\hat{F}_\mu(u,t,W)S = \mu \sum_{i=1}^{2} \tfrac{1}{(W-a_i(\hat{J}(u,\cdot)-t))^2}S.$$

Note that $\frac{1}{(W-a_i(\hat{J}(u,\cdot)-t))^2} \in L^\infty_{\mathbb{P}}(\Xi)$.

We solve (9.26) inexactly by applying a forward-backward block Gauss-Seidel iteration. This has the advantage that we can apply a standard low-rank tensor solver to compute the direction $S$ approximately. We select $(s^0,\tau_0,S_0) = (0,0,0)$ as intial guess for the solution of the barrier-Newton system. Let $(s^\ell,\tau_\ell,S_\ell)$ with $\ell \in \mathbb{N}_0$ be the current iterate. The next iterate $(s^{\ell+1},\tau_{\ell+1},S_{\ell+1})$ is computed as follows:

In the forward solve, a solution update $(\tilde{s}^{\ell+1},\tilde{\tau}_{\ell+1},\tilde{S}_{\ell+1})$ for (9.26) is computed based on the current residual and by replacing the exact Hessian by its lower block triangle. This means that we solve

$$\nabla^2_{uu}\hat{F}_\mu\,(\tilde{s}^{\ell+1}-s^\ell) = -\nabla_u\hat{F}_\mu - \nabla^2_{uu}\hat{F}_\mu\,s^\ell - \nabla^2_{tu}\hat{F}_\mu\,\tau_\ell - \nabla^2_{Wu}\hat{F}_\mu\,S_\ell. \tag{9.27}$$

to compute $\tilde{s}^{\ell+1}$. Again, we neglect the argument $(u,t,W)$ here and in the following. Equation (9.27) is rewritten as

$$\nabla^2_{uu}\hat{F}_\mu\,\tilde{s}^{\ell+1} = -\nabla_u\hat{F}_\mu - \nabla^2_{tu}\hat{F}_\mu\,\tau_\ell - \nabla^2_{Wu}\hat{F}_\mu\,S_\ell \tag{9.28}$$

and an iterative solver for it, e. g., a PCG method, is initialized with $s^\ell$. Then we compute $\tilde{\tau}_{\ell+1}$ by solving

$$\nabla^2_{tt}\hat{F}_\mu\,(\tilde{\tau}_{\ell+1}-\tau_\ell) = -\nabla_t\hat{F}_\mu - \nabla^2_{ut}\hat{F}_\mu\,s^\ell - \nabla^2_{tt}\hat{F}_\mu\,\tau_\ell - \nabla^2_{Wt}\hat{F}_\mu\,S_\ell - \nabla^2_{ut}\hat{F}_\mu\,(\tilde{s}^{\ell+1}-s^\ell),$$

which is

$$\nabla^2_{tt}\hat{F}_\mu\,\tilde{\tau}_{\ell+1} = -\nabla_t\hat{F}_\mu - \nabla^2_{ut}\hat{F}_\mu\,\tilde{s}^{\ell+1} - \nabla^2_{Wt}\hat{F}_\mu\,S_\ell. \tag{9.29}$$

Since $\tilde{\tau}_{\ell+1}$ is a scalar, this equation can be solved directly. Analogously, $\tilde{S}_{\ell+1}$ is computed by solving

$$\nabla^2_{WW}\hat{F}_\mu\,\tilde{S}_{\ell+1} = -\nabla_W\hat{F}_\mu - \nabla^2_{uW}\hat{F}_\mu\,\tilde{s}^{\ell+1} - \nabla^2_{tW}\hat{F}_\mu\,\tilde{\tau}_{\ell+1}. \tag{9.30}$$

An iterative tensor solver for this equation is initialized with $S_\ell$.

In the backward solve, the solution update and next iterate $(s^{\ell+1}, \tau_{\ell+1}, S_{\ell+1})$ is computed by solving the system with the upper block triangle and the new residual on the right-hand side. In fact, $S_{\ell+1}$ would be computed via

$$\nabla^2_{WW}\hat{F}_\mu\,(S_{\ell+1}-\tilde{S}_{\ell+1}) = -\nabla_W\hat{F}_\mu - \nabla^2_{uW}\hat{F}_\mu\,\tilde{s}^{\ell+1} - \nabla^2_{tW}\hat{F}_\mu\,\tilde{\tau}_{\ell+1} - \nabla^2_{WW}\hat{F}_\mu\,\tilde{S}_{\ell+1}.$$

This system has the solution $S_{\ell+1} = \tilde{S}_{\ell+1}$ so that it does not have to be solved numerically. Next, $\tau_{\ell+1}$ is computed by

$$\nabla^2_{tt}\hat{F}_\mu\,\tau_{\ell+1} = -\nabla_t\hat{F}_\mu - \nabla^2_{ut}\hat{F}_\mu\,\tilde{s}^{\ell+1} - \nabla^2_{Wt}\hat{F}_\mu\,S_{\ell+1}. \tag{9.31}$$

and $s^{\ell+1}$ is given as the solution of

$$\nabla^2_{uu}\hat{F}_\mu\,s^{\ell+1} = -\nabla_u\hat{F}_\mu - \nabla^2_{tu}\hat{F}_\mu\,\tau_{\ell+1} - \nabla^2_{Wu}\hat{F}_\mu\,S_{\ell+1}. \tag{9.32}$$

The iterative solver for this equation is initialized with $\tilde{s}^{\ell+1}$. In the next iteration, we have $\tilde{s}^{\ell+2} = s^{\ell+1}$ so that (9.28) does not have to be solved numerically.

### 9.3.4. Implementation and Numerical Results

Using the above considerations and ideas from our paper [46], we implement a Newton-type method for the log-barrier problem (9.14) or equivalently (9.18) with the reduced objective function taken from the example from Chapter 3, where we choose $\varphi \equiv 0$ to have convexity. Then the Hessian operator in (9.26) is at least positive semidefinite since we are solving a convex problem due to Lemma 9.25. We work with a fixed discretization using low-rank tensors as described in Chapter 6, i. e., we use linear FE functions to discretize $H^1_0(\Omega)$ and $L^2(\Omega)$ and represent each random variable by its values at the Gaussian quadrature nodes or—equivalently—a multivariate polynomial. In particular, the FE function $\mathrm{u} \in \mathrm{U} \subset U$ is represented by a vector $\mathsf{u} \in \mathbb{R}^{d_u}$ and the polynomial $\mathrm{W} \in \mathcal{P}_d(\Xi)$ is represented by a tensor $\mathbf{W} \in \mathbb{R}^{d_1\times\cdots\times d_m}$ of weighted function values. Since the weights are positive, pointwise constraints such as (9.25) translate to componentwise constraints on the tensor $\mathbf{W}$. The setup is the same as in Chapter 8 with the following exceptions: The domain is chosen to be the square $\Omega = (-1,1)^2$. It is divided into $m = 4$ strips $\Omega_i$ as shown in Figure 9.2, where also the used FE mesh consisting of 16641 nodes is drawn. We use the coefficient

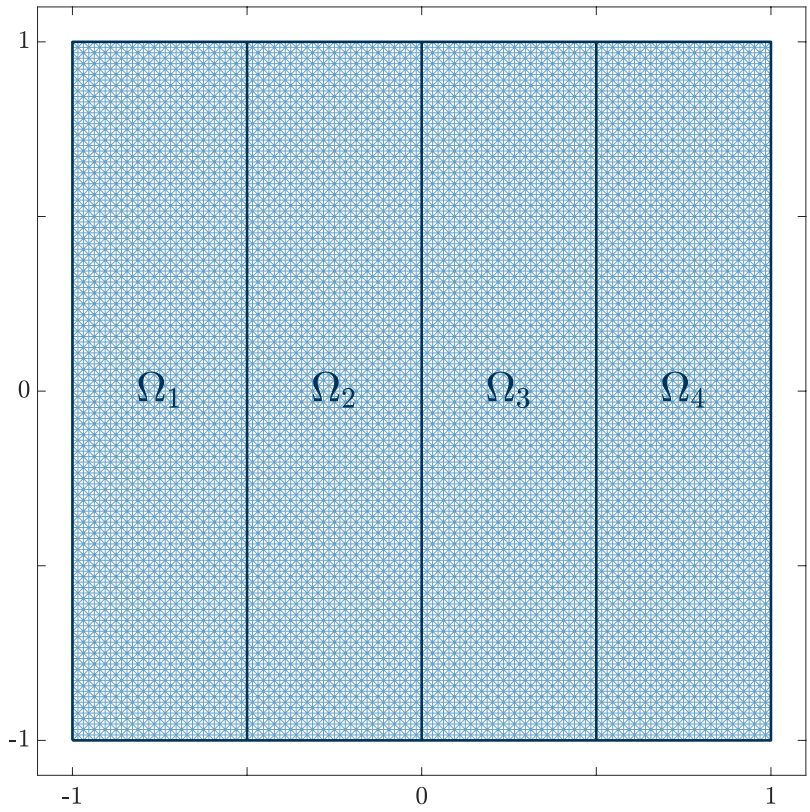



Figure 9.2.: The physical domain $\Omega$, its partition into subdomains $\Omega_i$, and the used finite element mesh

deviations $\sigma = (0.3, 0.4, 0.5, 0.6)^\top$ and discretize each parameter with polynomials of degree $d_i - 1 = 4i + 12$. A high resolution of the stochastic space is chosen because we want to be able to compute quantiles with high accuracy and generate meaningful plots of distribution functions. As mentioned, the "nonlinearity" is set to $\varphi \equiv 0$. The desired state is $\hat{q} \equiv 1$ and we have the set of admissible controls $U_{\text{ad}} = U = L^2(\Omega)$.

Algorithm 2 describes the implemented procedure. It is a Newton-type log-barrier method with stopping criteria stemming from practical experience with low-rank tensor implementations. As observed in [46], the approximate componentwise reciprocals computed by an iterative low-rank tensor method can become too inexact, especially during the last iterations of the log-barrier method. Then, the error in, e. g., the computed gradient of $\hat{F}_\mu$ is too large so that a stopping criterion based on its norm is not reliable anymore. Therefore, we also stop the algorithm if the error in the computed reciprocals is too large or if the stepsizes for retaining approximate feasibility become too small, which is a sign for not accurately enough computed Newton steps. Instead of working with a fixed value $\mu_k = \mu$, we use continuation and start with $\mu_0 \geq \mu$ and decrease this value in each iteration by a factor $\mu_{\text{fac}} \in (0, 1)$ until $\mu_k = \mu$ is reached. The algorithm is formulated such that the iterates $(\mathrm{u}^k, t_k, \mathrm{W}_k)$ are approximately feasible for (9.18) w. r. t. $c_\mu$ and $\mu_k$. We give some more details about the substeps in the following.

**Algorithm 2:** Log-Barrier Method for Solving Problem (9.14)

**Parameters** : $\mu_0 \geq \mu$, $\mu_{\text{fac}} \in (0,1)$, $c_\mu \in (0,1)$, $\varepsilon_{\text{grad}} > 0$

**Input**: Initial iterates $\mathrm{u}^0 \in \mathrm{U}$, $t_0 \in \mathbb{R}$, $\mathrm{W}_0 \in \mathcal{P}_d(\Xi)$ s. t. $(\mathrm{u}^0, t_0, \mathrm{W}_0)$ is approximately feasible for (9.18) w. r. t. $c_\mu$ and $\mu_0$.

**Output**: $\bar{\mathrm{u}} \in \mathrm{U}$, $\bar{t} \in \mathbb{R}$, $\bar{\mathrm{W}} \in \mathcal{P}_d(\Xi)$

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

Compute the tensors corresponding to $\hat{J}(\mathrm{u}^k, \cdot)$ and $\nabla_u \hat{J}(\mathrm{u}^k, \cdot)$.

Compute the tensors representing $Y_{k,i} := \mathrm{W}_k - a_i(\hat{J}(\mathrm{u}^k, \cdot) - t_k)$ and $Z_{k,i} = Y_{k,i}^{-1}$, $i \in \{1, 2\}$.

**if** *the tensors for $Z_{k,i}$ cannot be computed exactly enough,* **then**

STOP and return $(\bar{\mathrm{u}}, \bar{t}, \bar{\mathrm{W}}) = (\mathrm{u}^k, t_k, \mathrm{W}_k)$.

**end**

Compute representations of the partial gradients $\nabla_u \hat{F}^k_{\mu_k} = \nabla_u \hat{F}_{\mu_k}(\mathrm{u}^k, t_k, \mathrm{W}_k)$, $\nabla_t \hat{F}^k_{\mu_k} = \nabla_t \hat{F}_{\mu_k}(\mathrm{u}^k, t_k, \mathrm{W}_k)$, and $\nabla_W \hat{F}^k_{\mu_k} = \nabla_W \hat{F}_{\mu_k}(\mathrm{u}^k, t_k, \mathrm{W}_k)$, see (9.23).

**if** $\mu_k = \mu$ *and* $\|\nabla_u \hat{F}^k_{\mu_k}\|_U^2 + |\nabla_t \hat{F}^k_{\mu_k}|^2 + \|\nabla_W \hat{F}^k_{\mu_k}\|^2_{L^2_{\mathbb{P}}(\Xi)} < \varepsilon^2_{\text{grad}}$, **then**

STOP and return $(\bar{\mathrm{u}}, \bar{t}, \bar{\mathrm{W}}) = (\mathrm{u}^k, t_k, \mathrm{W}_k)$.

**end**

Compute the tensors corresponding to $Y^2_{k,i}$ and to $Z^2_{k,i}$.

To compute the update direction $(\mathrm{s}^k, \tau_k, \mathrm{S}_k) \in \mathrm{U} \times \mathbb{R} \times \mathcal{P}_d(\Xi)$, solve the barrier-Newton system (9.26) with $(u, t, W) = (\mathrm{u}^k, t_k, \mathrm{W}_k)$ and $\mu = \mu_k$ approximately by applying a forward-backward block Gauss-Seidel iteration.

Set $\mu_{k+1} := \max\{\mu, \mu_{\text{fac}} \cdot \mu_k\}$.

Compute stepsizes $\sigma^k_u, \sigma^k_t, \sigma^k_W \in (0, 1]$ such that $(\mathrm{u}^k + \sigma^k_u \mathrm{s}^k, t_k + \sigma^k_t \tau_k, \mathrm{W}_k + \sigma^k_W \mathrm{S}_k)$ is approximately feasible for (9.18) w. r. t. $c_\mu$ and $\mu_{k+1}$.

**if** *such stepsizes cannot be computed or are too small,* **then**

STOP and return $(\bar{\mathrm{u}}, \bar{t}, \bar{\mathrm{W}}) = (\mathrm{u}^k, t_k, \mathrm{W}_k)$.

**end**

Set $\mathrm{u}^{k+1} := \mathrm{u}^k + \sigma^k_u \mathrm{s}^k$, $t_{k+1} := t_k + \sigma^k_t \tau_k$, and $\mathrm{W}_{k+1} := \mathrm{W}_k + \sigma^k_W \mathrm{S}_k$.

**end**

The object $\mathrm{u}^k \in \mathrm{U}$ is a finite element function represented by a vector $\mathsf{u}^k \in \mathbb{R}^{d_u}$. When computing the tensors corresponding to $\hat{J}(\mathrm{u}^k, \cdot)$ and $\nabla_u \hat{J}(\mathrm{u}^k, \cdot)$, we solve the discretized state and adjoint equation by AMEn and store the low-rank tensors $\mathbf{y}_k$, $\mathbf{z}_k \in \mathbb{R}^{d_0 \times \cdots \times d_m}$ representing the discrete state and adjoint state, respectively, to initialize AMEn in the next iteration. The $\mathbb{R}^{d_1 \times \cdots \times d_m}$-tensor representing $\hat{J}(\mathrm{u}^k, \cdot)$ is

$$\left(\mathbb{1}^\top \circ_1 \left[(\mathsf{M}_H \circ_1 (\mathbf{Q}\mathbf{y}_k - \hat{\mathbf{q}})) \odot (\mathbb{1} \otimes \boldsymbol{\omega}) \odot (\mathbf{Q}\mathbf{y}_k - \hat{\mathbf{q}})\right]\right)(1, \bullet, \ldots, \bullet) + \left(\tfrac{\gamma}{2} (\mathsf{u}^k)^\top \tilde{\mathsf{M}} \mathsf{u}^k\right) \boldsymbol{\omega}.^{-1}, \quad (9.33)$$

with $\mathbb{1} \in \mathbb{R}^{d_H}$, cf. (6.18), and the one representing $\nabla_u \hat{J}(\mathrm{u}^k, \cdot)$ is given by

$$-(\tilde{\mathsf{M}}^{-1} \mathsf{B}^\top) \circ_1 \mathbf{z}_k + \gamma \mathsf{u}^k \otimes \boldsymbol{\omega}.^{-1} \in \mathbb{R}^{d_u \times d_1 \times \cdots \times d_m},$$

cf. (6.22). In (9.33) and in the following we use interpolation, i. e., we construct a tensor of function evaluations in the Gaussian quadrature nodes $\mathsf{a}_l$ multiplied by the weights $\boldsymbol{\omega}(l)^{-1}$.

The tensors $\mathbf{Y}_{k,i} \in \mathbb{R}^{d_1 \times \cdots \times d_m}$ and $\mathbf{Z}_{k,i} \in \mathbb{R}^{d_1 \times \cdots \times d_m}$ shall represent the functions $Y_{k,i}$ and $Z_{k,i}$, cf. (6.10). To compute $\mathbf{Z}_{k,i} \approx \boldsymbol{\omega}.^{-1} \odot (\boldsymbol{\omega} \odot \mathbf{Y}_{k,i}).^{-1}$, we first apply the Newton-Schulz method, see Subsection 2.1.3. If the relative error $\frac{\|(\boldsymbol{\omega} \odot \mathbf{Z}_{k,i}) \odot (\boldsymbol{\omega} \odot \mathbf{Y}_{k,i}) - \mathbb{1}\|_{\mathsf{F}}}{\|\mathbb{1}\|_{\mathsf{F}}}$ is greater than $10^{-4}$, we perform additional AMEn iterations to compute the elementwise reciprocal, cf. [46]. Algorithm 2 is stopped if the relative error in the elementwise reciprocal exceeds 0.1. Note that in theory we have $Y_{k,i}(\xi) \geq c_\mu \mu_k$ for a. e. $\xi \in \Xi$ because $(\mathrm{u}^k, t_k, \mathrm{W}_k)$ is approximately feasible for (9.18) w. r. t. $c_\mu$ and $\mu_k$. Therefore, the functions $Z_{k,i} \in L^\infty_{\mathbb{P}}(\Xi)$ are well-defined and fulfill $Z_{k,i}(\cdot) \in (0, \frac{1}{c_\mu \mu_k}]$ almost surely.

The evaluation of the gradient (9.23) becomes

$$\begin{aligned}
\nabla_u \hat{F}_{\mu_k}(\mathrm{u}^k, t_k, \mathrm{W}_k) &= \mu_k \sum_{i=1}^{2} a_i \, \mathbb{E}\big[\nabla_u \hat{J}(\mathrm{u}^k, \cdot)\, Z_{k,i}\big] \\
\nabla_t \hat{F}_{\mu_k}(\mathrm{u}^k, t_k, \mathrm{W}_k) &= 1 - \mu_k \sum_{i=1}^{2} a_i \, \mathbb{E}\big[Z_{k,i}\big] \\
\nabla_W \hat{F}_{\mu_k}(\mathrm{u}^k, t_k, \mathrm{W}_k) &= \mathbb{1} - \mu_k \sum_{i=1}^{2} Z_{k,i}.
\end{aligned}$$

The expectations can be evaluated using tensor contractions. To evaluate the gradient norm, we use the discrete $L^2_{\mathbb{P}}(\Xi)$-norm because it is also given as a tensor contraction.

The tensors corresponding to $Y^2_{k,i}$ are $\mathbf{Y}_{k,i} \odot (\boldsymbol{\omega} \odot \mathbf{Y}_{k,i})$ and are computed by truncated componentwise multiplication and $i$-mode matrix products, cf. (6.25). Analogously, we compute the tensors corresponding to $Z^2_{k,i}$.

To solve the barrier-Newton system (9.26) approximately, we perform at most 5 iterations of the described forward-backward block Gauss-Seidel method. We stop this iteration earlier if the relative residual is smaller than 0.1. In (9.28), we use an approximation $\tilde{\nabla}^2_{uu} \hat{F}_{\mu_k}(\mathrm{u}^k, t_k, \mathrm{W}_k)$ of the Hessian involving a constant (w. r. t. $\xi$) approximation $\tilde{\nabla}^2_{uu} \hat{J}(\mathrm{u}^k)$ of the Hessian $\nabla^2_{uu} \hat{J}(\mathrm{u}^k, \cdot)$, cf. (6.26). This leads to

$$\begin{aligned}
&\mu_k \sum_{i=1}^{2} \Big( a_i \, \mathbb{E}\big[Z_{k,i}\big] \tilde{\nabla}^2_{uu} \hat{J}(\mathrm{u}^k) \tilde{\mathrm{s}}^{k,\ell+1} + a_i^2 \, \mathbb{E}\big[(\nabla_u \hat{J}(\mathrm{u}^k, \cdot), \tilde{\mathrm{s}}^{k,\ell+1})_U \nabla_u \hat{J}(\mathrm{u}^k, \cdot)\, Z^2_{k,i}\big] \Big) = \\
&- \mu_k \sum_{i=1}^{2} a_i \, \mathbb{E}\big[\nabla_u \hat{J}(\mathrm{u}^k, \cdot)\, Z_{k,i}\big] + \mu_k \tau_{k,\ell} \sum_{i=1}^{2} a_i^2 \mathbb{E}[\nabla_u \hat{J}(\mathrm{u}^k, \cdot)\, Z^2_{k,i}] \\
&\quad + \mu \sum_{i=1}^{2} a_i^2 \mathbb{E}[\nabla_u \hat{J}(\mathrm{u}^k, \cdot)\, \mathrm{S}_{k,\ell}\, Z^2_{k,i}].
\end{aligned}$$

A PCG method is used to solve this equation because the discrete reference Hessian $\tilde{\nabla}^2_{\mathsf{uu}} \hat{\mathsf{J}}(\mathsf{u}^k)$ can only be applied efficiently, but should not be formed as an explicit, typically dense matrix. The expectation $\mathbb{E}\big[Z_{k,i}\big]$ and the right-hand side are precomputed before solving w. r. t. $\tilde{\mathrm{s}}^{k,\ell+1}$. Expected values are computed by tensor contractions and componentwise

multiplication. (9.32) is solved analogously. Equation (9.29) becomes

$$\begin{aligned}\tilde{\tau}_{k,\ell+1} &= \frac{-\nabla_t \hat{F}_{\mu_k}(\mathrm{u}^k, t_k, \mathrm{W}_k) - \nabla^2_{ut}\hat{F}_{\mu_k}(\mathrm{u}^k, t_k, \mathrm{W}_k)\tilde{\mathrm{s}}^{k,\ell+1} - \nabla^2_{Wt}\hat{F}_{\mu_k}(\mathrm{u}^k, t_k, \mathrm{W}_k)\mathrm{S}_{k,\ell}}{\nabla^2_{tt}\hat{F}_{\mu_k}(\mathrm{u}^k, t_k, \mathrm{W}_k)} \\ &= \Big(\mu_k \sum_{i=1}^{2} a_i^2\, \mathbb{E}\big[Z_{k,i}^2\big]\Big)^{-1}\Big(-1 + \mu_k \sum_{i=1}^{2} a_i\, \mathbb{E}\big[Z_{k,i}\big] \\ &\qquad + \mu_k \sum_{i=1}^{2} a_i^2\, \mathbb{E}[(\nabla_u \hat{J}(\mathrm{u}^k, \cdot), \tilde{\mathrm{s}}^{k,\ell+1})_U\, Z_{k,i}^2] - \mu_k \sum_{i=1}^{2} a_i \mathbb{E}[Z_{k,i}^2\, \mathrm{S}_{k,\ell}]\Big)\end{aligned}$$

and (9.31) is formulated analogously. We transform (9.30) to

$$\begin{aligned}\tilde{\mathrm{S}}_{k,\ell+1} &:= \nabla^2_{WW}\hat{F}_{\mu_k}(\mathrm{u}^k, t_k, \mathrm{W}_k)^{-1}\Big(-\nabla_W \hat{F}_{\mu_k}(\mathrm{u}^k, t_k, \mathrm{W}_k) \\ &\qquad - \nabla^2_{uW}\hat{F}_{\mu_k}(\mathrm{u}^k, t_k, \mathrm{W}_k)\tilde{\mathrm{s}}^{k,\ell+1} - \nabla^2_{tW}\hat{F}_{\mu_k}(\mathrm{u}^k, t_k, \mathrm{W}_k)\tilde{\tau}_{k,\ell+1}\Big) \\ &= \tfrac{1}{\mu_k}(Z_{k,1}^2 + Z_{k,2}^2)^{-1}\Big(-\mathbb{1} + \mu_k(Z_{k,1} + Z_{k,2}) \\ &\qquad + \mu_k(a_1 Z_{k,1}^2 + a_2 Z_{k,2}^2)(\nabla_u \hat{J}(\mathrm{u}^k, \cdot), \tilde{\mathrm{s}}^{k,\ell+1})_U - \mu_k \tilde{\tau}_{k,\ell+1}(a_1 Z_{k,1}^2 + a_2 Z_{k,2}^2)\Big) \\ &= \tfrac{1}{\mu_k}(Y_{k,2}^2 + Y_{k,1}^2)^{-1}\Big(-Y_{k,1}^2 Y_{k,2}^2 + \mu_k(Y_{k,1}Y_{k,2}^2 + Y_{k,1}^2 Y_{k,2}) \\ &\qquad + \mu_k(a_1 Y_{k,2}^2 + a_2 Y_{k,1}^2)(\nabla_u \hat{J}(\mathrm{u}^k, \cdot), \tilde{\mathrm{s}}^{k,\ell+1})_U - \mu \tilde{\tau}_{k,\ell+1}(a_1 Y_{k,2}^2 + a_2 Y_{k,1}^2)\Big).\end{aligned}$$

The advantage of the latter formulation is that is does not depend on the possibly inexactly computed $Z_{k,i}$. The tensor representing $\tilde{\mathrm{S}}_{k,\ell+1}$ is computed by AMEn with componentwise multiplication operators.

We propose a stepsize selection strategy, which uses the fact that (9.25) is a box constraint w. r. t. $W$ and $t$ and that it can be rewritten as

$$\operatorname*{ess\,inf}_{\xi\in\Xi} W(\xi) - a_i(\hat{J}(u,\xi) - t) \geq c_\mu \tilde{\mu} \quad \text{for } i \in \{1,2\}.$$

It tries to avoid too many evaluations of the random variable objective function $\hat{J}(u,\cdot)$ for different $u$ and works as follows:

S1. Choose factors $\sigma_{\mathrm{fac},u} \in (0,1)$, $\hat{\sigma}_{\mathrm{fac}} \in (0,1]$, and initial stepsizes $\sigma_u$, $\hat{\sigma} \in (0,1]$.

S2. Compute the tensor corresponding to $\hat{J}(\mathrm{u}^k + \sigma_u \mathrm{s}^k, \cdot) \in L^\infty_{\mathbb{P}}(\Xi)$ and store this quantity and the corresponding state for the next iteration.

S3. Compute $\delta_{k,i} := \min_{\xi\in\Xi} \mathrm{W}_k(\xi) + \hat{\sigma}\,\mathrm{S}_k(\xi) - a_i \hat{J}(\mathrm{u}^k + \sigma_u \mathrm{s}^k, \xi)$ for $i \in \{1,2\}$. This is done by the multilevel coordinate search (MCS) method [65], a non-rigorous global optimization routine, which works with evaluations of the respective polynomial only, see Subsection 2.1.3.

S4. If $\delta_{k,i} \geq c_\mu \mu_{k+1} - \hat{\sigma} a_i \tau_k - a_i t_k$ for $i \in \{1,2\}$, choose $\sigma_u^k = \sigma_u$ and $\sigma_t^k = \sigma_W^k = \hat{\sigma}$ and STOP. The new iterate $(\mathrm{u}^k + \sigma_u^k \mathrm{s}^k, t_k + \sigma_t^k \tau_k, \mathrm{W}_k + \sigma_W^k \mathrm{S}_k)$ is then approximately feasible for (9.18) w. r. t. $c_\mu$ and $\mu_{k+1}$.
Otherwise, compute $\nu_{k,i} := \min_{\xi\in\Xi} \mathrm{W}_k(\xi) - a_i \hat{J}(\mathrm{u}^k + \sigma_u \mathrm{s}^k, \xi)$ for $i \in \{1,2\}$ by MCS.

S5. If $\nu_{k,i} < c_\mu \mu_{k+1} - a_i t_k$ for some $i \in \{1,2\}$, decrease $\sigma_u$ by the factor $\sigma_{\mathrm{fac},u}$ and go to S2 because $(\mathrm{u}^k + \sigma_u \mathrm{s}^k, t_k, \mathrm{W}_k)$ is not approximately feasible.
Otherwise, i. e., $\nu_{k,i} \geq c_\mu \mu_{k+1} - a_i t_k$ for all $i \in \{1,2\}$ and $c_\mu \mu_{k+1} - a_i t_k > \delta_{k,i} + \hat{\sigma} a_i \tau_k$ for at least one $i \in \{1,2\}$ giving $\nu_{k,i} > \delta_{k,i} + \hat{\sigma} a_i \tau_k$ for at least one $i \in \{1,2\}$, compute

$$\sigma^k := \min_{i\,:\,\nu_{k,i} > \delta_{k,i} + \hat{\sigma} a_i \tau_k} \frac{c_\mu \mu_{k+1} - \nu_{k,i} - a_i t_k}{-\frac{\nu_{k,i}}{\hat{\sigma}} + \frac{\delta_{k,i}}{\hat{\sigma}} + a_i \tau_k}.$$

This stepsize is non-negative because the denominator is negative and the numerator is non-positive.

S6. If $\sigma^k < \hat{\sigma}_{\mathrm{fac}}\, \sigma_u$, decrease $\sigma_u$ by the factor $\sigma_{\mathrm{fac},u}$ and go to S2 to have stepsizes of approximately the same magnitude.
Otherwise, choose $\sigma_u^k = \sigma_u$ and $\sigma_t^k = \sigma_W^k = \sigma^k$ and STOP. The new iterate $(\mathrm{u}^k + \sigma_u^k \mathrm{s}^k, t_k + \sigma_t^k \tau_k, \mathrm{W}_k + \sigma_W^k \mathrm{S}_k)$ is approximately feasible:

$$\begin{aligned}
&\mathrm{W}_k(\cdot) + \sigma^k \mathrm{S}_k(\cdot) - a_i \hat{J}(\mathrm{u}^k + \sigma_u^k \mathrm{s}^k, \cdot) + a_i t_k + a_i \sigma^k \tau_k \\
&= (1 - \tfrac{\sigma^k}{\hat{\sigma}})\big(\mathrm{W}_k(\cdot) - a_i \hat{J}(\mathrm{u}^k + \sigma_u^k \mathrm{s}^k, \cdot)\big) \\
&\qquad + \tfrac{\sigma^k}{\hat{\sigma}}\big(\mathrm{W}_k(\cdot) + \hat{\sigma} \mathrm{S}_k(\cdot) - a_i \hat{J}(\mathrm{u}^k + \sigma_u \mathrm{s}^k, \cdot)\big) + a_i t_k + \sigma^k a_i \tau_k \\
&\geq (1 - \tfrac{\sigma^k}{\hat{\sigma}}) \nu_{k,i} + \tfrac{\sigma^k}{\hat{\sigma}} \delta_{k,i} + a_i t_k + \sigma^k a_i \tau_k \\
&= \sigma^k (-\tfrac{\nu_{k,i}}{\hat{\sigma}} + \tfrac{\delta_{k,i}}{\hat{\sigma}} + a_i \tau_k) + \nu_{k,i} + a_i t_k.
\end{aligned}$$

If $-\frac{\nu_{k,i}}{\hat{\sigma}} + \frac{\delta_{k,i}}{\hat{\sigma}} + a_i \tau_k < 0$, we have

$$\sigma^k (-\tfrac{\nu_{k,i}}{\hat{\sigma}} + \tfrac{\delta_{k,i}}{\hat{\sigma}} + a_i \tau_k) + \nu_{k,i} + a_i t_k \geq c_\mu \mu_{k+1} - \nu_{k,i} - a_i \tau_k + \nu_{k,i} + a_i t_k = c_\mu \mu_{k+1}$$

by the definition of $\sigma^k$. Otherwise, we estimate

$$\underbrace{\sigma^k (-\tfrac{\nu_{k,i}}{\hat{\sigma}} + \tfrac{\delta_{k,i}}{\hat{\sigma}} + a_i \tau_k)}_{\geq 0} + \nu_{k,i} + a_i t_k \geq \nu_{k,i} + a_i t_k \geq c_\mu \mu_{k+1}.$$

Algorithm 2 is stopped if the stepsize $\sigma_u$ becomes smaller than $10^{-3}$ during the iteration. In the numerical tests, we choose $\mu_0 = 10$, $\mu_{\mathrm{fac}} = 0.8$, $c_\mu = 0.1$, $\varepsilon_{\mathrm{grad}} = 0.01$, $\sigma_u = \hat{\sigma} = 0.7$, $\sigma_{\mathrm{fac},u} = \hat{\sigma}_{\mathrm{fac}} = 0.5$, and bound all appearing tensor ranks by 200. This rank bound is chosen quite large to be able to obtain accurate results for difficult setups. In easier cases such as $\beta = 0.5$, $\lambda = 1.00$, $\mu = 0.10$, see below, the largest used rank for the elementwise reciprocals $Z_{k,i}$ is 57 and the largest rank for the steps $\mathrm{S}_k$ is 83, i. e., the rank bound is not restrictive. The initial control $\mathrm{u}^0$ is chosen to be the deterministic control and we take $t_0 = 0$ and $\mathrm{W}_0(\cdot) := a_2 \hat{J}(\mathrm{u}^0, \cdot) + \mu_0$. Since $\hat{J}(\mathrm{u}^0, \cdot) \geq 0$ and $a_2 > a_1 \geq 0$, this yields that $(\mathrm{u}^0, t_0, \mathrm{W}_0)$ is approximately feasible.

Clearly, Algorithm 2 is not as rigorous as, e. g., Algorithm 1 and the implementation presented in Chapter 8. The stopping criteria based on the quality of the elementwise reciprocal and the computed stepsize, for instance, come from practical experience with low-rank tensor implementations. In theory, developing a convergence theory for such an algorithm based

on results on inexact Newton methods [41] or inexact interior point methods [19] would be possible. But it is questionable if the requirements of such convergence results can be met in practice. Even a very small relative error in the tensor $\mathbf{Y}_{k,i}$ measured in a norm induced by an inner product can cause the reciprocal tensor $\mathbf{Z}_{k,i}$ to be not well-defined anymore because, e. g., one entry of $\mathbf{Y}_{k,i}$ could become zero, cf. the discussion in Subsection 2.1.3. Furthermore, the discretization relies on interpolation in the multivariate Gaussian quadrature nodes. As noted at the end of Section 6.2, this can be interpreted as a discretization based on samples in the quadrature nodes. This is a suitable interpretation in this case because the polynomials of arbitrary degree do not form a dense subset of $L^\infty_{\mathbb{P}}(\Xi)$ in general so that a discretization by polynomials is not appropriate.

In the following, we present results for different combinations of the quantile parameter $\beta \in \{0.5, 0.9\}$ in $\mathrm{CVaR}_\beta$, the convex combination parameter $\lambda \in \{1.00, 0.75, 0.50\}$, see (9.5), and the log-barrier parameter $\mu \in \{0.1, 0.05, 0.02\}$. The results are summed up in Table 9.1.

| $\beta$ | $\lambda$ | $\mu$ | number of iterations | achieved gradient norm | computing time (hours) |
|---|---|---|---|---|---|
| 0.9 | 1.00 | 0.05 | 24 | 0.3012 | 3.83 |
| 0.9 | 1.00 | 0.10 | 25 | 0.0062 | 4.38 |
| 0.9 | 0.75 | 0.05 | 26 | 0.2869 | 1.97 |
| 0.9 | 0.50 | 0.05 | 28 | 0.0067 | 2.69 |
| 0.5 | 1.00 | 0.02 | 34 | 0.0071 | 5.56 |
| 0.5 | 1.00 | 0.05 | 28 | 0.0010 | 1.02 |
| 0.5 | 1.00 | 0.10 | 25 | 0.0083 | 0.62 |
| 0.5 | 0.75 | 0.02 | 33 | 0.0072 | 3.77 |
| 0.5 | 0.50 | 0.02 | 34 | 0.0068 | 2.51 |

Table 9.1.: Results of the implemented log-barrier method.

If the achieved gradient norm is greater than 0.01, the algorithm has stopped because the error in the elementwise reciprocal is too large. Stopping due to too small stepsizes has never been occurred although the full stepsize is not always taken, especially in the last iterations.

In general, it can be observed that it is easier to solve problems which are less risk-averse in the sense that smaller values of $\beta$ or $\lambda$, or larger values of $\mu$ are chosen. In two cases, namely $\beta = 0.9$, $\mu = 0.05$, and $\lambda \in \{0.75, 1.00\}$, it is not possible to obtain a gradient norm smaller than 0.01. Nevertheless, we will see that the obtained solutions yield a smaller CVaR than the ones with with $\lambda = 0.50$ or $\mu = 0.10$, for which a better accuracy is achieved. In contrast to the case $\beta = 0.9$, the algorithm computes a solution with a small gradient for $\beta = 0.5$ even with $\mu = 0.02$. The variability in the computing times is due to various influences: Depending on the obtained residual, possibly more Gauss-Seidel iterations have to be performed, especially for ill-conditioned systems. Furthermore, we use costly direct solves with full linear algebra in AMEn subproblems if the computation of the elementwise reciprocal with an iterative subsolver does not yield satisfactory results. The ranks needed to

compute, e. g., the reciprocal tensor accurately enough can vary depending on the problem data. Subsequent computations with this tensor are then more costly.

An interesting result, where the controls differ relatively much from each other, is obtained for $\beta = 0.5$, $\lambda = 1.00$, and $\mu = 0.02$, i. e., for minimization of the smoothed $\mathrm{CVaR}_{0.5}$ with a rather small value of the barrier parameter $\mu$. Figure 9.3 shows the distribution function of the "random variable objective function" [72] $\hat{J}(u, \cdot)$ for the deterministic, the risk-neutral, and the respective risk-averse control. In all distribution function plots, the expected value $\mathbb{E}[\hat{J}(u, \cdot)]$ of the respective random variable is marked by "$*$". $\mathrm{CVaR}_{0.5}[\hat{J}(u, \cdot)]$ and $\mathrm{CVaR}_{0.9}[\hat{J}(u, \cdot)]$ are marked by "$+$" and "$\times$", respectively. As depicted in Figure 9.3, the risk-neutral control reduces the expectation as well as the $\mathrm{CVaR}_{0.5}$ and the $\mathrm{CVaR}_{0.9}$ of the random variable objective function compared to the values achieved by the deterministic control. Using the risk-averse approach, it is possible to decrease the $\mathrm{CVaR}_{0.5}$ further while the expectation increases only a bit. Even the $\mathrm{CVaR}_{0.9}$ is decreased in this setting although it is not minimized explicitly. The obtained controls and their differences are plotted in Figure 9.4. Especially the difference plots show that the uncertainty in the system is increasing from left to right. Additionally, the strips $\Omega_i$ can be recognized in these plots.

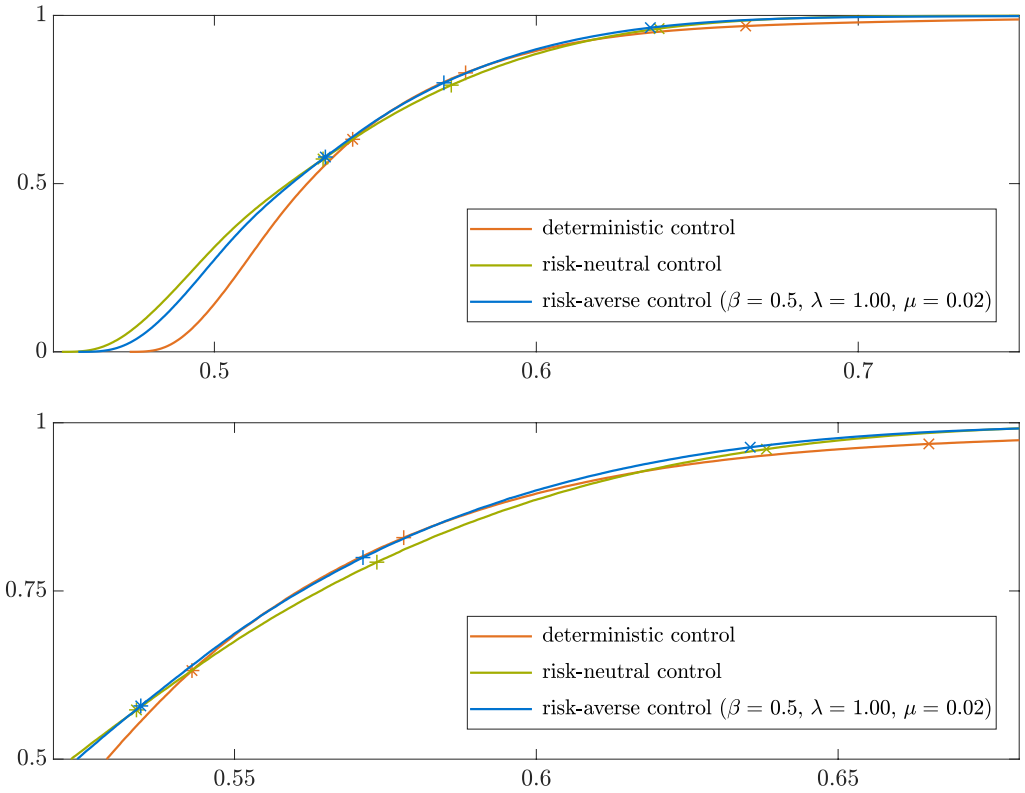



Figure 9.3.: Distribution function of the random variable objective function for the deterministic, the risk-neutral, and the risk-averse control with $\beta = 0.5$, $\lambda = 1.00$, and $\mu = 0.02$, details at the bottom

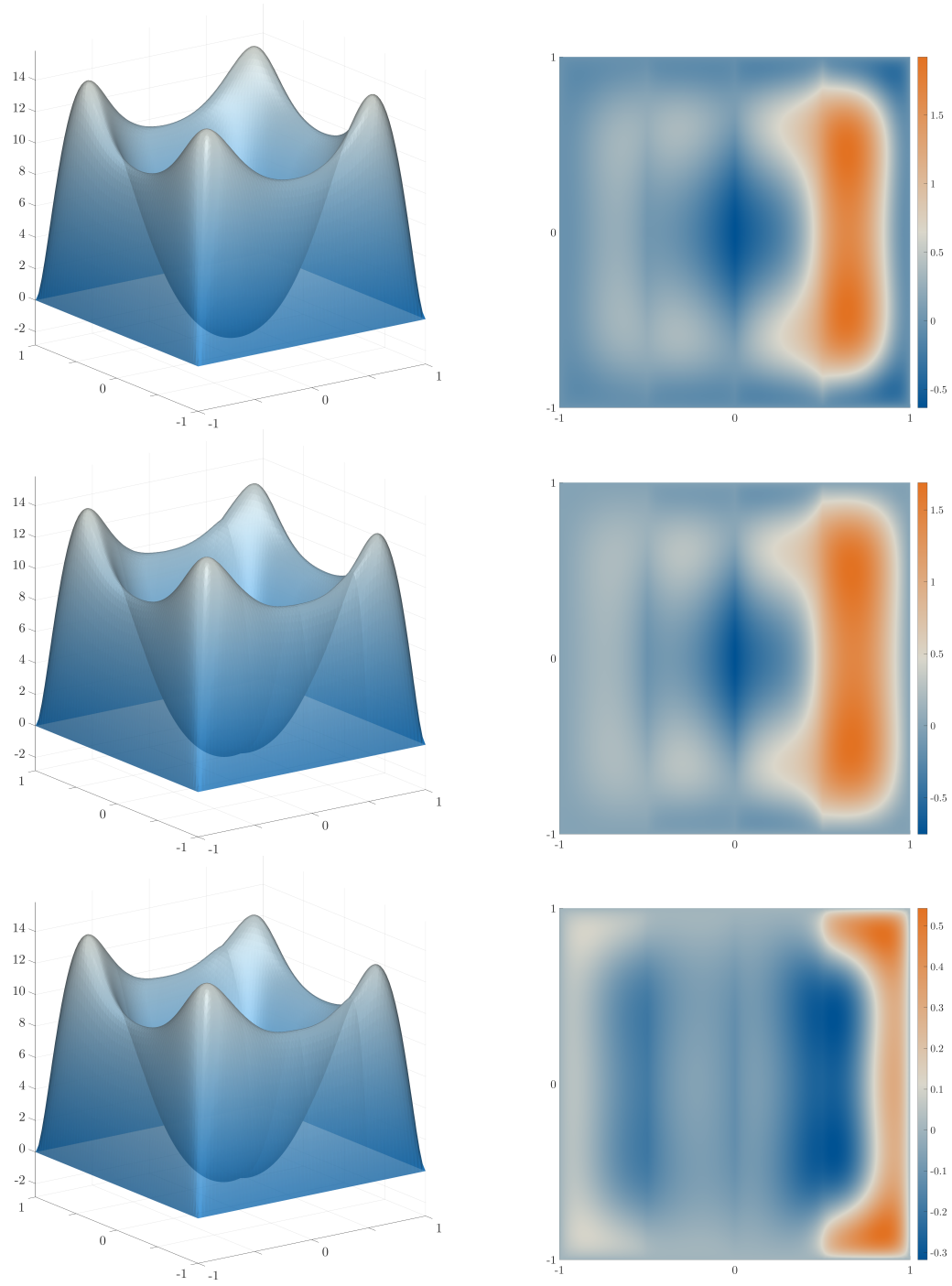


Figure 9.4.: Plots of the deterministic, the risk-neutral, and the risk-averse control (from top left to bottom left) for $\beta = 0.$, $\lambda = 1.00$, and $\mu = 0.02$. Differences between the deterministic and the risk-neutral, the deterministic and the risk-averse, and the risk-neutral and the risk-averse control (from top right to bottom right).

In the following, we investigate how the choice of one of the parameters $\beta$, $\lambda$, $\mu$ influences the result. Sometimes, the differences between the distribution functions are small so that we show regions of interest in the respective plots.

We start with the role of the log-barrier parameter $\mu$. Figure 9.5 shows the distribution function of the random variable objective function for the risk-neutral and the risk-averse controls with $\beta = 0.5$ and $\lambda = 1.00$, i. e., we aim for minimizing smoothed versions of $\mathrm{CVaR}_{0.5}$. The log-barrier parameter is decreased from 0.10 to 0.02. This setup is chosen because the algorithm is capable of computing a solution with small gradient, see Table 9.1, which is not the case for $\beta = 0.9$. We see that the $\mathrm{CVaR}_{0.5}[\hat{J}(u,\cdot)]$ increases and approaches the value achieved by the risk-neutral control for larger values of $\mu$ while the expected value decreases slightly. This becomes clear by taking a look at Figure 9.1. For larger values of $\mu$, the function $\hat{v}_\mu$ gets locally closer to the identity $v(t) = t$, which induces $\mathcal{R}_v = \mathbb{E}$.

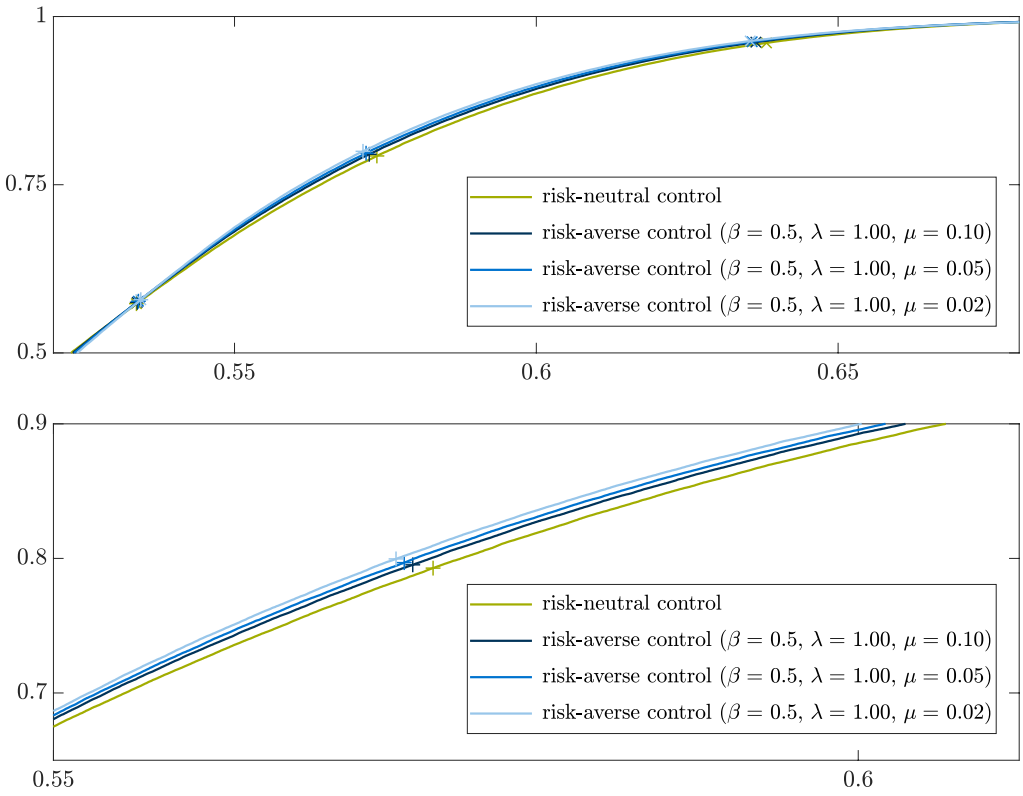



Figure 9.5.: Distribution function of the random variable objective function for the risk-neutral and the risk-averse controls with $\mu \in \{0.10, 0.05, 0.02\}$, $\beta = 0.5$, $\lambda = 1.00$

Figures 9.6 and 9.6 show the distribution functions as the combination parameter $\lambda$ varies. Here we have two different setups, namely $(\beta, \mu) = (0.5, 0.02)$ and $(\beta, \mu) = (0.9, 0.05)$. As expected, the respective $\mathrm{CVaR}_\beta$ is smaller for larger values of $\lambda$ and the expectation increases. It is remarkable that we make this observation also in the case $\beta = 0.9$ although the algorithm has to stop with a gradient norm of about 0.3 for $\lambda = 1.00$ and $\lambda = 0.75$ because the componentwise reciprocal cannot be computed accurately enough, see Table 9.1.

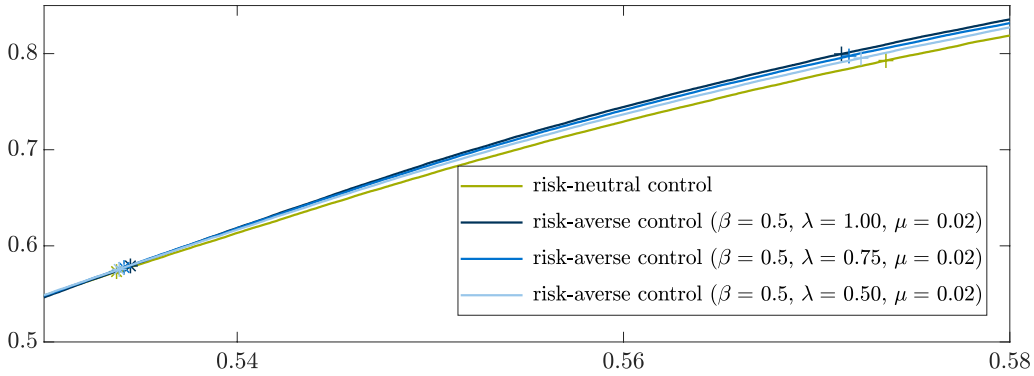



Figure 9.6.: Distribution function of the random variable objective function for the risk-neutral and the risk-averse controls with $\lambda \in \{1.00, 0.75, 0.50\}$ and $(\beta, \mu) = (0.5, 0.02)$

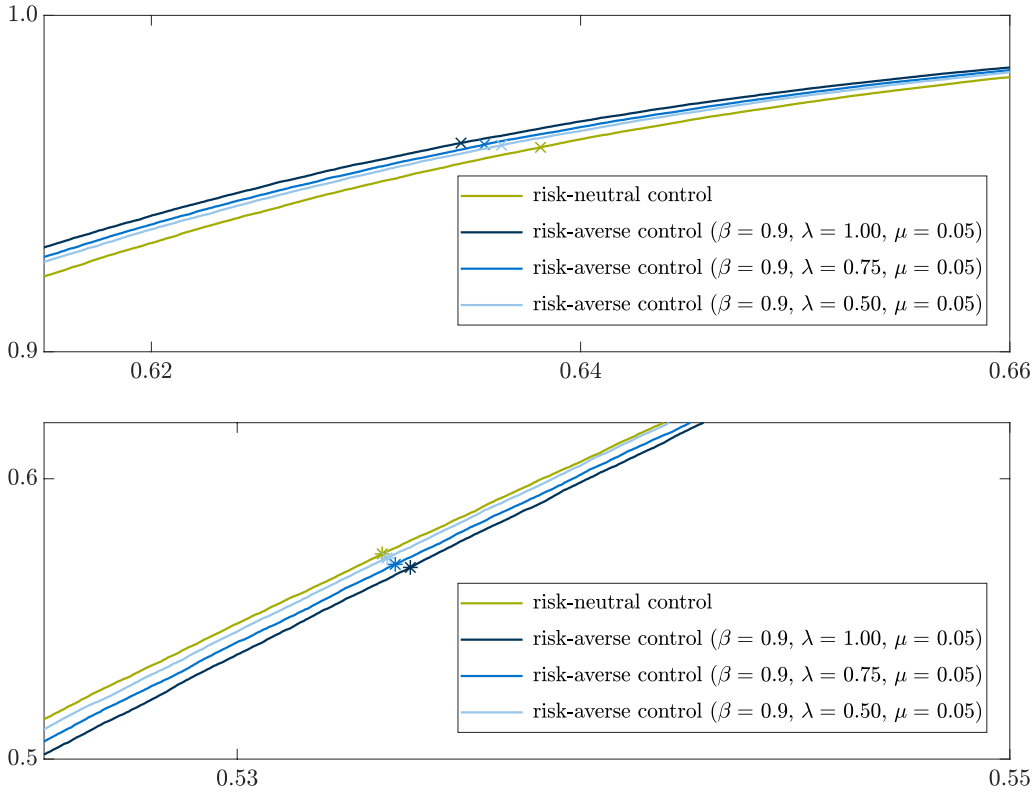



Figure 9.7.: Distribution function of the random variable objective function for the risk-neutral and the risk-averse controls with $\lambda \in \{1.00, 0.75, 0.50\}$ and $(\beta, \mu) = (0.9, 0.05)$

In Figure 9.8 we show the distribution function for different quantile parameters $\beta \in \{0.5, 0.9\}$ with $\lambda = 1.00$, i. e., we minimize a smoothed version of $\mathrm{CVaR}_\beta$ with $\mu = 0.10$ and $\mu = 0.05$, respectively. In both cases, the control resulting from a minimization of the smoothed $\mathrm{CVaR}_{0.9}$ yields a smaller $\mathrm{CVaR}_{0.9}$, but a larger $\mathrm{CVaR}_{0.5}$ than the one which minimizes the the smoothed $\mathrm{CVaR}_{0.5}$. Both are more risk-averse in terms of $\mathrm{CVaR}_{0.5}$ and $\mathrm{CVaR}_{0.9}$ than the risk-neutral control. In the case $\beta = 0.9$, $\mu = 0.05$ this behavior can be observed although the gradient norm is rather large in the computed solution. This solution achieves a smaller $\mathrm{CVaR}_{0.9}$ than the one computed with $\beta = 0.9$ and $\mu = 0.10$.

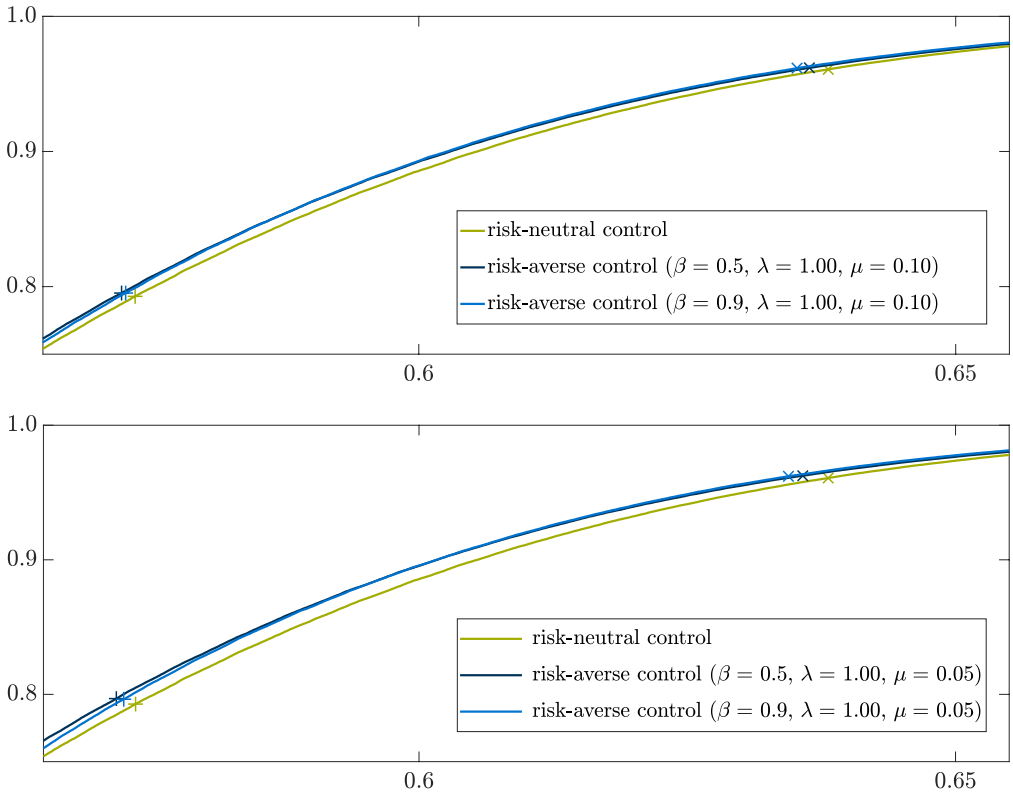



Figure 9.8.: Distribution function of the random variable objective function for the risk-neutral and the risk-averse controls with $\beta \in \{0.5, 0.9\}$, $\lambda = 1.00$, $\mu = 0.10$ (top), $\mu = 0.05$ (bottom)

In conclusion, we have proposed a log-barrier method implemented with low-rank tensors for solving risk-averse optimal control problems under uncertainty with a convex combination of the mean and the CVaR as risk measure. We have shown that keeping the barrier parameter $\mu > 0$ on a fixed level results in the minimization of a smoothed, monotonic, regular risk measure. We have formulated an existence result for convex problems as well as an optimality condition and the barrier-Newton equation in function space. The numerical results have shown that this is a hard problem class if implemented with low-rank tensors as already observed in [46]. Nevertheless, we are capable of computing risk-averse controls by this method, which yield a random variable objective with smaller CVaR than the risk-neutral control.

# 10. Conclusions and Perspectives

In this thesis, we have discussed optimal control problems of semilinear, elliptic PDEs with uncertain inputs and different risk measures. We have provided the necessary theory to formulate the PDEs in weak form in a tensor Banach space and have derived adjoint-based expressions for the derivatives of the reduced objective function in the risk-neutral and in the mean-variance case. A trust-region framework allowing for inexact objective function, gradient, and criticality measure evaluations has been established. It features global convergence while error bounds up to unknown multiplicative constants are sufficient. In the risk-neutral and in the mean variance case, we have investigated how the required error tolerances can be fulfilled based on the error in the state, in the adjoint state, and in the inexact projection onto the set of admissible controls. If the risk measure is the expected value and the problem data is essentially bounded w. r. t. the random parameters, error control of the state and adjoint state in the $L^2_{\mathbb{P}}(\Xi; H^1_0(\Omega))$-norm is sufficient. We have derived an a posteriori estimator for this error based on a stochastic Galerkin discretization with polynomials in tensor product form. It has been discussed how the PDEs with uncertain inputs can be solved adaptively using low-rank tensor calculus and methods. This is necessary to make the computation efficient. Numerical results have shown the adaptivity of the proposed approach to different problem data in the sense that the constructed FE mesh resolves the PDE solution and the active set well and the polynomial grades are chosen dependent on the influence of the respective uncertain parameter on the controlled system.

The error estimation theory is rigorous such that global convergence of the algorithm can be established, but the practical implementation lacks this property at some points. We have used the AMEn method to solve tensor equations because of its efficiency in practice although its convergence theory is still limited to our knowledge. Additionally, it is not clear a priori if approximate solutions to the respective equations with moderate tensor ranks exist and can be computed by AMEn. In our experiments, the required small error tolerance has not been met by AMEn sometimes in the last iterations of the trust-region algorithm. Furthermore, we have bounded the number of used FE nodes due to efficiency reasons. Some terms in the error estimation—appearing only in the case that a nonlinear PDE is considered—have been neglected, namely the interpolation error and the $L^\infty_{\mathbb{P}}(\Xi; H^1_0(\Omega))$-norm of the computed state.

In addition to the mentioned smooth risk measures, we also have discussed risk measures which are convex combinations of the mean and the conditional value-at-risk. They have many favorable properties, but are nonsmooth. We have proposed to apply a log-barrier method to a smooth reformulation of optimal control problems involving such risk measures and have investigated how this procedure affects the underlying risk measures if the barrier parameter is not driven to zero, but kept on a fixed level. In fact, useful properties, such as convexity and monotonicity, are preserved. Numerical results of a low-rank tensor implementation of this method working on a fixed discretization have been shown. The barrier-Newton

system has been solved approximately by a forward-backward block Gauss-Seidel method, which makes use of the structure of the system. The results reveal how the distribution function of the random variable objective function of the optimal control problem can be shaped by this procedure. Although the convergence of Newton's method in the proposed function space setting is well-known, it is hard to derive a practically relevant convergence result for a low-rank tensor implementation of it. Even very small perturbations due to truncation may cause the barrier terms to be not well-defined anymore so that iterative solvers for computing them may fail to converge.

Based on this work, several directions of future research can be pursued. The proposed approach for the analysis (Chapter 3) and the error estimation (Chapters 5 and 7) can be adapted to different settings of optimal control problems with semilinear, elliptic PDEs such as boundary control or different boundary conditions. Furthermore, low-rank methods are also suitable for time-dependent PDEs [105], where the time yields an additional tensor mode. The a posteriori error estimation procedure from Chapter 7 can be generalized to the 3D case and to more general coefficient functions. Additionally, the convergence of the discretization for the considered class of semilinear PDEs should be analyzed. Regarding the mean-variance risk measure, the question arises whether error estimation in $L^q_{\mathbb{P}}(\Xi; H^1_0(\Omega))$ with $q > 2$ is possible. For the ease of implementation, we have bounded all errors based on the $L^2_{\mathbb{P}}(\Xi; H^1_0(\Omega))$-error in the state and adjoint state. By this procedure, we certainly overestimate the error in some cases. As discussed, weaker error control can be sufficient, for instance, an $L^2_{\mathbb{P}}(\Xi; L^2(\Omega))$-error bound for the adjoint state to control the gradient error. In addition, alternative a posteriori error estimation techniques, such as the dual-weighted-residual method [18] for the objective function evaluation, can be employed, but have to be adapted to the stochastic setting.

In our numerical experiments, we have refrained from testing the scaling of the proposed algorithms w. r. t. to the number of the parameters because such tests can be found in [46] and our tests have been dedicated to adaptivity, which itself has required to run different setups. Regarding computing time and exactness of the obtained results, a comprehensive comparison to different treatments of the stochasticity, such as the use of adaptive sparse grids or multilevel Monte Carlo methods, would be interesting. For a fair comparison, one should find a setup for which the exact solution can be constructed, and the sample-based methods should also use adaptive grid refinement. Since they can benefit from parallelization, parallel codes for low-rank tensors, which are investigated currently [52], should be applied.

Based on the derived barrier-Newton system from Section 9.3 or a similar primal-dual system, one can derive an interior-point method for such problems. As shown, the necessary derivatives are Fréchet derivatives if $L^\infty_{\mathbb{P}}(\Xi)$ is considered as underlying function space. If one wants to design an adaptive algorithm for such problems, this space would have to be discretized suitably, which is a hard task because, e. g., the set of polynomials of arbitrary degree is not dense in $L^\infty_{\mathbb{P}}(\Xi)$. To overcome this issue, the space of continuous functions could be used alternatively.

To sum up, it shall be mentioned that many different methods and concepts from, e. g., multilinear algebra, functional analysis and PDE theory, numerical optimization and optimization in Banach spaces, numerical analysis and approximation theory as well as probability theory and uncertainty quantification, have been used in this dissertation. The interplay of them makes the topic very interesting and motivates future research in various directions.

# Danksagung

Ich möchte an dieser Stelle mehreren Personen meinen Dank aussprechen, die zum Gelingen dieser Arbeit beigetragen haben.

Zuerst gilt mein Dank Prof. Dr. Michael Ulbrich, der mich im Rahmen des „TopMath"-Studiengangs bereits seit meinem dritten Studienjahr betreut hat und mich schon im Bachelor-Studium an Themen der Unsicherheitsquantifizierung und der Optimierung mit partiellen Differentialgleichungen heranführte. Sein Vorschlag, Niedrigrangtensormethoden im Bereich der Optimalsteuerung unter Unsicherheit und der Unsicherheitsquantifizierung zu verwenden, führte zu einem spannenden, vielseitigen Promotionsthema. Durch die Anstellung an seinem Lehrstuhl bot sich mir die Möglichkeit, neben der Forschungserfahrung auch umfangreiche Lehrerfahrung zu sammeln.

Prof. Dr. Thomas Surowiec danke ich für die spannende Zusammenarbeit im Bezug auf alternative Risikomaße, was die Basis für Kapitel 9 dieser Arbeit gebildet hat.

Außerdem freue ich mich, dass sich Prof. Dr. Christian Clason und Prof. Dr. Matthias Heinkenschloss bereiterklärt haben, diese Arbeit zu begutachten.

Für die Förderung ideeller und finanzieller Art möchte ich mich beim Elitestudienprogramm „TopMath – Mathematik mit Promotion" und bei der International Research Training Group „IGDK Munich – Graz: Optimization and Numerical Analysis for Partial Differential Equations with Nonsmooth Structures" und allen Verantwortlichen bedanken. Durch TopMath konnte ich schon früh sehr frei und forschungsorierientiert studieren. Beide Programme ermöglichten mir die Teilnahme an interessanten Konferenzen, Workshops und Kursen.

Außerdem sollen sämtliche Kollegen und Kommilitonen aus dem Studium, von TopMath, dem IGDK und vom Lehrstuhl für Mathematische Optimierung nicht unerwähnt bleiben, ohne dass hier jeder persönlich genannt wird. Ich bedanke mich für hilfreiche fachliche Diskussionen und für die gute Zeit und Zusammenarbeit. Gesondert erwähnen will ich Johannes Milz und Johannes Haubner, die Teile dieser Arbeit genau gelesen und mit Anmerkungen versehen haben, und meinen Mentor Dr. Andre Milzarek, der mir gerade zu Beginn der Promotionszeit mit hilfreichen Ratschlägen zur Seite stand.

Zuletzt gebührt mein Dank meiner Familie, meinen Freunden und Gela. Danke, dass ihr für mich da seid, mich unterstützt und dass wir so viel gemeinsam erleben dürfen.

# A. Appendix

## A.1. Tensor Spaces

We give the proof of Proposition 2.6 here since it does not contribute to the main content of the thesis.

**Proposition A.1** (Characterization of equivalence classes)**.** *Let $V$ and $W$ be vector spaces over a field $\mathbb{K}$ and let $v, \tilde{v} \in V$ and $w, \tilde{w} \in W$ be given. Consider the algebraic tensor space $V \otimes_a W$. The pair $(\tilde{v}, \tilde{w})$ belongs to the equivalence class $v \otimes w$ if and only if the following holds:*

$$(v = 0 \vee w = 0) \wedge (\tilde{v} = 0 \vee \tilde{w} = 0) \tag{A.1}$$

*or*

$$(v \neq 0 \wedge w \neq 0) \vee (\tilde{v} \neq 0 \wedge \tilde{w} \neq 0) \quad \textit{and} \quad \exists\, c \in \mathbb{K}\backslash\{0\} \textit{ s. t. } \tilde{v} = cv \textit{ and } w = c\tilde{w}. \tag{A.2}$$

*Case* (A.1) *holds exactly for pairs belonging to the equivalence class* $0 \otimes 0$.

*Proof.* First we show that (A.1) holds if and only if both tensors $v \otimes w$ and $\tilde{v} \otimes \tilde{w}$ are in fact zero, i. e., they belong to the equivalence class $0 \otimes 0$: For some $v \in V$ we have $(v, 0) \in \mathcal{N}$ with $\mathcal{N}$ defined as in Definition 2.5 taking $m = n = 1$, $\alpha_1 = 1$, $\beta_1 = 0$, $v_1 = v$, and $w_1 = 0$, and multiplying the resulting tensor by $-1$. Analogously we get $(0, w) \in \mathcal{N}$ for any $w \in W$.

Now we prove that the condition $v = 0 \vee w = 0$, cf. (A.1), is also necessary for a pair $(v, w)$ to belong to $0 \otimes 0 = \mathcal{N}$. First observe that it is enough to allow only vectors from given bases $B \subset V$ and $C \subset W$ in the definition of the set $\mathcal{N}$: We have $\mathcal{N} = \tilde{\mathcal{N}}$ with

$$\tilde{\mathcal{N}} := \operatorname{span}\Big\{ \sum_{i=1}^{m}\sum_{j=1}^{n} \alpha_i\beta_j(v_i, w_j) - \Big(\sum_{i=1}^{m} \alpha_i v_i, \sum_{j=1}^{n} \beta_j w_j\Big) : \\ m, n \in \mathbb{N}, \alpha_i, \beta_j \in \mathbb{K}, v_i \in B, w_j \in C \Big\}$$

We only show that linear combinations of pairs of the form

$$\sum_{i=1}^{m}\sum_{j=1}^{n} \alpha_i\beta_j(v_i, w_j) - \Big(\sum_{i=1}^{m} \alpha_i v_i, \sum_{j=1}^{n} \beta_j w_j\Big)$$

with general $v_i \in V$ and $w_i \in W$ belong to the set $\tilde{\mathcal{N}}$ because the inclusion $\tilde{\mathcal{N}} \subset \mathcal{N}$ is obvious. We insert the representations $v_i = \sum_{v \in B} \lambda_v^i v$ and $w_j = \sum_{w \in C} \mu_w^j w$, which are in fact finite

sums, into the formula and compute

$$\begin{aligned}
&\sum_{i=1}^{m}\sum_{j=1}^{n}\alpha_i\beta_j(v_i,w_j)-\Big(\sum_{i=1}^{m}\alpha_i v_i,\sum_{j=1}^{n}\beta_j w_j\Big)\\
&=\sum_{i=1}^{m}\sum_{j=1}^{n}\alpha_i\beta_j\big(\sum_{v\in B}\lambda_v^i v,\sum_{w\in C}\mu_w^j w\big)-\Big(\sum_{i=1}^{m}\alpha_i\sum_{v\in B}\lambda_v^i v,\sum_{j=1}^{n}\beta_j\sum_{w\in C}\mu_w^j w\Big)\\
&=\sum_{i=1}^{m}\sum_{j=1}^{n}\alpha_i\beta_j\Big(\sum_{v\in B}\lambda_v^i v,\sum_{w\in C}\mu_w^j w\Big)-\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{v\in B}\sum_{w\in C}\alpha_i\beta_j\lambda_v^i\mu_w^j(v,w)\\
&\quad+\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{v\in B}\sum_{w\in C}\alpha_i\beta_j\lambda_v^i\mu_w^j(v,w)-\Big(\sum_{i=1}^{m}\sum_{v\in B}\alpha_i\lambda_v^i v,\sum_{j=1}^{n}\sum_{w\in C}\beta_j\mu_w^j w\Big).
\end{aligned}$$

This is a linear combination belonging to the set $\tilde{\mathcal{N}}$. Now let $(v,w)\in\mathcal{N}$, i. e.,

$$\begin{aligned}
(v,w)&=\sum_{k=1}^{K}\gamma_k\Big(\sum_{i=1}^{m}\sum_{j=1}^{n}\alpha_i^k\beta_j^k(v_i^k,w_j^k)-\Big(\sum_{i=1}^{m}\alpha_i^k v_i^k,\sum_{j=1}^{n}\beta_j^k w_j^k\Big)\Big)\\
&=\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n}\gamma_k\alpha_i^k\beta_j^k(v_i^k,w_j^k)-\sum_{k=1}^{K}\gamma_k\Big(\sum_{i=1}^{m}\alpha_i^k v_i^k,\sum_{j=1}^{n}\beta_j^k w_j^k\Big)
\end{aligned}\tag{A.3}$$

for some $m,n,K\in\mathbb{N}$, $\alpha_i^k,\beta_j^k,\gamma_k\in\mathbb{K}$, $v_i^k\in B$, $w_j^k\in C$. We assume that $v\neq 0$ and $w\neq 0$ holds. Thus, we can take the bases $B$ and $C$ such that $v\in B$ and $w\in C$. Furthermore, we take the shortest possible representation, i. e., $K$ to be minimal. That gives us that $\gamma_k\neq 0$ and that $\alpha^k$ and $\beta^k$ cannot both be unit vectors for all $k$ because both cases would lead to a zero summand and the sum could be shortened. Therefore, $(v,w)$ is not contained in the second sum over $k$ in (A.3), which yields $(v,w)=\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n}\gamma_k\alpha_i^k\beta_j^k(v_i^k,w_j^k)$ and $\sum_{k=1}^{K}\gamma_k\left(\sum_{i=1}^{m}\alpha_i^k v_i^k,\sum_{j=1}^{n}\beta_j^k w_j^k\right)=0$. Now assume that a non-zero pair is contained in the second sum. Since $\gamma_k\neq 0$, it has to be contained again at least once, such that it cancels out. Because of the unique basis representation the respective terms would also cancel in the first sum over $k$, which is a contradiction to the minimality of $K$. Therefore we have $K=1$ and w. l. o. g. $\gamma_1=1$. This gives $\left(\sum_{i=1}^{m}\alpha_i^k v_i^k,\sum_{j=1}^{n}\beta_j^k w_j^k\right)=0$ and thus $\sum_{i=1}^{m}\alpha_i^1 v_i^1=0$ and $\sum_{j=1}^{n}\beta_j^1 w_j^1=0$ and therefore $\alpha^1=0$ and $\beta^1=0$ due to the linear independence. We obtain $(v,w)=0$, which is a contradiction to the assumption that $v\neq 0$ and $w\neq 0$.

In the second case (A.2) we have that $(v,w)-(\tilde{v},\tilde{w})=(v,c\tilde{w})-(cv,\tilde{w})\in\mathcal{N}$ by taking $m=n=1$, $\alpha_1=c$, $\beta_1=c^{-1}$, $v_1=v$ and $w_1=c\tilde{w}$.

It remains to prove the "only if" part. Let $(\tilde{v},\tilde{w})\in v\otimes w$, i. e., $(v,w)-(\tilde{v},\tilde{w})\in\mathcal{N}$. Assume that (A.1) does not hold meaning that the first part of (A.2) holds and we have to show the second part. If $v=\tilde{v}$ and $w=\tilde{w}$, we can take $c=1$ and are done; now we assume $c\notin\{0,1\}$.

First observe that if $\tilde{v}=cv$ holds, we get $w=c\tilde{w}$: Represent $(v,w)-(cv,\tilde{w})$ as a shortest linear combination of the form (A.3) with $v\in B$, $w\in C$. Since $cv\notin B$ we can argue as above that $(v,w)$ is contained in the first part and $(cv,\tilde{w})$ is contained in the second part. Assuming that other pairs than $(cv,\tilde{w})$ are contained in the second sum, we see again that the terms containing these would cancel in both sums and that we can take $K=1$ and

$\gamma_1 = 1$. We have $(cv, \tilde{w}) = \left(\sum_{i=1}^m \alpha_i^1 v_i^1, \sum_{j=1}^n \beta_j^1 w_j^1\right)$ and thus $\alpha^1 = ce^1$ if we take w. l. o. g. $v_1^1 = v$, where $e^i$ is the $i$-th unit vector. Since $(v, w)$ is contained in the first part, we get $(v, w) = \sum_{j=1}^n c\beta_j (v, w_j^1)$ and therefore $\beta^1 = c^{-1}e^1$ if $w_1^1 = w$. This gives us finally that $(cv, \tilde{w}) = (cv, c^{-1}w)$ and hence $w = c\tilde{w}$. Analogously it follows from $(v, w) - (\tilde{v}, \tilde{w}) \in \mathcal{N}$ and $w = c\tilde{w}$ that $\tilde{v} = cv$.

Now we assume that the second part of (A.2) does not hold. That means that w. l. o. g. $v$ and $\tilde{v}$ are linearly independent. We get from the considerations above that also $w$ and $\tilde{w}$ are linearly independent. In the representation of $(v, w) - (\tilde{v}, \tilde{w})$ as in (A.3) we can take bases $B$ and $C$ containing both vectors, respectively: $v, \tilde{v} \in B$, $w, \tilde{w} \in C$. Again we argue that $(v, w) - (\tilde{v}, \tilde{w})$ is contained in the first sum and that we get $K = 1$ and $\gamma_1 = 1$ leading to $0 = \left(\sum_{i=1}^m \alpha_i^1 v_i^1, \sum_{j=1}^n \beta_j^1 w_j^1\right)$. From that we obtain $\alpha^1 = 0$ and $\beta^1 = 0$, which gives $(v, w) - (\tilde{v}, \tilde{w}) = (0, 0)$, which contradicts the assumption that $v$ and $\tilde{v}$ are linearly independent. Since $v$ and $\tilde{v}$ are linearly dependent and both non-zero, there exists a constant $c \in \mathbb{K}\backslash\{0\}$ such that $\tilde{v} = cv$ holds. As shown above, we get that $w = c\tilde{w}$ is also true. □

## A.2. General $L^p$ Spaces and Operator Theory

We discuss some general results about embeddings and interpolation of $L^p$ spaces, which are used frequently for error and regularity estimates. Additionally, we state some results on operators between Banach spaces used for existence theory and error estimation of the considered PDEs.

**Proposition A.2** (Estimating $L^p$-norms on finite measure spaces)**.** *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space with $0 < \mu(\Omega) < \infty$. Let $p \in [1, \infty]$ and $q \in [1, p]$ be given and let $v \in L_\mu^p(\Omega)$. Then, $v \in L_\mu^q(\Omega)$ and we have the estimate*

$$\|v\|_{L_\mu^q(\Omega)} \leq C\|v\|_{L_\mu^p(\Omega)}$$

*with $C = \mu(\Omega)^{1/q - 1/p}$. In case that $\mu$ is a probability measure, $\|v\|_{L_\mu^q(\Omega)} \leq \|v\|_{L_\mu^p(\Omega)}$ holds.*

*Proof.* By Hölder's inequality we get

$$\|v\|_{L_\mu^q(\Omega)}^q = \|1 \cdot |v|^q\|_{L_\mu^1(\Omega)} \leq \|1\|_{L_\mu^{(1-q/p)^{-1}}(\Omega)} \cdot \||v|^q\|_{L_\mu^{p/q}(\Omega)} = \mu(\Omega)^{1-q/p} \cdot \|v\|_{L_\mu^p(\Omega)}^q$$

and therefore the desired result. As usual, we define $0^{-1} = \infty$ in the case $q = p$ and get $\|1\|_{L_\mu^\infty(\Omega)} = 1 = \mu(\Omega)^0$. If $\mu$ is a probability measure, $\mu(\Omega) = 1$ and thus $C = 1$ holds. □

**Proposition A.3.** *Let $A, A_0 : Y \to Y^*$ be bounded, self-adjoint and boundedly invertible linear operators between a Hilbert space $Y$ and its dual space. We define the inner product $(y, v)_A := \langle Ay, v\rangle_{Y^*,Y}$ and norm $\|y\|_A := \sqrt{(y, y)_A}$ for $y, v \in Y$, and analogously the inner products induced by $A_0$ on $Y$ and by $A^{-1}$, $A_0^{-1}$ on $Y^*$. The norms induced by $A$ and $A_0$ shall be equivalent with the constants $\sqrt{\lambda}$ and $\sqrt{\Lambda}$ for $0 < \lambda \leq \Lambda$, i. e.,*

$$\lambda\|y\|_A^2 \leq \|y\|_{A_0}^2 \leq \Lambda\|y\|_A^2 \tag{A.4}$$

*holds for every $y \in Y$.*

*Then, the estimate*

$$\tfrac{1}{\Lambda}\|b\|_{A^{-1}}^2 \leq \|b\|_{A_0^{-1}}^2 \leq \tfrac{1}{\lambda}\|b\|_{A^{-1}}^2 \tag{A.5}$$

*holds for every* $b \in Y^*$.

*Proof.* From the first inequality in (A.4), it follows that

$$\lambda\|y\|_A^2 \leq \langle A_0 y, y\rangle_{Y^*,Y} = \langle AA^{-1}A_0 y, y\rangle_{Y^*,Y} \leq \|A^{-1}A_0 y\|_A \|y\|_A = \|A_0 y\|_{A^{-1}}\|y\|_A$$

and therefore $\|y\|_A \leq \frac{1}{\lambda}\|A_0 y\|_{A^{-1}}$ holds for all $y \in Y$. This induces with $b = A_0 y$, that $\|b\|_{A_0^{-1}}^2 = \|y\|_{A_0}^2 \leq \frac{1}{\lambda}\|A_0 y\|_{A^{-1}}^2 = \frac{1}{\lambda}\|b\|_{A^{-1}}^2$ holds for every $b \in Y^*$, which is the second inequality in (A.5). We have used the Cauchy-Schwarz inequality and that the operators $A$ and $A_0$ are invertible. The rest of the prerequisites is only needed to make sure that the respective operators induce inner products and that (A.4) holds.

In the same fashion, the second inequality in (A.4), written as $\frac{1}{\Lambda}\|y\|_{A_0}^2 \leq \|y\|_A^2$, induces the estimate $\|b\|_{A^{-1}}^2 \leq \Lambda\|b\|_{A_0^{-1}}^2$, which is equivalent to the first inequality in (A.5). □

**Proposition A.4.** *Let* $N : Y \to Y^*$ *be a monotone operator between a real, reflexive Banach space* $Y$ *and its dual space. Let* $N$ *be Gâteaux-differentiable at* $y \in Y$. *Then, the derivative* $N'(y) : Y \to Y^*$ *is also a monotone operator.*

*Proof.* Since $N'(y)$ is linear, it is sufficient to prove that $\langle N'(y)v, v\rangle_{Y^*,Y} \geq 0$ holds for all $v \in Y$. By the definition of the Gâteaux derivative and the continuity of $N'(y)$, we have

$$\begin{aligned}\langle N'(y)v, v\rangle_{Y^*,Y} &= \lim_{t\to 0^+} \tfrac{1}{t}\langle N(y+tv) - N(y), v\rangle_{Y^*,Y} \\ &= \lim_{t\to 0^+} \tfrac{1}{t^2} \underbrace{\langle N(y+tv) - N(y), y + tv - y\rangle_{Y^*,Y}}_{\geq 0} \geq 0,\end{aligned}$$

which is non-negative because of the monotonicity of $N$. □

## A.3. Superposition Operators between $L^p$ Spaces

In this section, we discuss the properties of nonlinear superposition operators between general $L^p$ spaces. We use this theory later to show required properties of the superposition operators between the state space $L^p_{\mathbb{P}}(\Xi; H_0^1(\Omega))$ and its dual.

**Theorem A.5** (Well-definedness and continuity of superposition operators between $L^p$ spaces)**.** *Let* $(\Omega, \mathcal{A}, \mu)$ *be a measure space with a* $\sigma$*-finite measure* $\mu$ *and let* $p, q \in [1, \infty)$. *Let* $\varphi : \Omega \times \mathbb{R} \to \mathbb{R}$ *be a Carathéodory function, i. e., the function* $\varphi(\cdot, t)$ *is measurable for all* $t \in \mathbb{R}$ *and the function* $\varphi(\omega, \cdot)$ *is continuous for a. e.* $\omega \in \Omega$. *Furthermore, let* $\varphi$ *fulfill the growth condition*

$$|\varphi(\omega, t)| \leq c_q(\omega) + c_\varphi |t|^{p/q} \quad \text{for all } t \in \mathbb{R} \text{ and a. e. } \omega \in \Omega$$

*with some function* $c_q \in L_\mu^q(\Omega)$ *and a constant* $c_\varphi \geq 0$.

*Then, the superposition operator*

$$N : L^p_\mu(\Omega) \to L^q_\mu(\Omega),\ N(y)(\omega) := \varphi(\omega, y(\omega))$$

*is well-defined and continuous.*

*Proof.* The theorem is a consequence of [5, Thm. 3.1, Lem. 1.5, Thm. 1.1, Thm. 3.7]. [5, Thm. 3.1] provides a necessary and sufficient condition for the operator $N$ to be well-defined. We use only the sufficient part and tighten the conditions a bit by requiring the growth conditions for (almost) every $\omega$, $t$ and using the Carathéodory property of $\varphi$ and [5, Lem. 1.5, Thm. 1.1]. [5, Thm. 3.7] provides the continuity of the operator (and also some equivalence relation in a more general framework). □

**Remark A.6.** Note that [5] is a suitable reference for this quite known statement although it covers the topic in a very general setting so that it is necessary to combine various theorems and lemmas from this book. Other references such as [48] restrict the discussion to domains equipped with the Lebesgue measure so that, e. g., the atomic part of the probability measure $\mathbb{P}$ is not taken into account. This is especially important when providing necessary conditions for, e. g., boundedness, which can be seen in the following theorem, whose analog [48, Thm. 3] lacks this generality.

**Theorem A.7** (Boundedness of superposition operators between $L^p$ spaces)**.** *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space with a $\sigma$-finite measure $\mu$ and let $p, q \in [1, \infty)$. Let $\varphi : \Omega \times \mathbb{R} \to \mathbb{R}$ be a sup-measurable function and let $N : L^p_\mu(\Omega) \to L^q_\mu(\Omega)$ (defined as in Theorem A.5) be well-defined.*
*Then, $N$ is locally bounded if and only if the function $\varphi(\omega, \cdot)$ is bounded for each $\omega \in \Omega_d$, where $\Omega_d \subset \Omega$ is the purely atomic part of the measure $\mu$.*

*Proof.* See [5, Thm. 3.2]. □

**Remark A.8.**

- If $\varphi : \Omega \times \mathbb{R} \to \mathbb{R}$ is a Carathéodory function, it is sup-measurable [5, Thm. 1.1].
- The fact that $N$ is locally bounded means that $\limsup_{v \to y} \|N(v)\|_{L^q_\mu(\Omega)} < \infty$ holds for all $y \in L^p_\mu(\Omega)$.
- If $\mu$ is non-atomic, Theorem A.7 gives that the superposition operator $N$ is automatically locally bounded if it is well-defined.

**Theorem A.9** (Differentiability of superposition operators between $L^p$ spaces)**.** *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space with a $\sigma$-finite measure $\mu$ and let $p, q \in [1, \infty)$ with $p > q$. Let $\varphi : \Omega \times \mathbb{R} \to \mathbb{R}$ be a Carathéodory function (see Theorem A.7) and let the restriction $\varphi(\omega, \cdot)$ be continuously differentiable for a. e. $\omega \in \Omega$. Let the function $\varphi$ fulfill the growth condition*

$$|\varphi(\omega, t)| \leq c_q(\omega) + c_\varphi |t|^{p/q} \quad \text{for all } t \in \mathbb{R} \text{ and a. e. } \omega \in \Omega \tag{A.6}$$

*with some function* $c_q \in L^q_\mu(\Omega)$ *and a constant* $c_\varphi \geq 0$. *Furthermore, let the partial derivative* $\varphi_t : \Omega \times \mathbb{R} \to \mathbb{R}$ *fulfill the growth condition*

$$|\varphi_t(\omega, t)| \leq c'_{pq/(p-q)}(\omega) + c'_{\varphi_t}|t|^{(p-q)/q} \quad \text{for all } t \in \mathbb{R} \text{ and a. e. } \omega \in \Omega \tag{A.7}$$

*with some function* $c'_{pq/(p-q)} \in L^{pq/(p-q)}_\mu(\Omega)$ *and a constant* $c'_{\varphi_t} \geq 0$.
*Then, the superposition operator* $N : L^p_\mu(\Omega) \to L^q_\mu(\Omega)$ *induced by* $\varphi$ *(defined as in Theorem A.5) is continuously Fréchet-differentiable with derivative*

$$[N'(y)v](\omega) = \varphi_t(\omega, y(\omega))v(\omega).$$

*Proof.* We adapt the proof from the very similar Proposition A.11 in [111].

Since $\varphi$ is a Carathéodory function and condition (A.6) holds, the operator $N$ is well-defined and continuous by Theorem A.5.

Moreover, $\varphi_t$ is a Carathéodory function: Since $\varphi(\omega, \cdot)$ is continuously differentiable for a. e. $\omega \in \Omega$, the partial derivative $\varphi_t(\omega, \cdot)$ is continuous for a. e. $\omega \in \Omega$. By definition, $\varphi_t(\omega, t) = \lim_{h\to 0} \frac{1}{h}(\varphi(\omega, t+h) - \varphi(\omega, t))$ holds. Now, for a sequence $(h_n)_{n\in\mathbb{N}} \subset \mathbb{R} \setminus \{0\}$ converging to zero, we have that the difference quotient $\varphi_t^{(n)}(\omega, t) := \frac{1}{h_n}(\varphi(\omega, t+h_n) - \varphi(\omega, t))$ is measurable w. r. t. $\omega$ for all $t \in \mathbb{R}$ as sum of measurable functions. The pointwise limit $\varphi_t(\cdot, t)$ of these measurable functions is thus also measurable.

Due to the Carathéodory property and condition (A.7), the superposition operator $N_{\varphi_t} : L^p_\mu(\Omega) \to L^{pq/(p-q)}_\mu(\Omega)$ generated by $\varphi_t$ is also well-defined and continuous by Theorem A.5. Note that the identification $\mathcal{L}(L^p_\mu(\Omega), L^q_\mu(\Omega)) \cong L^{pq/(p-q)}_\mu(\Omega)$ can be made by the Riesz representation theorem and Hölder's inequality. This gives $N'(y) \cong N_{\varphi_t}(y)$ and that the derivative $N' : L^p_\mu(\Omega) \to \mathcal{L}(L^p_\mu(\Omega), L^q_\mu(\Omega))$ as defined in the theorem is well-defined and continuous, cf. the proof of [111, Prop. A.11] for more details.

Moreover,

$$\begin{aligned}
&\|N(y+v) - N(y) - N'(y)v\|_{L^q_\mu(\Omega)} \\
&= \|\varphi(\cdot, y(\cdot) + v(\cdot)) - \varphi(\cdot, y(\cdot)) - \varphi_t(\cdot, y(\cdot))v(\cdot)\|_{L^q_\mu(\Omega)} \\
&= \Big\| \int_0^1 [\varphi_t(\cdot, y(\cdot) + \sigma v(\cdot)) - \varphi_t(\cdot, y(\cdot))]v(\cdot)\,\mathrm{d}\sigma \Big\|_{L^q_\mu(\Omega)} \\
&\leq \int_0^1 \|[\varphi_t(\cdot, y(\cdot) + \sigma v(\cdot)) - \varphi_t(\cdot, y(\cdot))]v(\cdot)\|_{L^q_\mu(\Omega)}\,\mathrm{d}\sigma \\
&\leq \Big( \int_0^1 \|\varphi_t(\cdot, y(\cdot) + \sigma v(\cdot)) - \varphi_t(\cdot, y(\cdot))\|_{L^{pq/(p-q)}_\mu(\Omega)}\,\mathrm{d}\sigma \Big) \|v(\cdot)\|_{L^p_\mu(\Omega)} \\
&\leq \big( \sup_{\sigma\in[0,1]} \|\varphi_t(\cdot, y(\cdot) + \sigma v(\cdot)) - \varphi_t(\cdot, y(\cdot))\|_{L^{pq/(p-q)}_\mu(\Omega)} \big) \|v(\cdot)\|_{L^p_\mu(\Omega)} \\
&= o(\|v\|_{L^p_\mu(\Omega)}) \quad \text{as } \|v\|_{L^p_\mu(\Omega)} \to 0
\end{aligned}$$

follows from Hölder's inequality and the fact that $N_{\varphi_t}$ is continuous. This shows that $N$ is Fréchet-differentiable with the given derivative $N'$. □

Alternatively, if $\mu$ is a finite measure, it is sufficient to require $\varphi(\cdot, 0) \in L^q_\mu(\Omega)$ and a suitable growth condition only on $\varphi_t(\omega, \cdot)$. Then, condition (A.6) follows directly:

**Lemma A.10.** *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space with $\mu(\Omega) < \infty$ and let $p, q \in [1, \infty)$ with $p > q$. Let $\varphi : \Omega \times \mathbb{R} \to \mathbb{R}$ be a Carathéodory function, see Theorem A.7, and let the restriction $\varphi(\omega, \cdot)$ be continuously differentiable for a. e. $\omega \in \Omega$. Furthermore, let the partial derivative $\varphi_t : \Omega \times \mathbb{R} \to \mathbb{R}$ fulfill the $\omega$-independent growth condition*

$$|\varphi_t(\omega, t)| \leq a'_{\varphi_t} + c'_{\varphi_t}|t|^{(p-q)/q} \quad \text{for all } t \in \mathbb{R} \text{ and a. e. } \omega \in \Omega,$$

*with two constants $a'_{\varphi_t}, c'_{\varphi_t} \geq 0$ and let $\varphi(\cdot, 0) \in L^q_\mu(\Omega)$.*
*Then, the function $\varphi(\omega, \cdot)$ fulfills*

$$|\varphi(\omega, t)| \leq c_q(\omega) + c_\varphi |t|^{p/q}$$

*for a. e. $\omega \in \Omega$ with $c_q \in L^q(\Omega)$ and a constant $c_\varphi \in \mathbb{R} \geq 0$.*

*Proof.* We use a part of the proof of [111, Prop. A.11]. For a. e. $\omega \in \Omega$, we have

$$\begin{aligned}
|\varphi(\omega, t)| &\leq |\varphi(\omega, 0)| + \int_0^1 |\varphi_t(\omega, \sigma t) t| \, \mathrm{d}\sigma \\
&\leq |\varphi(\omega, 0)| + |t| \int_0^1 a'_{\varphi_t} + c'_{\varphi_t} |\sigma t|^{(p-q)/q} \, \mathrm{d}\sigma \\
&\leq |\varphi(\omega, 0)| + a'_{\varphi_t}|t| + \tfrac{q c'_{\varphi_t}}{p} |t|^{p/q} \leq c_q(\omega) + c_\varphi |t|^{p/q}
\end{aligned}$$

with some $c_\varphi \geq 0$ and $c_q \in L^q_\mu(\Omega)$ because $\varphi(\cdot, 0) \in L^q(\Omega)$ and constant functions belong to every $L^q_\mu(\Omega)$ since $\mu$ is finite. □

## A.4. Superposition Operators from $L^p_{\mathbb{P}}(\Xi; H^1_0(\Omega))$ to Its Dual

For an open, bounded Lipschitz domain $\Omega \subset \mathbb{R}^n$ ($n \in \{2, 3\}$) equipped with the Lebesgue measure $\lambda$, we define the deterministic state space $Y := H^1_0(\Omega)$ and its dual $Y^* = H^{-1}(\Omega)$. Furthermore, let $\Xi \subset \mathbb{R}^m$ be measurable and equipped with the probability measure $\mathbb{P}$. For $p \in [2, \infty)$ (if $n = 2$) and $p \in [2, 6]$ (if $n = 3$) we define the Bochner space $\boldsymbol{Y} := L^p_{\mathbb{P}}(\Xi; Y)$. We have the Sobolev embedding $H^1_0(\Omega) \hookrightarrow L^q(\Omega)$, i. e., $\|y\|_{L^q(\Omega)} \leq C_q \|y\|_{H^1_0(\Omega)}$ with some constant $C_q > 0$ for $q \in [1, \infty)$ if $n = 2$ and $q \in [1, 6]$ if $n = 3$. Now, for $p = q$ we can embed the space $\boldsymbol{Y} \hookrightarrow L^p_{\mathbb{P}}(\Xi; L^p(\Omega)) \cong L^p_{\lambda \otimes \mathbb{P}}(\Omega \times \Xi)$ and then use existing $L^p$-theory. The respective dual spaces are identified with $L^{p*}_{\lambda \otimes \mathbb{P}}(\Omega \times \Xi) \cong L^{p^*}_{\mathbb{P}}(\Xi; L^{p^*}(\Omega)) \hookrightarrow L^{p^*}_{\mathbb{P}}(\Xi; H^{-1}(\Omega)) = \boldsymbol{Y}^*$ with the conjugate exponent $p^* = \frac{p}{p-1} \in (1, 2]$ if $n = 2$ and $p^* \in [\frac{1}{6}, 1]$ if $n = 3$.

Now we consider a function $\varphi : \Omega \times \Xi \times \mathbb{R} \to \mathbb{R}$ with some of the following properties:

**Assumption A.11.**

1. $\varphi$ satisfies the Carathéodory property, i. e., $\varphi(\cdot, \cdot, t)$ is measurable for all $t \in \mathbb{R}$ and $\varphi(x, \xi, \cdot)$ is continuous for a. e. $x \in \Omega$, $\xi \in \Xi$.

2. $\varphi$ satisfies the growth condition $|\varphi(x, \xi, t)| \leq c_{p*}(x, \xi) + c_\varphi |t|^{p-1}$ for all $t \in \mathbb{R}$ and a. e. $x \in \Omega$, $\xi \in \Xi$ with some constant $c_\varphi \geq 0$ and a function $c_{p^*} \in L^{p^*}_{\lambda \otimes \mathbb{P}}(\Omega \times \Xi)$.

3. The function $\varphi(x, \xi, \cdot)$ is monotonically increasing for a. e. $x \in \Omega$, $\xi \in \Xi$.

4. For a. e. $x \in \Omega$, $\xi \in \Xi$, the function $\varphi(x, \xi, \cdot)$ is continuously differentiable with partial derivative $\varphi_t : \Omega \times \Xi \times \mathbb{R} \to \mathbb{R}$.

5. It holds that $p > 2$ and the first partial derivative satisfies the growth condition

$$|\varphi_t(x, \xi, t)| \leq c'_{p/(p-2)}(x, \xi) + c'_{\varphi_t}|t|^{p-2} \text{ for all } t \in \mathbb{R} \text{ and a. e. } x \in \Omega,\ \xi \in \Xi$$

with some constant $c'_{\varphi_t} \geq 0$ and a function $c'_{p/(p-2)} \in L^{p/(p-2)}_{\lambda \otimes \mathbb{P}}(\Omega \times \Xi)$.

6. For a. e. $x \in \Omega$, $\xi \in \Xi$, the function $\varphi(x, \xi, \cdot)$ is twice continuously differentiable with first partial derivative $\varphi_t : \Omega \times \Xi \times \mathbb{R} \to \mathbb{R}$ and second partial derivative $\varphi_{tt} : \Omega \times \Xi \times \mathbb{R} \to \mathbb{R}$.

7. It holds that $p > 3$ and the second partial derivative satisfies the growth condition

$$|\varphi_{tt}(x, \xi, t)| \leq c''_{p/(p-3)}(x, \xi) + c''_{\varphi_{tt}}|t|^{p-3} \text{ for all } t \in \mathbb{R} \text{ and a. e. } x \in \Omega,\ \xi \in \Xi$$

with some constant $c''_{\varphi_{tt}} \geq 0$ and a function $c''_{p/(p-3)} \in L^{p/(p-3)}_{\lambda \otimes \mathbb{P}}(\Omega \times \Xi)$.

Using $\varphi$, we define the nonlinear operator

$$\boldsymbol{N} : \boldsymbol{Y} \to \boldsymbol{Y}^*, \quad \langle \boldsymbol{N}(\boldsymbol{y}), \boldsymbol{v} \rangle_{\boldsymbol{Y}^*, \boldsymbol{Y}} := \int_\Xi \int_\Omega \varphi(x, \xi, \boldsymbol{y}(x, \xi)) \boldsymbol{v}(x, \xi) \, \mathrm{d}x \, \mathrm{d}\mathbb{P} \quad \forall\, \boldsymbol{y}, \boldsymbol{v} \in \boldsymbol{Y}.$$

We investigate its properties:

**Proposition A.12.** *Under Assumption A.11:1-2, the operator $\boldsymbol{N} : \boldsymbol{Y} \to \boldsymbol{Y}^*$ is well-defined and continuous.*

*Proof.* We have that $\boldsymbol{N}$ is well-defined and continuous as an operator from $L^p_{\lambda \otimes \mathbb{P}}(\Omega \times \Xi)$ to $L^{p^*}_{\lambda \otimes \mathbb{P}}(\Omega \times \Xi)$ by Theorem A.5. Using the continuous embeddings $\boldsymbol{Y} \hookrightarrow L^p_{\lambda \otimes \mathbb{P}}(\Omega \times \Xi)$ and $\boldsymbol{Y}^* \hookleftarrow L^{p^*}_{\lambda \otimes \mathbb{P}}(\Omega \times \Xi)$ as described above, we get that it is also well-defined and continuous from $\boldsymbol{Y}$ to $\boldsymbol{Y}^*$. □

**Proposition A.13.** *Given Assumption A.11:1-3, the operator $\boldsymbol{N} : \boldsymbol{Y} \to \boldsymbol{Y}^*$ is monotone.*

*Proof.* By Proposition A.12, the operator $\boldsymbol{N}$ is well-defined. Since $\varphi(x, \xi, \cdot)$ is increasing for a. e. $x, \xi$ by Assumption A.11:3, $\varphi(x, \xi, t) \leq \varphi(x, \xi, \tilde{t})$ holds if $t \leq \tilde{t}$. This gives $\varphi(x, \xi, \tilde{t}) - \varphi(x, \xi, t) \geq 0$ if $\tilde{t} - t \geq 0$ and $\varphi(x, \xi, t) - \varphi(x, \xi, \tilde{t}) \leq 0$ of $t - \tilde{t} \leq 0$, giving $(\varphi(x, \xi, t) - \varphi(x, \xi, \tilde{t}))(t - \tilde{t}) \geq 0$ for all $t, \tilde{t} \in \mathbb{R}$ and a. e. $x, \xi$. Hence, we get

$$\langle \boldsymbol{N}(\boldsymbol{y}) - \boldsymbol{N}(\tilde{\boldsymbol{y}}), \boldsymbol{y} - \tilde{\boldsymbol{y}} \rangle_{\boldsymbol{Y}^*, \boldsymbol{Y}} =$$
$$\int_\Xi \int_\Omega (\varphi(x, \xi, \boldsymbol{y}(x, \xi)) - \varphi(x, \xi, \tilde{\boldsymbol{y}}(x, \xi)))(\boldsymbol{y}(x, \xi) - \tilde{\boldsymbol{y}}(x, \xi)) \, \mathrm{d}x \, \mathrm{d}\mathbb{P} \geq 0$$

for all $\boldsymbol{y}, \tilde{\boldsymbol{y}} \in \boldsymbol{Y}$, showing the monotonicity of $\boldsymbol{N}$. □

**Proposition A.14.** *Under Assumption A.11:1-2,4-5, the operator $\boldsymbol{N} : \boldsymbol{Y} \to \boldsymbol{Y}^*$ is continuously Fréchet-differentiable with derivative*

$$[\boldsymbol{N}'(\boldsymbol{y})\boldsymbol{v}](x,\xi) = \varphi_t(x,\xi,\boldsymbol{y}(x,\xi))\boldsymbol{v}(x,\xi).$$

*Proof.* We have that $\boldsymbol{N}$ is continuously Fréchet-differentiable with the given derivative as an operator from $L^p_{\lambda\otimes\mathbb{P}}(\Omega\times\Xi)$ to $L^{p^*}_{\lambda\otimes\mathbb{P}}(\Omega\times\Xi)$ by Theorem A.9, having $q = p^* = \frac{p}{p-1}$, and thus $\frac{pq}{p-q} = \frac{p}{p-2}$ and $\frac{p-q}{q} = p-2$. Note that we require $p > p^*$ which is fulfilled if and only if $p > 2$. Again, the continuous embeddings $\boldsymbol{Y} \hookrightarrow L^p_{\lambda\otimes\mathbb{P}}(\Omega\times\Xi)$ and $L^{p^*}_{\lambda\otimes\mathbb{P}}(\Omega\times\Xi) \hookrightarrow \boldsymbol{Y}^*$ and the chain rule yield continuous F-differentiability from $\boldsymbol{Y}$ to $\boldsymbol{Y}^*$. □

**Remark A.15.** For $p = 2$ we can consider the subspace of $L^2_{\mathbb{P}}(\Xi;Y)$ of functions of the form $\boldsymbol{y}(x,\xi) = y(x)\upsilon(\xi)$ with some fixed $y \in Y$ with $\|y\|_Y = 1$ and arbitrary $\upsilon \in L^2_{\mathbb{P}}(\Xi)$. This subspace is isomorphic to $L^2_{\mathbb{P}}(\Xi)$. [5, Thm. 3.13] states that if a Carathéodory function $\varphi : \Xi\times\mathbb{R}\to\mathbb{R}$ generates a well-defined, F-differentiable superposition operator acting from $L^2_{\mathbb{P}}(\Xi)$ into itself, the function $\varphi$ has the form $\varphi(\xi,t) = a(\xi) + b(\xi)t$ for a. e. $\xi \in \Xi_c$ with $a \in L^2_{\mathbb{P}}(\Xi)$ and $b \in L^\infty_{\mathbb{P}}(\Xi)$. $\Xi_c \subset \Xi$ is the non-atomic part of the probability measure $\mathbb{P}$. Hence, we need to require $p > 2$ if we want to work with "true" nonlinearities.

**Proposition A.16.** *Under Assumption A.11:1-2,5-7, the operator $\boldsymbol{N} : \boldsymbol{Y} \to \boldsymbol{Y}^*$ is twice continuously Fréchet-differentiable with second derivative*

$$[[\boldsymbol{N}''(\boldsymbol{y})\boldsymbol{v}]\tilde{\boldsymbol{v}}](x,\xi) = \varphi_{tt}(x,\xi,\boldsymbol{y}(x,\xi))\boldsymbol{v}(x,\xi)\tilde{\boldsymbol{v}}(x,\xi).$$

*Proof.* It follows from Proposition A.14 that $\boldsymbol{N}$ is continuously F-differentiable. We show that $N'$ is again continuously F-differentiable with the given derivative. For this purpose we identify $\mathcal{L}(L^p, L^{p^*}) \cong L^{p/(p-2)}$ and consider the continuously F-differentiable superposition operator $\boldsymbol{N}_{\varphi_t} : L^p_{\lambda\otimes\mathbb{P}}(\Omega\times\Xi) \to L^{p/(p-2)}_{\lambda\otimes\mathbb{P}}(\Omega\times\Xi)$ induced by the Carathéodory function $\varphi_t : \Omega\times\Xi\times\mathbb{R}\to\mathbb{R}$, cf. the proof of Theorem A.9. It is well-defined and continuous, see Proposition A.12. Assumption A.11:6-7 and Theorem A.9 yield that it is even continuously F-differentiable. There we use $q = \frac{p}{p-2}$ and thus $p > q$ if and only if $p > 3$, $\frac{p}{q} = p-2$, $\frac{pq}{p-q} = \frac{p}{p-3}$, and $\frac{p-q}{q} = p-3$. Its derivative is $[\boldsymbol{N}'_{\varphi_t}(\boldsymbol{y})\boldsymbol{v}](x,\xi) = \varphi_{tt}(x,\xi,\boldsymbol{y}(x,\xi))\boldsymbol{v}(x,\xi)$. It is identified with the derivative given above. □

## A.5. Alternative Approach for the Discussion of the Semilinear, Elliptic PDE

We present an alternative approach for the existence of a unique solution of the PDE discussed in Section 3.2 for the more concrete case $\varphi(t) = \hat{\varphi}(x,\xi,t) = t^3$ and the choice $p = 4$. Recall that the state space is $\boldsymbol{Y} = L^4_{\mathbb{P}}(\Xi; H^1_0(\Omega))$. Since $\hat{\varphi}$ fulfills Assumption A.11 with $c_{p*} \equiv 0$, $c_\varphi = 1$, $c'_{p/(p-2)} \equiv 0$, $c'_{\varphi_t} = 3$, $c''_{p/(p-3)} \equiv 0$, $c''_{\varphi_{tt}} = 6$, the superposition operator $\boldsymbol{N} : \boldsymbol{Y} \to \boldsymbol{Y}^*$ is twice continuously differentiable and monotone by the considerations in Section A.4. It follows that the operator $\tilde{\boldsymbol{N}} \equiv \boldsymbol{A} + \boldsymbol{N} : \boldsymbol{Y} \to \boldsymbol{Y}^*$ is a strictly monotone and twice continuously differentiable.

Unfortunately, the operator $\tilde{\boldsymbol{N}}$ is not coercive in $\boldsymbol{Y}$ in the sense of Theorem 3.7: For a counterexample we choose $\Omega := (0,1)^2 \subset \mathbb{R}^2$ and $\Xi := (0,1) \subset \mathbb{R}$ with the uniform distribution. For $x \in \Omega$ we define

$$y_\varepsilon(x) := \begin{cases} \varepsilon^q x_1, & x_1 \le x_2 \le 2\varepsilon - x_1, \\ \varepsilon^q x_2, & x_2 < x_1 \text{ and } x_2 \le 2\varepsilon - x_1, \\ -\varepsilon^q x_1 + 2\varepsilon^{q+1}, & 2\varepsilon - x_1 < x_2 \le x_1 \text{ and } x_1 < 2\varepsilon, \\ -\varepsilon^q x_2 + 2\varepsilon^{q+1}, & x_1 < x_2 < 2\varepsilon \text{ and } 2\varepsilon - x_1 < x_2, \\ 0, & \text{otherwise} \end{cases}$$

for $\varepsilon \in (0, \frac{1}{2})$ and $q \in (-\frac{5}{4}, -\frac{3}{4})$, e. g., $q = -1$. Note that $y_\varepsilon$ is continuous and belongs to $H_0^1(\Omega)$. In fact it is a multiple of a finite element ansatz function defined on triangles of equal area $\varepsilon^2$. The weak derivative is, up to sets of measure zero,

$$\nabla y_\varepsilon(x) := \begin{cases} (\varepsilon^q, 0)^\top, & x_1 \le x_2 \le 2\varepsilon - x_1, \\ (0, \varepsilon^q)^\top, & x_2 < x_1 \text{ and } x_2 \le 2\varepsilon - x_1, \\ (-\varepsilon^q, 0)^\top, & 2\varepsilon - x_1 < x_2 \le x_1 \text{ and } x_1 < 2\varepsilon, \\ (0, -\varepsilon^q)^\top, & x_1 < x_2 < 2\varepsilon \text{ and } 2\varepsilon - x_1 < x_2, \\ (0,0)^\top, & \text{otherwise.} \end{cases}$$

This yields

$$\|y_\varepsilon\|_{H_0^1(\Omega)}^2 = \int_\Omega \|\nabla y_\varepsilon(x)\|_{L^2(\Omega)^2}^2 \,\mathrm{d}x = 4\varepsilon^{2q+2}$$

and

$$\begin{aligned} \|y_\varepsilon\|_{L^4(\Omega)}^4 &= \int_\Omega y_\varepsilon(x)^4 \,\mathrm{d}x = 4 \int_0^\varepsilon \int_{x_1}^{2\varepsilon - x_1} \varepsilon^{4q} x_1^4 \,\mathrm{d}x_2 \,\mathrm{d}x_1 \\ &= 8\varepsilon^{4q} \int_0^\varepsilon x_1^4 (\varepsilon - x_1) \,\mathrm{d}x_1 = \tfrac{4}{15}\varepsilon^{4q+6}. \end{aligned}$$

Furthermore, we define $\upsilon_\varepsilon(\xi) := \xi^{-1/2+\varepsilon}$ and get $\|\upsilon_\varepsilon\|_{L_{\mathbb{P}}^1(\Xi)} = \int_0^1 \xi^{-1/2+\varepsilon} \,\mathrm{d}\xi = \frac{2}{1+2\varepsilon}$ as well as $\|\upsilon_\varepsilon\|_{L_{\mathbb{P}}^2(\Xi)}^2 = \int_0^1 \xi^{-1+2\varepsilon} \,\mathrm{d}\xi = \frac{1}{2\varepsilon}$. For $\boldsymbol{y}_\varepsilon(x,\xi) := y_\varepsilon(x)\,\upsilon_\varepsilon(\xi)^{1/2}$ we obtain

$$\begin{aligned} \|\boldsymbol{y}_\varepsilon\|_{L_{\mathbb{P}}^2(\Xi;H_0^1(\Omega))}^2 &= \int_0^1 \upsilon_\varepsilon(\xi) \|y_\varepsilon\|_{H_0^1(\Omega)}^2 \,\mathrm{d}\xi = \tfrac{8\varepsilon^{2q+2}}{1+2\varepsilon}, \\ \|\boldsymbol{y}_\varepsilon\|_{L_{\mathbb{P}}^4(\Xi;H_0^1(\Omega))}^4 &= \int_0^1 \upsilon_\varepsilon(\xi)^2 \|y_\varepsilon\|_{H_0^1(\Omega)}^4 \,\mathrm{d}\xi = 8\varepsilon^{4q+3}, \\ \|\boldsymbol{y}_\varepsilon\|_{L_{\mathbb{P}}^4(\Xi;L^4(\Omega))}^4 &= \int_0^1 \upsilon_\varepsilon(\xi)^2 \|y_\varepsilon\|_{L^4(\Omega)}^4 \,\mathrm{d}\xi = \tfrac{2}{15}\varepsilon^{4q+5}. \end{aligned}$$

This gives

$$\begin{aligned} \frac{\langle \boldsymbol{A}\boldsymbol{y}_\varepsilon + \boldsymbol{N}(\boldsymbol{y}_\varepsilon), \boldsymbol{y}_\varepsilon \rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}}}{\|\boldsymbol{y}_\varepsilon\|_{\boldsymbol{Y}}} &\le \frac{\overline{\kappa}\|\boldsymbol{y}_\varepsilon\|_{L_{\mathbb{P}}^2(\Xi;H_0^1(\Omega))}^2 + \|\boldsymbol{y}_\varepsilon\|_{L_{\mathbb{P}}^4(\Xi;L^4(\Omega))}^4}{\|\boldsymbol{y}_\varepsilon\|_{L_{\mathbb{P}}^4(\Xi;H_0^1(\Omega))}} \\ &= \frac{\overline{\kappa}\frac{8\varepsilon^{2q+2}}{1+2\varepsilon} + \frac{2}{15}\varepsilon^{4q+5}}{8^{1/4}\varepsilon^{q+3/4}} = \tfrac{8^{3/4}\overline{\kappa}}{1+2\varepsilon}\varepsilon^{q+5/4} + \tfrac{2^{1/4}}{15}\varepsilon^{3q+17/4}. \end{aligned}$$

By $q+\frac{3}{4}<0$ we have $\|\boldsymbol{y}_\varepsilon\|_{L^4_{\mathbb{P}}(\Xi;H^1_0(\Omega))} = 8^{1/4}\varepsilon^{q+3/4} \overset{\varepsilon\to 0+}{\longrightarrow} \infty$, but

$$0 \le \frac{\langle \boldsymbol{A}\boldsymbol{y}_\varepsilon + \boldsymbol{N}(\boldsymbol{y}_\varepsilon), \boldsymbol{y}_\varepsilon\rangle_{\boldsymbol{Y}^*,\boldsymbol{Y}}}{\|\boldsymbol{y}_\varepsilon\|_{\boldsymbol{Y}}} \le \tfrac{8^{3/4}\overline{\kappa}}{1+2\varepsilon}\varepsilon^{q+5/4} + \tfrac{2^{1/4}}{15}\varepsilon^{3(q+17/12)} \overset{\varepsilon\to 0+}{\longrightarrow} 0$$

holds by $q+\frac{5}{4}>0$ and $q+\frac{17}{12} > q+\frac{15}{12} > 0$. Therefore, the operator $\tilde{\boldsymbol{N}}$ is *not* coercive on $\boldsymbol{Y}$.

To be able to apply Theorem 3.7 we show coercivity and the other prerequisites on a larger space $\tilde{\boldsymbol{Y}} \supset \boldsymbol{Y}$, namely $\tilde{\boldsymbol{Y}} := L^2_{\mathbb{P}}(\Xi; H^1_0(\Omega)) \cap L^4_{\mathbb{P}}(\Xi; L^4(\Omega))$, to infer that (3.12) has a unique solution in $\tilde{\boldsymbol{Y}}$. After that, we will see that this solution actually belongs to $\boldsymbol{Y}$. The space $\tilde{\boldsymbol{Y}}$ is equipped with the norm $\|\boldsymbol{y}\|_{\tilde{\boldsymbol{Y}}} := \|\boldsymbol{y}\|_{L^2_{\mathbb{P}}(\Xi;H^1_0(\Omega))} + \|\boldsymbol{y}\|_{L^4_{\mathbb{P}}(\Xi;L^4(\Omega))}$. By $L^4_{\mathbb{P}}(\Xi) \hookrightarrow L^2_{\mathbb{P}}(\Xi)$ (Proposition A.2) and $H^1_0(\Omega) \hookrightarrow L^4(\Omega)$ (Sobolev embedding) we have $\boldsymbol{Y} \subset \tilde{\boldsymbol{Y}}$.

The operator $\boldsymbol{A} : \hat{\boldsymbol{Y}} := L^2_{\mathbb{P}}(\Xi; H^1_0(\Omega)) \to L^2_{\mathbb{P}}(\Xi; H^{-1}(\Omega)) = \hat{\boldsymbol{Y}}^*$ is well-defined since $A \in L^\infty_{\mathbb{P}}(\Xi; \mathcal{L}(Y, Y^*))$, see Proposition 3.1, and thus also as an operator from $\tilde{\boldsymbol{Y}}$ to $\tilde{\boldsymbol{Y}}^*$: Since $\tilde{\boldsymbol{Y}} \subset \hat{\boldsymbol{Y}}$, meaning that $\|\boldsymbol{y}\|_{\hat{\boldsymbol{Y}}} \le c\|\boldsymbol{y}\|_{\tilde{\boldsymbol{Y}}}$ for all $\boldsymbol{y} \in \tilde{\boldsymbol{Y}}$ (here with constant $c=1$), we get

$$\|\boldsymbol{A}\|_{\mathcal{L}(\tilde{\boldsymbol{Y}},\tilde{\boldsymbol{Y}}^*)} = \sup_{\substack{\boldsymbol{y},\boldsymbol{v}\in\tilde{\boldsymbol{Y}},\\ \|\boldsymbol{y}\|_{\tilde{\boldsymbol{Y}}}\le 1, \|\boldsymbol{v}\|_{\tilde{\boldsymbol{Y}}}\le 1}} \langle \boldsymbol{A}\boldsymbol{y}, \boldsymbol{v}\rangle_{\tilde{\boldsymbol{Y}}^*,\tilde{\boldsymbol{Y}}} \le \sup_{\substack{\boldsymbol{y},\boldsymbol{v}\in\hat{\boldsymbol{Y}},\\ \|\boldsymbol{y}\|_{\hat{\boldsymbol{Y}}}\le c, \|\boldsymbol{v}\|_{\hat{\boldsymbol{Y}}}\le c}} \langle \boldsymbol{A}\boldsymbol{y}, \boldsymbol{v}\rangle_{\hat{\boldsymbol{Y}}^*,\hat{\boldsymbol{Y}}} = c^2\|\boldsymbol{A}\|_{\mathcal{L}(\hat{\boldsymbol{Y}},\hat{\boldsymbol{Y}}^*)}.$$

From the Theorem A.5 we have that $\boldsymbol{N} : L^4_{\mathbb{P}}(\Xi; L^4(\Omega)) \to L^{4/3}_{\mathbb{P}}(\Xi; L^{4/3}(\Omega))$ is well-defined and continuous and thus also well-defined and continuous as an operator from $\tilde{\boldsymbol{Y}}$ to $\tilde{\boldsymbol{Y}}^*$ by the same argument. This proves that $\tilde{\boldsymbol{N}} : \tilde{\boldsymbol{Y}} \to \tilde{\boldsymbol{Y}}^*$ is well-defined and continuous, in particular, hemicontinuous. Strict monotonicity of $\tilde{\boldsymbol{N}}$ is shown exactly as in Proposition A.13.

Now we prove coercivity. For $\|\boldsymbol{y}\|_{L^2_{\mathbb{P}}(\Xi;H^1_0(\Omega))} + \|\boldsymbol{y}\|_{L^4_{\mathbb{P}}(\Xi;L^4(\Omega))} \to \infty$ we have

$$\frac{\langle \boldsymbol{A}\boldsymbol{y} + \boldsymbol{N}(\boldsymbol{y}), \boldsymbol{y}\rangle_{\tilde{\boldsymbol{Y}}^*,\tilde{\boldsymbol{Y}}}}{\|\boldsymbol{y}\|_{\tilde{\boldsymbol{Y}}}} \ge \frac{\underline{\kappa}\|\boldsymbol{y}\|^2_{L^2_{\mathbb{P}}(\Xi;H^1_0(\Omega))} + \|\boldsymbol{y}\|^4_{L^4_{\mathbb{P}}(\Xi;L^4(\Omega))}}{\|\boldsymbol{y}\|_{L^2_{\mathbb{P}}(\Xi;H^1_0(\Omega))} + \|\boldsymbol{y}\|_{L^4_{\mathbb{P}}(\Xi;L^4(\Omega))}} \longrightarrow \infty.$$

This can be shown as follows: Let $(a_n)_{n\in\mathbb{N}} \subset \mathbb{R}_{\ge 0}$ and $(b_n)_{n\in\mathbb{N}} \subset \mathbb{R}_{\ge 0}$ be two sequences such that $a_n + b_n > 0$ for all $n$, $a_n + b_n \overset{n\to\infty}{\longrightarrow} \infty$ and let $c > 0$. Then, if $a_n \ge b_n$, giving $a_n > 0$, we have

$$d_n := \tfrac{ca_n^2+b_n^4}{a_n+b_n} \ge \tfrac{ca_n^2}{2a_n} = \tfrac{ca_n}{2} \ge \tfrac{c}{4}(a_n+b_n),$$

and if $a_n \le b_n$, giving $b_n > 0$, we have

$$d_n = \tfrac{ca_n^2+b_n^4}{a_n+b_n} \ge \tfrac{b_n^4}{2b_n} = \tfrac{b_n^3}{2} \ge \tfrac{1}{16}(a_n+b_n)^3.$$

Overall, this gives

$$d_n \ge \min\{\tfrac{c}{4}(a_n+b_n), \tfrac{1}{16}(a_n+b_n)^3\} \ge \min\{\tfrac{c}{4}, \tfrac{1}{16}\}(a_n+b_n)$$

as long as $a_n + b_n \ge 1$. The assumption $a_n + b_n \overset{n\to\infty}{\longrightarrow} \infty$ yields $d_n \overset{n\to\infty}{\longrightarrow} \infty$.

Overall, we can apply Theorem 3.7 to deduce that (3.12) has a unique solution $\boldsymbol{y} \in \tilde{\boldsymbol{Y}} = L^2_{\mathbb{P}}(\Xi; H^1_0(\Omega)) \cap L^4_{\mathbb{P}}(\Xi; L^4(\Omega))$. By constructing and analyzing $\boldsymbol{y}$ more explicitly as in Section 3.2, we see that this solution even belongs to $L^4_{\mathbb{P}}(\Xi; H^1_0(\Omega))$ if $r_f \ge 4$.

## A.6. Concrete Computations for Section 9.3

In Section 9.3 we use some concretely computed results, which are presented in the following.

### Computation of the Unique Stationary Point for Proposition 9.22

Let

$$\frac{\partial}{\partial w} f_\mu(x, \bar{w}, t) = 1 - \frac{\mu}{\bar{w} - a_1(x-t)} - \frac{\mu}{\bar{w} - a_2(x-t)} = 0, \quad \bar{w} > a_i(x-t) \quad (i \in \{1, 2\}).$$

For $s_1 := a_1(x-t)$, $s_2 := a_2(x-t)$, i. e., $\bar{w} > s_1$ and $\bar{w} > s_2$, we perform the following equivalent transformations:

$$\begin{aligned}
& 1 - \frac{\mu}{\bar{w} - s_1} - \frac{\mu}{\bar{w} - s_2} = 0, && \bar{w} > s_1, \bar{w} > s_2 \\
\Leftrightarrow\ & (\bar{w} - s_1)(\bar{w} - s_2) - \mu(\bar{w} - s_2) - \mu(\bar{w} - s_1) = 0, && \bar{w} > s_1, \bar{w} > s_2 \\
\Leftrightarrow\ & \bar{w}^2 - (s_1 + s_2 + 2\mu)\bar{w} + s_1 s_2 + \mu(s_1 + s_2) = 0, && \bar{w} > s_1, \bar{w} > s_2 \\
\Leftrightarrow\ & \bar{w} = \frac{s_1 + s_2 + 2\mu \pm \sqrt{(s_1 + s_2 + 2\mu)^2 - 4(s_1 s_2 + \mu(s_1 + s_2))}}{2}, && \bar{w} > s_1, \bar{w} > s_2 \\
\Leftrightarrow\ & \bar{w} = \mu + \frac{s_1 + s_2 + \sqrt{(s_1 - s_2)^2 + 4\mu^2}}{2}.
\end{aligned}$$

In the last expression, taking "$-$" in front of the square root would result in $\bar{w} - s_2 = \frac{s_1 - s_2 + 2\mu - \sqrt{(s_1 - s_2)^2 + 4\mu^2}}{2} \leq \frac{s_1 - s_2}{2}$ and in the same manner $\bar{w} - s_1 \leq \frac{s_2 - s_1}{2}$. This gives that one of the two conditions $\bar{w} > s_1$ and $\bar{w} > s_2$ would be violated then which is why it is necessary to take "$+$".

### Computation of the Derivatives of $v_\mu$

We compute the derivatives of the function $v_\mu$, needed in the proof of Theorem 9.23. Recall that

$$v_\mu(s) = w_\mu(s) - \mu \ln(w_\mu(s) - a_1 s) - \mu \ln(w_\mu(s) - a_2 s) + \zeta(\mu)$$

holds with $w_\mu(s) = \mu + \frac{(a_1 + a_2)s + \sqrt{(a_1 - a_2)^2 s^2 + 4\mu^2}}{2}$. Clearly, by the chain rule,

$$v'_\mu(s) = w'_\mu(s) - \mu \frac{w'_\mu(s) - a_1}{w_\mu(s) - a_1 s} - \mu \frac{w'_\mu(s) - a_2}{w_\mu(s) - a_2 s}$$

with $w'_\mu(s) = \frac{a_1 + a_2}{2} + \frac{(a_1 - a_2)^2 s}{2\sqrt{(a_1 - a_2)^2 s^2 + 4\mu^2}}$. Indeed, the function $w_\mu$ is twice continuously differentiable on $\mathbb{R}$ with the second derivative

$$\begin{aligned}
w''_\mu(s) &= \frac{2\sqrt{(a_1 - a_2)^2 s^2 + 4\mu^2}(a_1 - a_2)^2 - (a_1 - a_2)^2 s \frac{2(a_1 - a_2)^2 s}{\sqrt{(a_1 - a_2)^2 s^2 + 4\mu^2}}}{4((a_1 - a_2)^2 s^2 + 4\mu^2)} \\
&= \frac{\frac{4\mu^2 (a_1 - a_2)^2}{\sqrt{(a_1 - a_2)^2 s^2 + 4\mu^2}}}{2(a_1 - a_2)^2 s^2 + 8\mu^2} = \frac{2\mu^2 (a_1 - a_2)^2}{((a_1 - a_2)^2 s^2 + 4\mu^2)^{3/2}} > 0 \quad \forall\, s \in \mathbb{R}.
\end{aligned}$$

With this expression, the second derivative of $v_\mu$ can be computed defining $\beta := a_2 - a_1 > 0$ and $\gamma(s) := \sqrt{\beta^2 s^2 + 4\mu^2} \geq 2\mu$:

$$\begin{aligned}
\tilde{v}_\mu''(s) &= w_\mu''(s) - \mu \frac{(w_\mu(s)-a_1 s)w_\mu''(s)-(w_\mu'(s)-a_1)^2}{(w_\mu(s)-a_1 s)^2} - \mu \frac{(w_\mu(s)-a_2 s)w_\mu''(s)-(w_\mu'(s)-a_2)^2}{(w_\mu(s)-a_2 s)^2} \\
&= \frac{2\mu^2\beta^2}{\gamma^3(s)} - \mu \frac{(\mu+\frac{\beta s}{2}+\frac{\gamma(s)}{2})\frac{2\mu^2\beta^2}{\gamma^3(s)}-(\frac{\beta}{2}+\frac{\beta^2 s}{2\gamma(s)})^2}{(\mu+\frac{\beta s}{2}+\frac{\gamma(s)}{2})^2} - \mu \frac{(\mu-\frac{\beta s}{2}+\frac{\gamma(s)}{2})\frac{2\mu^2\beta^2}{\gamma^3(s)}-(-\frac{\beta}{2}+\frac{\beta^2 s}{2\gamma(s)})^2}{(\mu-\frac{\beta s}{2}+\frac{\gamma(s)}{2})^2} \\
&= \frac{\mu\beta^2}{\gamma^3(s)(\mu+\frac{\beta s}{2}+\frac{\gamma(s)}{2})^2(\mu-\frac{\beta s}{2}+\frac{\gamma(s)}{2})^2} \Big(2\mu(\mu+\tfrac{\beta s}{2}+\tfrac{\gamma(s)}{2})^2(\mu-\tfrac{\beta s}{2}+\tfrac{\gamma(s)}{2})^2 \\
&\quad - 2\mu^2(\mu+\tfrac{\beta s}{2}+\tfrac{\gamma(s)}{2})(\mu-\tfrac{\beta s}{2}+\tfrac{\gamma(s)}{2})^2 + \gamma(s)(\tfrac{\gamma(s)}{2}+\tfrac{\beta s}{2})^2(\mu-\tfrac{\beta s}{2}+\tfrac{\gamma(s)}{2})^2 \\
&\quad - 2\mu^2(\mu-\tfrac{\beta s}{2}+\tfrac{\gamma(s)}{2})(\mu+\tfrac{\beta s}{2}+\tfrac{\gamma(s)}{2})^2 + \gamma(s)(-\tfrac{\gamma(s)}{2}+\tfrac{\beta s}{2})^2(\mu+\tfrac{\beta s}{2}+\tfrac{\gamma(s)}{2})^2\Big).
\end{aligned}$$

The first factor is always positive and well-defined by $\frac{\pm\beta s+\gamma(s)}{2} > \frac{\pm\beta s+\beta|s|}{2} \geq 0$. Defining $\alpha_+(s) := \frac{\beta s+\gamma(s)}{2} > 0$ and $\alpha_-(s) := \frac{-\beta s+\gamma(s)}{2} > 0$, the second factor can be transformed to

$$\begin{aligned}
&2\mu(\mu+\alpha_+(s))^2(\mu+\alpha_-(s))^2 - 2\mu^2[(\mu+\alpha_+(s))(\mu+\alpha_-(s))^2 + (\mu+\alpha_-(s))(\mu+\alpha_+(s))^2] \\
&\quad + \gamma(s)[\alpha_+(s)^2(\mu+\alpha_-(s))^2 + \alpha_-(s)^2(\mu+\alpha_+(s))^2] \\
&= 2\mu(\mu+\alpha_+(s))(\mu+\alpha_-(s))\left[(\mu+\alpha_+(s))(\mu+\alpha_-(s)) - \mu(\mu+\alpha_-(s)) - \mu(\mu+\alpha_+(s))\right] \\
&\quad + \gamma(s)[\mu^2(\beta^2 s^2 + 2\mu^2) + 2\mu^3\gamma(s) + 2\mu^4] \\
&= 2\mu(\mu+\alpha_+(s))(\mu+\alpha_-(s))\underbrace{\left[\alpha_+(s)(\mu+\alpha_-(s)) - \mu^2 - \mu\alpha_+(s)\right]}_{=0} \\
&\quad + \gamma(s)\mu^2[\beta^2 s^2 + 2\mu\gamma(s) + 4\mu^2] \\
&= \gamma(s)\mu^2[\gamma(s)^2 + 2\mu\gamma(s)] = \gamma(s)^2\mu^2(\gamma(s)+2\mu) \geq 16\mu^5 > 0
\end{aligned}$$

using $\alpha_+(s)+\alpha_-(s) = \gamma(s)$, $\alpha_+(s)^2+\alpha_-(s)^2 = \beta^2 s^2 + 2\mu^2$ and $\alpha_+(s)\alpha_-(s) = \mu^2$. We obtain a simplified version of the second derivative with $(\mu+\alpha_+(s))^2(\mu+\alpha_-(s))^2 = \mu^2(2\mu+\gamma(s))^2$:

$$v_\mu''(s) = \frac{\mu\beta^2}{\gamma(s)(2\mu+\gamma(s))} = \frac{\mu(a_2-a_1)^2}{2\mu\sqrt{(a_2-a_1)^2 s^2 + 4\mu^2} + (a_2-a_1)^2 s^2 + 4\mu^2}.$$

### Computation of Limits

For the proof of Theorem 9.23, we compute the limits $\lim_{s\to-\infty} v_\mu'(s)$ and $\lim_{s\to+\infty} v_\mu'(s)$. First, observe that

$$\lim_{s\to-\infty} w_\mu'(s) = \frac{a_1+a_2}{2} - \frac{|a_2-a_1|}{2} = a_1 \text{ and } \lim_{s\to+\infty} w_\mu'(s) = \frac{a_1+a_2}{2} + \frac{|a_2-a_1|}{2} = a_2$$

holds ($a_2 > a_1$). Moreover,

$$\frac{w_\mu'(s)-a_1}{w_\mu(s)-a_1 s} = \frac{\frac{a_2-a_1}{2} + \frac{(a_2-a_1)^2 s}{2\sqrt{(a_2-a_1)^2 s^2+4\mu^2}}}{\mu + \frac{(a_2-a_1)s}{2} + \frac{\sqrt{(a_2-a_1)^2 s^2+4\mu^2}}{2}} \longrightarrow 0 \quad \text{as } s \to \pm\infty.$$

As $s \to +\infty$, the numerator tends to $a_2 - a_1$, but the denominator goes to $+\infty$. For $s \to -\infty$, the numerator becomes 0 and the denominator tends to $\mu$. An analogous argumentation for the term $\frac{w'_\mu(s)-a_2}{w_\mu(s)-a_2 s}$ yields the limits

$$\lim_{s\to-\infty} v'_\mu(s) = a_1 \text{ and } \lim_{s\to+\infty} v'_\mu(s) = a_2.$$

## Computation of $\hat{v}_\mu(0)$ and $\hat{v}'_\mu(0)$ for Theorem 9.23

In the proof of Theorem 9.23, $d := \frac{2-a_1-a_2}{(1-a_1)(a_2-1)}\mu = \frac{1}{a_2-1}\mu - \frac{1}{1-a_1}\mu$ and $\hat{v}_\mu(s) = v_\mu(s+d) - d$ are defined. We compute

$$\begin{aligned}
\hat{v}_\mu(0) &= v_\mu(d) - d = w_\mu(d) - \mu \ln(w_\mu(d) - a_1 d) - \mu \ln(w_\mu(d) - a_2 d) + \zeta(\mu) - d \\
&= \tfrac{a_2}{a_2-1}\mu - \tfrac{a_1}{1-a_1}\mu - \mu\left(\ln(\tfrac{a_2-a_1}{a_2-1}\mu) + \ln(\tfrac{a_2-a_1}{1-a_1}\mu)\right) \\
&\quad + \mu\left(\ln(\tfrac{a_2-a_1}{a_2-1}\mu) + \ln(\tfrac{a_2-a_1}{1-a_1}\mu) - 2\right) - \tfrac{1}{a_2-1}\mu + \tfrac{1}{1-a_1}\mu \\
&= \tfrac{a_2}{a_2-1}\mu - \tfrac{a_1}{1-a_1}\mu - 2\mu - \tfrac{1}{a_2-1}\mu + \tfrac{1}{1-a_1}\mu = 0
\end{aligned}$$

using

$$\begin{aligned}
\tfrac{2}{\mu} w_\mu(d) &= 2 + \tfrac{(a_1+a_2)d + \sqrt{(a_1-a_2)^2 d^2 + 4\mu^2}}{\mu} \\
&= 2 + \tfrac{(a_1+a_2)(2-a_1-a_2)}{(1-a_1)(a_2-1)} + \sqrt{\tfrac{(a_1-a_2)^2(2-a_1-a_2)^2}{(1-a_1)^2(a_2-1)^2} + 4} \\
&= \tfrac{2(1-a_1)(a_2-1)+(a_1+a_2)(2-a_1-a_2)}{(1-a_1)(a_2-1)} + \tfrac{\sqrt{(a_2-a_1)^2(2-a_1-a_2)^2+4(1-a_1)^2(a_2-1)^2}}{(1-a_1)(a_2-1)} \\
&= \tfrac{2(1-a_1)(a_2-1+a_1+a_2)+(a_2-1)(1-a_1-a_1-a_2)}{(1-a_1)(a_2-1)} \\
&\quad + \tfrac{\sqrt{((1-a_1)+(a_2-1))^2((1-a_1)-(a_2-1))^2+4(1-a_1)^2(a_2-1)^2}}{(1-a_1)(a_2-1)} \\
&= \tfrac{-(1-a_1)^2+2a_2(1-a_1)-(a_2-1)^2-2a_1(a_2-1)+\sqrt{((1-a_1)^2+(a_2-1)^2)^2}}{(1-a_1)(a_2-1)} \\
&= \tfrac{2a_2(1-a_1)-2a_1(a_2-1)}{(1-a_1)(a_2-1)} = \tfrac{2a_2}{a_2-1} - \tfrac{2a_1}{1-a_1}
\end{aligned}$$

and $w_\mu(d) - a_i d = \left(\frac{a_2-a_i}{a_2-1} - \frac{a_1-a_i}{1-a_1}\right)\mu$. We see that the shift $\zeta(\mu)$ is defined such that the shifted function $\hat{v}_\mu$ has value 0 at $s = 0$.

Furthermore,

$$\begin{aligned}
\hat{v}'_\mu(0) = v'_\mu(d) &= w'_\mu(d) - \mu\,\frac{w'_\mu(d) - a_1}{w_\mu(d) - a_1 d} - \mu\,\frac{w'_\mu(d) - a_2}{w_\mu(d) - a_2 d} \\
&= \tfrac{a_2}{(a_2-1)^2} + \tfrac{a_1}{(1-a_1)^2} - \mu\,\frac{\frac{a_2}{(a_2-1)^2} + \frac{a_1}{(1-a_1)^2} - a_1}{\frac{a_2-a_1}{a_2-1}\mu} - \mu\,\frac{\frac{a_2}{(a_2-1)^2} + \frac{a_1}{(1-a_1)^2} - a_2}{\frac{a_2-a_1}{1-a_1}\mu} \\
&= \left(\tfrac{a_2}{(a_2-1)^2} + \tfrac{a_1}{(1-a_1)^2}\right)\left(\tfrac{a_2-a_1-(a_2-1)-(1-a_1)}{a_2-a_1}\right) + \tfrac{a_1(a_2-1)}{a_2-a_1} + \tfrac{a_2(1-a_1)}{a_2-a_1} = 1
\end{aligned}$$

is computed using

$$\begin{aligned} w'_\mu(d) &= \tfrac{a_1+a_2}{2} + \frac{(a_1-a_2)^2 d}{2\sqrt{(a_1-a_2)^2 d^2+4\mu^2}} = \tfrac{a_1+a_2}{2} + \frac{\frac{(a_2-a_1)^2(2-a_1-a_2)}{(1-a_1)(a_2-1)}}{2\frac{(1-a_1)^2+(a_2-1)^2}{(1-a_1)(a_2-1)}} \\ &= \tfrac{a_1+a_2}{2} + \tfrac{(a_2-a_1)^2((1-a_1)-(a_2-1))}{2((1-a_1)^2+(a_2-1)^2)} = \tfrac{a_1+a_2}{2} + \tfrac{(a_2-a_1)((1-a_1)^2-(a_2-1)^2)}{2((1-a_1)^2+(a_2-1)^2)} \\ &= \tfrac{a_2(1-a_1)^2+a_1(a_2-1)^2}{(1-a_1)^2+(a_2-1)^2} = \tfrac{a_2}{(a_2-1)^2} + \tfrac{a_1}{(1-a_1)^2}. \end{aligned}$$

Now, we see what has been the idea behind the definition of the shift $\zeta(\mu)$ and the value of $d$: First, $d \in \mathbb{R}$ is chosen as the unique point where $v'_\mu(d) = 1$ holds. Then, the shift $\zeta(\mu)$ is defined such that $\mathcal{R}_\mu[0] = 0$ holds, which is provided by $\hat{v}_\mu(0) = 0$ and $\hat{v}'_\mu(0) = 1$.

# Bibliography


[1] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, 2008.

[2] M. Ainsworth and J. T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, Pure and Applied Mathematics, John Wiley & Sons, New York, 2000.

[3] A. Alexanderian, N. Petra, G. Stadler, and O. Ghattas, *Mean-Variance Risk-Averse Optimal Control of Systems Governed by PDEs with Random Parameter Fields Using Quadratic Approximations*, SIAM/ASA J. Uncertain. Quantif., 5 (2017), pp. 1166–1192.

[4] A. Ali, E. Ullmann, and M. Hinze, *Multilevel Monte Carlo Analysis for Optimal Control of Elliptic PDEs with Random Coefficients*, SIAM/ASA J. Uncertain. Quantif., 5 (2017), pp. 466–492.

[5] J. Appell and P. Zabreĭko, *Nonlinear Superposition Operators*, vol. 95 of Cambridge Tracts in Mathematics, Cambrige University Press, Cambridge, 1990.

[6] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, *Coherent Measures of Risk*, Math. Finance, 9 (1999), pp. 203–228.

[7] I. Babuška, F. Nobile, and R. Tempone, *A stochastic collocation method for elliptic partial differential equations with random input data*, SIAM Rev., 52 (2010).

[8] M. Bachmayr, A. Cohen, and W. Dahmen, *Parametric PDEs: sparse or low-rank approximations?*, IMA J. Numer. Anal., (2017).

[9] M. Bachmayr and W. Dahmen, *Adaptive near-optimal rank tensor approximation for high-dimensional operator equations*, Found. Comput. Math., 15 (2015), pp. 839–898.

[10] ——, *Adaptive low-rank methods for problems on Sobolev spaces with error control in $L_2$*, ESAIM Math. Model. Numer. Anal., 50 (2016), pp. 1107–1136.

[11] M. Bachmayr and R. Schneider, *Iterative Methods Based on Soft Thresholding of Hierarchical Tensors*, Found. Comput. Math., 17 (2017), pp. 1037–1083.

[12] M. Bachmayr, R. Schneider, and A. Uschmajew, *Tensor Networks and Hierarchical Tensors for the Solution of High-Dimensional Partial Differential Equations*, Found. Comput. Math., 16 (2016), pp. 1423–1472.

[13] J. Ballani and L. Grasedyck, *A projection method to solve linear systems in tensor format*, Numer. Linear Algebra Appl., 20 (2013), pp. 27–43.

[14] ——, *Hierarchical tensor approximation of output quantities of parameter-dependent PDEs*, SIAM/ASA J. Uncertain. Quantif., 3 (2015), pp. 852–872.

[15] J. Ballani, L. Grasedyck, and M. Kluge, *Black box approximation of tensors in hierarchical tucker format*, Linear Algebra Appl., 438 (2013), pp. 639–657.

[16] A. Barth, C. Schwab, and N. Zollinger, *Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients*, Numer. Math., 119 (2011), pp. 123–161.

[17] H. Bauschke and P. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, New York, 2011.

[18] R. Becker and R. Rannacher, *An optimal control approach to a posteriori error estimation in finite element methods*, Acta Numer., 10 (2001), pp. 1–102.

[19] S. Bellavia, *Inexact interior-point method*, J. Optim. Theory Appl., 96 (1998), pp. 109–121.

[20] P. Benner, S. Dolgov, A. Onwunta, and M. Stoll, *Low-rank solvers for unsteady Stokes–Brinkman optimal control problem with random data*, Comput. Methods Appl. Mech. Engrg., 304 (2016), pp. 26–54.

[21] P. Benner, A. Onwunta, and M. Stoll, *Block-Diagonal Preconditioning for Optimal Control Problems Constrained by PDEs with Uncertain Inputs*, SIAM J. Matrix Anal. Appl., 37 (2013), pp. 491–518.

[22] A. Bespalov, C. E. Powell, and D. Silvester, *Energy norm a posteriori error estimation for parametric operator equations*, SIAM J. Sci. Comput., 36 (2014), pp. A339 – A363.

[23] A. Bespalov and D. Silvester, *Efficient adaptive stochastic Galerkin methods for parametric operator equations*, SIAM J. Sci. Comput., 38 (2016), pp. A2118 – A2140.

[24] A. Borzì, V. Schulz, C. Schillings, and G. von Winckel, *On the treatment of distributed uncertainties in PDE-constrained optimization*, GAMM-Mitt., 33 (2010), pp. 230–246.

[25] A. Borzì and G. von Winckel, *A POD framework to determine robust controls in PDE optimization*, Comput. Vis. Sci., 14 (2011), pp. 91–103.

[26] R. G. Carter, *Numerical optimization in hilbert space using inexact function and gradient evaluations*, Technical report 89-45, ICASE, Langley, VA, 1989.

[27] ——, *On the Global Convergence of Trust Region Algorithms Using Inexact Gradient Information*, SIAM J. Numer. Anal., 28 (1991), pp. 251–265.

[28] P. Chen and A. Quarteroni, *Weighted Reduced Basis Method for Stochastic Optimal Control Problems with Elliptic PDE Constraint*, SIAM/ASA J. Uncertain. Quantif., 2 (2014), pp. 364–396.

[29] P. Chen, A. Quarteroni, and G. Rozza, *A Weighted Reduced Basis Method for Elliptic Partial Differential Equations with Random Input Data*, SIAM J. Numer. Anal., 51 (2013), pp. 3163–3185.

[30] K. Cliffe, M. Giles, R. Scheichl, and A. Teckentrup, *Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients*, Comput. Vis. Sci., 14 (2011), pp. 3–15.

[31] A. Conn, N. Gould, and P. Toint, *Trust-Region Methods*, SIAM and MOS, Philadelphia, 2000.

[32] C. Da Silva and F. Herrmann, *Optimization on the hierarchical Tucker manifold – Applications to tensor completion*, Linear Algebra Appl., 481 (2015), pp. 131–173.

[33] V. de Silva and L.-H. Lim, *Tensor rank and the ill-posedness of the best low-rank approximation problem*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 1084–1127.

[34] A. Defant and K. Floret, *Tensor Norms and Operator Ideals*, Elsevier Science Publishers, Amsterdam, 1993.

[35] S. V. Dolgov, *Alternating minimal energy approach to ODEs and conservation laws in tensor product formats.* preprint, arXiv:1403.8085, 2014.

[36] S. V. Dolgov and D. V. Savostyanov, *Alternating Minimal Energy Methods for Linear Systems in Higher Dimensions*, SIAM J. Sci. Comput., 36 (2014), pp. A2248–A2271.

[37] W. Dörfler, *A Convergent Adaptive Algorithm for Poisson's Equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

[38] M. Eigel, C. J. Gittelson, C. Schwab, and E. Zander, *Adaptive stochastic Galerkin FEM*, Comput. Methods Appl. Mech. Engrg., 270 (2014), pp. 247–269.

[39] ——, *A convergent adaptive stochastic Galerkin finite element method with quasi-optimal spatial meshes*, ESAIM Math. Model. Numer. Anal., 49 (2015), pp. 1367–1398.

[40] M. Eigel, M. Pfeffer, and R. Schneider, *Adaptive stochastic Galerkin FEM with hierarchical tensor representations*, Numer. Math., 136 (2017), pp. 765–803.

[41] S. C. Eisenstat and H. F. Walker, *Globally convergent inexact Newton methods*, SIAM J. Optim., 4 (1994), pp. 393–422.

[42] H. C. Elman and Q. Liao, *Reduced basis collocation methods for partial differential equations with random coefficients*, SIAM/ASA J. Uncertain. Quantif., 1 (2013), pp. 192–217.

[43] M. Espig, W. Hackbusch, A. Litvinenko, H. G. Matthies, and E. Zander, *Efficient analysis of high dimensional data in tensor formats*, in Sparse Grids and Applications, J. Garcke and M. Griebel, eds., Springer, Berlin, 2013, pp. 31–56.

[44] R. Forster and R. Kornhuber, *A polynomial chaos approach to stochastic variational inequalities*, J. Numer. Math., 18 (2010), pp. 235–255.

[45] J. Garcke, *Sparse grids in a nutshell*, in Sparse Grids and Applications, J. Garcke and M. Griebel, eds., Springer, Berlin, 2013, pp. 57–80.

[46] S. Garreis and M. Ulbrich, *Constrained Optimization with Low-Rank Tensors and Applications to Parametric Problems with PDEs*, SIAM J. Sci. Comput., 39 (2017), pp. A25–A54. The first version of this article has been acknowledged as Master's thesis for Sebastian Garreis within the Master's with Honours Examination for Mathematics at the Technical University of Munich.

[47] T. Gerstner and M. Griebel, *Numerical integration using sparse grids*, Numer. Algorithms, 18 (1998), pp. 209–232.

[48] H. Goldberg, W. Kampowsky, and F. Tröltzsch, *On* Nemytskij *Operators in $l_p$-Spaces of Abstract Functions*, Math. Nachr., 155 (1992), pp. 127–140.

[49] G. H. Golub and C. F. van Loan, *Matrix Computations*, Johns Hopkins Univ. Press, Baltimore and London, 1996.

[50] L. Grasedyck, *Hierarchical singular value decomposition of tensors*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2029–2054.

[51] L. Grasedyck, D. Kressner, and C. Tobler, *A literature survey of low-rank tensor approximation techniques*, GAMM-Mitt., 36 (2013), pp. 53–78.

[52] L. Grasedyck and C. Löbbert, *Distributed hierarchical SVD in the Hierarchical Tucker format*, Numer. Linear Algebra Appl., (2018), p. e2174.

[53] M. Gunzburger, C. Webster, and G. Zhang, *Stochastic finite element methods for partial differential equations with random input data*, Acta Numer., 23 (2014), pp. 521–650.

[54] W. Hackbusch, *Tensor Spaces and Numerical Tensor Calculus*, Springer, Heidelberg, 2012.

[55] W. Hackbusch, B. N. Khoromskij, and E. E. Tyrtyshnikov, *Approximate iterations for structured matrices*, Numer. Math., 109 (2008), pp. 365–383.

[56] W. Hackbusch and S. Kühn, *A new scheme for the tensor representation*, J. Fourier Anal. Appl., 15 (2009), pp. 706–722.

[57] M. Heinkenschloss and L. N. Vicente, *Analysis of inexact trust-region SQP algorithms*, SIAM J. Optim., 12 (2001), pp. 283–302.

[58] M. Hintermüller, K. Ito, and K. Kunisch, *The Primal-Dual Active Set Strategy as a Semismooth Newton Method*, SIAM J. Optim., 13 (2002), pp. 865–888.

[59] M. Hinze, *A Variational Discretization Concept in Control Constrained Optimization: The Linear-Quadratic Case*, Comput. Optim. Applic., 30 (2005), pp. 45–61.

[60] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE Constraints*, Springer, New York, 2009.

[61] F. L. Hitchcock, *The expression of a tensor or a polyadic as a sum of products*, J. Math. Phys., 6 (1927), pp. 164–189.

[62] S. Holtz, T. Rohwedder, and R. Schneider, *The alternating linear scheme for tensor optimization in the tensor train format*, SIAM J. Sci. Comput., 34 (2012), pp. A683–A713.

[63] L.S. Hou, J. Lee, and H. Manouzi, *Finite element approximations of stochastic optimal control problems constrained by stochastic elliptic PDEs*, J. Math. Anal. Appl., 384 (2011), pp. 87–103.

[64] T. Huckle, K. Waldherr, and T. Schulte-Herbrüggen, *Computations in Quantum Tensor Networks*, Linear Algebra Appl., 438 (2012), pp. 750–781.

[65] W. Huyer and A. Neumaier, *Global optimization by multilevel coordinate search*, J. Global Optim., 14 (1999), pp. 331–355.

[66] T. Hytönen, M. Veraar, J. van Neerven, and L. Weis, *Analysis in Banach Spaces*, vol. I, Springer, Cham, 2016.

[67] T. G. Kolda and B. W. Bader, *Tensor Decompositions and Applications*, SIAM Rev., 51 (2009), pp. 455–500.

[68] D. P. Kouri, *A Multilevel Stochastic Collocation Algorithm for Optimization of PDEs with Uncertain Coefficients*, SIAM/ASA J. Uncertain. Quantif., 2 (2014), pp. 55–81.

[69] D. P. Kouri, M. Heinkenschloss, D. Ridzal, and B. G. van Bloemen Waanders, *A Trust-Region Algorithm with Adaptive Stochastic Collocation for PDE Optimization under Uncertainty*, SIAM J. Sci. Comput., 35 (2013), pp. A1847–A1879.

[70] ——, *Inexact Objective Function Evaluations in a Trust-Region Algorithm for PDE-Constrained Optimization under Uncertainty*, SIAM J. Sci. Comput., 36 (2014), pp. A3011–A3029.

[71] D. P. Kouri and T. M. Surowiec, *Risk-Averse PDE-Constrained Optimization Using the Conditional Value-At-Risk*, SIAM J. Optim., 26 (2016), pp. 365–396.

[72] ——, *Existence and Optimality Conditions for Risk-Averse PDE-Constrained Optimization*, SIAM/ASA J. Uncertain. Quantif., 6 (2018), pp. 787–815.

[73] D. Kressner, M. Steinlechner, and B. Vandereycken, *Low-rank tensor completion by Riemannian optimization*, BIT, 54 (2014), pp. 447–468.

[74] ——, *Preconditioned Low-rank Riemannian Optimization for Linear Systems with Tensor Product Structure*, SIAM J. Sci. Comput., 38 (2016), pp. A2018–A2044.

[75] D. Kressner and C. Tobler, *Low-rank tensor Krylov subspace methods for parametrized linear systems*, SIAM J. Matrix Anal. Appl., 32 (2011), pp. 1288–1316.

[76] ——, *Algorithm 941: htucker—a Matlab toolbox for tensors in hierarchical Tucker format*, ACM Trans. Math. Software, 40 (2014).

[77] P. Krokhmal, J. Palmquist, and S. Uryasev, *Portfolio Optimization With Conditional Value-at-Risk Objective and Constraints*, Journal of Risk, 4 (2002), pp. 43–68.

[78] A. Kunoth and C. Schwab, *Analytic Regularity and GPC Approximation for Control Problems Constrained by Linear Parametric Elliptic and Parabolic PDEs*, SIAM J. Control Optim., 51 (2013), pp. 2442–2471.

[79] ——, *Sparse Adaptive Tensor Galerkin Approximations of Stochastic PDE-Constrained Control Problems*, SIAM/ASA J. Uncertain. Quantif., 4 (2016), pp. 1034–1059.

[80] F. Y. Kuo, C. Schwab, and I. H. Sloan, *Quasi-Monte Carlo Finite Element Methods for a Class of Elliptic Partial Differential Equations with Random Coefficients*, SIAM J. Numer. Anal., 50 (2012), pp. 3351–5574.

[81] L. De Lathauwer, B. De Moor, and J. Vandewalle, *A multilinear singular value decomposition*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1253–1278.

[82] H.-C. Lee and J. Lee, *A Stochastic Galerkin Method for Stochastic Control Problems*, Commun. Comput. Phys., 14 (2013), pp. 77–106.

[83] V. Murg, F. Verstraete, Ö. Legeza, and R. M. Noack, *Simulating strongly correlated quantum systems with tree tensor network*, Phys. Rev. B, 82 (2010).

[84] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, vol. 63 of CBMS-NSF Regional Conf. Ser. in Appl. Math., SIAM, Philadelphia, 1992.

[85] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, second ed., 2006.

[86] E. Novak and K. Ritter, *High dimensional integration of smooth functions over cubes*, Numer. Math., 75 (1996), pp. 79–97.

[87] I. V. Oseledets, *DMRG approach to fast linear algebra in the TT-format*, Comput. Methods Appl. Math., 11 (2011), pp. 382–393.

[88] ——, *Tensor-Train Decomposition*, SIAM J. Sci. Comput., 33 (2011), pp. 2295–2317.

[89] ——, *TT-toolbox 2.2: Fast multidimensional array operations in TT format*, 2012. https://github.com/oseledets/TT-Toolbox.

[90] I. V. Oseledets and S. V. Dolgov, *Solution of Linear Systems and Matrix Inversion in the TT-format*, SIAM J. Sci. Comput., 34 (2012), pp. A2718–A2739.

[91] I. V. Oseledets and E. E. Tyrtyshnikov, *TT-cross approximation for multidimensional arrays*, Linear Algebra Appl., 432 (2010), pp. 70–88.

[92] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Program., 58 (1993), pp. 353–367.

[93] M. Sørensen, L. de Lathauwer, P. Comon, S. Icart, and L. Deneire, *Canonical polyadic decomposition with a columnwise orthonormal factor matrix*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 1190–1213.

[94] R. T. Rockafellar and S. Uryasev, *Optimization of conditional value-at-risk*, Journal of Risk, 2 (2000), pp. 21–41.

[95] ——, *Conditional value-at-risk for general loss distributions*, J. Bank. Finance, 26 (2002), pp. 1443–1471.

[96] ——, *The fundamental risk quadrangle in risk management, optimization and statistical estimation*, Surv. Oper. Res. Manag. Sci., 18 (2013), pp. 33–53.

[97] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, Berlin Heidelberg, 1998.

[98] T. Rohwedder and A. Uschmajew, *On local convergence of alternating schemes for optimization of convex problems in the tensor train format*, SIAM J. Numer. Anal., (2013).

[99] A. Ruszczyński and A. Shapiro, *Optimization of Convex Risk Functions*, Math. Oper. Res., 31 (2006), pp. 433–452.

[100] B. Savas and L.-H. Lim, *Quasi-Newton Methods on Grassmannians and Multilinear Approximation of Tensors*, SIAM J. Sci. Comput., 32 (2010), pp. 3352–3393.

[101] G. Schulz, *Iterative Berechnung der reziproken Matrix*, ZAMM, 13 (1933), pp. 57–59.

[102] C. Schwab and C. J. Gittelson, *Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs*, Acta Numer., 20 (2011), pp. 291–467.

[103] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, SIAM, Philadelphia, 2009.

[104] M. Steinlechner, *Riemannian optimization for high-dimensional tensor completion*, SIAM J. Sci. Comput., 38 (2016), pp. S461–S484.

[105] M. Stoll and T. Breiten, *A Low-Rank in Time Approach to PDE-Constrained Optimization*, SIAM J. Sci. Comput., 37 (2015), pp. B1–B29.

[106] T. J. Sullivan, *Introduction to Uncertainty Quantification*, Springer, Cham, 2015.

[107] A. Teckentrup, R. Scheichl, M. Giles, and E. Ullmann, *Further analysis of multilevel monte carlo methods for elliptic pdes with random coefficients*, Numer. Math., 125 (2013), pp. 569–600.

[108] H. Tiesler, R. M. Kirby, D. Xiu, and T. Preusser, *Stochastic Collocation for Optimal Control Problems with Stochastic PDE Constraints*, SIAM J. Control Optim., 50 (2012), pp. 2659–2682.

[109] L. Tucker, *Some mathematical notes on three-mode factor analysis*, Psychometrika, 31 (1966).

[110] M. Ulbrich, *Semismooth Newton Methods for Operator Equations in Function Spaces*, SIAM J. Optim., 13 (2002), pp. 805–841.

[111] M. Ulbrich, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, vol. 11 of MOS-SIAM Ser. Optim., SIAM, Philadelphia, 2011.

[112] M. Ulbrich and S. Ulbrich, *Primal-dual interior-point methods for PDE-constrained optimization*, Math. Program., Ser. B, 117 (2009), pp. 435–485.

[113] A. Uschmajew and B. Vandereycken, *The geometry of algorithms using hiearchical tensors*, Linear Algebra Appl., 439 (2013), pp. 133–166.

[114] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, Chichester and Stuttgart, 1996.

[115] F. Verstraete, V. Murg, and J. I. Cirac, *Matrix product states, projected entangled pair states, and variational renormalization group methods for quantum spin systems*, Adv. Phys., 57 (2008), pp. 143–224.

[116] J. Weidmann, *Linear Operators in Hilbert Spaces*, Springer, New York, 1980.

[117] E. Zeidler, *Nonlinear Functional Analysis and its Applications*, vol. II/B: Nonlinear Monotone Operators, Springer, New York, 1990.

[118] J. C. Ziems, *Adaptive Multilevel Inexact SQP Methods for PDE-Constrained Optimization with Control Constraints*, SIAM J. Optim., 23 (2013), pp. 1257–1283.

[119] J. C. Ziems and S. Ulbrich, *Adaptive Multilevel Inexact SQP Methods for PDE-Constrained Optimization*, SIAM J. Optim., 21 (2011), pp. 1–40.