Efficient Methods for Optimal Control Problems Subject to Partial Differential Equations With Uncertain Coefficients

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
der Universität Mannheim

vorgelegt von

Philipp Arthur Guth
aus Mosbach

Mannheim, 2022

Dekan: Dr. Bernd Lübcke, Universität Mannheim
Referent: Prof. Claudia Schillings, Freie Universität Berlin
Korreferent: Prof. Stefan Vandewalle, Katholieke Universiteit Leuven
Korreferent: Prof. Ian Sloan, University of New South Wales Sydney

Tag der mündlichen Prüfung: 21. Dezember 2022

Abstract

In this thesis, we develop and analyze methods to efficiently solve optimization problems under uncertainty, constrained by partial differential equations (PDEs). The uncertainties may arise due to noisy measurements, unknown or unobservable parameters, model ambiguity, or intrinsic randomness of systems. The goal is to find a control which is robust with respect to variations in the uncertain parameters. We prove error bounds and convergence rates for the developed methods, confirm the theoretically derived results through numerical experiments, and examine the developed concepts with regard to their efficiency.
The focus of this work is the application and analysis of quasi-Monte Carlo methods, as well as the use of surrogate models for computationally intensive systems in conjunction with a penalty strategy.
We first analyze a general formulation of the optimal control problem for the existence and uniqueness of solutions, and then focus on three example problems of optimal control under uncertainty. The regularity of the problems with respect to the uncertain parameters plays a crucial role in the development and the error analysis of the methods.
The numerical treatment of the considered problems requires different approximation methods. The total approximation error is decomposed into its components and each error contribution is then studied separately in a chapter. The error estimates and convergence results developed in these chapters are not limited to problems of optimal control subject to PDE constraints with uncertain coefficients.
In addition, further strategies to increase the efficiency of the methods are investigated, such as multilevel strategies and the simultaneous solving of the optimal control problem and learning of surrogate models for computationally intensive models.

Zusammenfassung

In dieser Arbeit entwickeln und analysieren wir effiziente Methoden zur Lösung von Problemen der optimalen Steuerung mit partiellen Differentialgleichungen (pDGL), die unsichere oder zufällige Koeffizienten haben, als Nebenbedingungen. Die Unsicherheiten können aufgrund von verrauschten Messungen, unbekannten oder nicht beobachtbaren Parametern, Mehrdeutigkeit des Modells oder intrinsischer Zufälligkeit von Systemen entstehen. Gesucht ist dann eine Steuerung, die robust gegenüber Variationen der unsicheren Parameter ist. Ziel der Arbeit ist es, einerseits Fehlerschranken und Konvergenzraten für die entwickelten Methoden zu beweisen und andererseits die theoretisch hergeleiteten Resultate durch numerische Experimente zu bestätigen und die entwickelten Konzepte hinsichtlich ihrer Effizienz zu untersuchen.
Schwerpunkte dieser Arbeit sind die Anwendung und Analyse von quasi-Monte Carlo Methoden, sowie das Verwenden von Ersatzmodellen für kostenintensive Systeme in Verbindung mit einer Penalisierungsstrategie.
Wir untersuchen zunächst eine allgemeine Formulierung des Optimalsteuerungsproblems auf Existenz und Eindeutigkeit von Lösungen. Anschließend werden drei beispielhafte Probleme der optimalen Steuerung unter Unsicherheit betrachtet. Eine entscheidende Rolle für die Entwicklung und die Fehleranalyse der Methoden spielt die Regularität der Probleme bezüglich den unsicheren Parametern.
Für das numerische Lösen der betrachteten Probleme werden verschiedene Approximationsverfahren benötigt. Der Gesamtfehler der Approximation wird in dessen Bestandteile

zerlegt und jeweils separat in einem Kapitel untersucht. Die in diesen Kapiteln hergeleiteten Fehlerabschätzungen und Konvergenzresultate sind nicht beschränkt auf Probleme der optimalen Steuerung mit pDGL-Nebenbedingungen mit unsicheren Koeffizienten.
Zudem werden weitere Strategien zur zusätzlichen Steigerung der Effizienz untersucht, wie beispielsweise multilevel Strategien und das simultane Lösen des Optimalsteuerungsproblems und Lernen von Ersatzmodellen für rechenintensive Modelle.

Acknowledgements

I would like to express my deep gratitude to Claudia Schillings for providing me with the opportunity to write this thesis. Claudia has always been approachable when I needed her expert advice and had the faith in me to provide me with the freedom to pursue my research interests. I am deeply thankful for her academic and personal support and guidance since I joined her group as a PhD student.
Furthermore, I would like to thank Stefan Vandewalle and Ian Sloan for refereeing this thesis.
During my time as a PhD student, I had the pleasure to present and discuss my research at conferences and workshops in Zürich, Trier, Linz, Sydney, Berlin, Trondheim, Poznan, Konstanz, Oulu, Luminy, Bonn, and Oslo, and in many online events. I became friends with numerous great researchers. Especially, I want to thank Vesa Kaarnioja, Frances Kuo, Claudia Schillings, Ian Sloan, Andreas Van Barel, and Simon Weissmann for the joint work and a wonderful time, which gave rise to several research articles and finally this thesis.
I would also like to thank all colleagues at the Institute of Mathematics, and especially my office mates Niklas, Lukas, Simon, Matei, Vicky, Eneas, and Vesa for interesting discussions on mathematical subjects, and for entertainment during our breaks.
My special thanks goes to my family. In particular, I am deeply grateful for my parents Anita and Erwin, who encouraged me to always be curious, for Martin and for my siblings Katharina and Michael for all their support and especially for my wonderful girlfriend Paola.
Finally, I am very grateful to the "RTG 1953 - Probability & Statistics group Heidelberg-Mannheim" funded by the Deutsche Forschungsgesellschaft for funding my research. I am thankful for the travel support I was granted by the HCM Bonn, RICAM Linz, ALOP Trier, and especially the IPID4all grant for funding towards travel and expenses related to a research visit at University of New South Wales, Sydney, Australia. Moreover, I acknowledge support by the state of Baden-Württemberg through bwHPC and the Research Technology Service at UNSW Sydney through the Katana cluster.

Contents

1 Introduction

Many complex systems in science and engineering can be modeled as partial differential equations (PDEs). For instance, PDEs are foundational in the understanding of sound, heat, diffusion, electrostatics, electro-, thermo-, and fluid dynamics, elasticity, and many more. Hence, they are ubiquitous in airfoil, in problems concerning groundwater flow, in weather simulations, in computer tomography, and in microelectronics – to name only a few applications. The mathematical optimization of processes and systems that can be modeled by PDEs is an essential task for scientists and engineers across disciplines.

If not taken into account, limited knowledge or intrinsic randomness of parameters in the PDE model, such as material properties, external conditions, or reaction constants, have the potential to render worthless any solutions obtained using state-of-the-art methods for deterministic problems. The careful analysis of the uncertainty in PDE-constrained optimization problems is hence indispensable and has become a growing field of research.

Supposing that the practitioner has some control over the uncertain state of the system, the goal is to determine the optimal control input (if it exists) for the uncertain system. The quality measure is given by a cost functional which is composed with a risk measure taking the uncertainties into account.

The uncertainties often manifest themselves as random fields, which can be represented by a countably infinite number of random parameters. For the numerical treatment of such problems, a natural first step is the truncation of the representation at a finite (possibly very large) number of random parameters. The resulting error is analyzed in Chapter 5.

In order to robustify the optimal control with respect to the uncertainty, a risk measure is applied which involves integrals over the high-dimensional domain of the parameters. While being dimension independent, Monte Carlo methods obtain a notoriously slow convergence rate. Moreover, for sufficiently regular integrands in the presented setting it is possible to construct quasi-Monte Carlo (QMC) rules with error bounds not depending on the number of stochastic variables, while attaining faster convergence rates compared to Monte Carlo methods. Moreover, QMC approximations are particularly well suited for optimization since they preserve convexity due to their nonnegative (equal) cubature weights as opposed to sparse grid methods, for instance. QMC methods are discussed in Chapter 6.

To further reduce the computational cost we consider a multilevel QMC method that efficiently distributes the number of samples across different discretization levels of the underlying PDE. Moreover, the simultaneous learning of surrogates, such as polynomial expansions or neural networks, for the computational intensive PDE solution is investigated in a one-shot optimization framework.

1.1 Outline

We describe the structure of this thesis together with a brief outline of the following chapters.

Chapter 2

We start this thesis with a collection of definitions, notational conventions, and well-known results in order to embed the results in the following chapters into a rigorous mathematical setting. Thereby we focus on functional analysis and integration theory.

Chapter 3

We formulate the optimal control problem with PDE constraints under uncertainties in a very general setting. We list popular risk measures and classify them according to desirable properties. We then derive moderate conditions on the risk measure, the random variable objective function, and the PDE constraint for the existence and optimality of solutions in the general setting. For well-posed forward problems one can reformulate the optimal control problem in the so-called reduced formulation. Moreover, in the setting of parametric linear forward operators we show equivalence between the almost sure formulation of the constraints and the weak formulation in the parameter space.

Chapter 4

We consider three examples of optimal control problems. In all three example problems, we consider a tracking type objective functional composed with different risk measures:

- In Section 4.1 the risk measure is the expected value and the optimal control problem is subject to an elliptic PDE with a random diffusion coefficient. We suppose to have control over the source term of the PDE.
- In Section 4.2 the risk measure is either the expected value or the entropic risk measure. The constraint is a parabolic PDE with an uncertain diffusion coefficient, and we control the source term of the PDE for a given initial condition.
- In Section 4.3 the risk measure is again either the expected value or the entropic risk measure. The constraint is an abstract parametric linear operator equation with affine parameter dependence. In particular, the problems Section 4.1 and Section 4.2 fit into this framework. Nevertheless, we chose to present the elliptic and parabolic examples for better illustration.

In all examples we discuss the function space setting of the PDEs (operator equation, respectively), present their parametric weak formulation, and derive optimality conditions, that are based on the adjoint states, of the reduced formulation of the problem. We note that all examples fit into the abstract framework presented in Chapter 3, but the results are derived for the examples for clarity. Furthermore, we present a well-known optimization algorithm which can be used to solve the three example problems, and illustrate how the total discretization error can be decomposed into its contributions. The different error contributions are then analyzed separately in Chapter 5, Chapter 6, and Section 7.1.

The heart of the following error analysis is the parametric structure and the parametric regularity of the PDEs or operators, respectively. To this end, we investigate the regularity of the example problems with respect to the uncertain parameters.
This chapter is based on joint work Vesa Kaarnioja, Frances Y. Kuo, Claudia Schillings, and Ian H. Sloan and the two corresponding articles:

- A Quasi-Monte Carlo Method for Optimal Control Under Uncertainty. *SIAM/ASA J. Uncertain. Quantif.*, 9(2): 354–383, 2021. `https://doi.org/10.1137/19M1294952`.
- Parabolic PDE-constrained optimal control under uncertainty with entropic risk measure using quasi-Monte Carlo integration, 2022. Preprint at `https://arxiv.org/abs/2208.02767`.

Chapter 5

This chapter is devoted to the dimension truncation error, which is a natural error contribution arising in the discretization of infinite-dimensional integration problems. The problem is formulated in the general setting of integrands that belong to separable Banach spaces with generalized β-Gaussian distributed random parameters. Hence, the results presented in this chapter are not at all restricted to optimal control problems. In particular, we derive a set of sufficient conditions that guarantee convergence with a rate that appears to be superior to the existing literature for some values of β. Furthermore, since our results are stated in separable Banach spaces, they directly apply to PDE solutions discretized by conforming finite elements. Moreover, the setting is not restricted to PDEs, but only based on the parametric regularity of the Banach space-valued integrands. We can thus, for instance, compose an element in a separable Banach space (possibly a PDE solution) with an arbitrary nonlinear quantity of interest as long as the composition with the quantity of interest satisfies the hypothesis of our results.
This chapter is based on joint work with Vesa Kaarnioja and the corresponding article:

- Generalized dimension truncation error analysis for high-dimensional numerical integration: lognormal setting and beyond, 2022. Preprint at `https://arxiv.org/abs/2209.06176`.

This chapter and the corresponding article were motivated by the joint work with Vesa Kaarnioja, Frances Y. Kuo, Claudia Schillings, and Ian H. Sloan and the two corresponding articles listed in the outline of Chapter 4.

Chapter 6

We provide a brief introduction to quasi-Monte Carlo methods and particularly to randomly shifted rank-1 lattice rules in Chapter 6. The main contribution of this chapter is the generalization of existing error bounds and convergence rates for real-valued integrands to the general setting of integrands in separable Banach spaces. This generalization opens up many new areas of application for QMC methods, such as optimal control problems with PDE constraints under uncertainty. This chapter is based on the joint work with Vesa Kaarnioja, Frances Y. Kuo, Claudia Schillings, and Ian H. Sloan and the corresponding article

- Parabolic PDE-constrained optimal control under uncertainty with entropic risk measure using quasi-Monte Carlo integration, 2022. Preprint at `https://arxiv.org/abs/2208.02767`.

Chapter 7

This chapter concentrates on the spatial discretization of the PDE constraints. In Section 7.1, we provide a brief overview of the finite element method (FEM) that is used in the numerical experiments throughout this thesis. We derive an error bound and a convergence rate for the elliptic example, which – together with the truncation error and cubature error – completes the error analysis presented in Section 4.5. Parts of this chapter are based on joint work with Vesa Kaarnioja, Frances Y. Kuo, Claudia Schillings, and Ian H. Sloan and the corresponding article:

- A Quasi-Monte Carlo Method for Optimal Control Under Uncertainty. *SIAM/ASA J. Uncertain. Quantif.*, 9(2): 354–383, 2021. `https://doi.org/10.1137/19M1294952`.

The second part of this chapter is about a multilevel strategy that efficiently distributes samples across different FE discretization levels. More precisely, we analyze the application of a multilevel quasi-Monte Carlo (MLQMC) method to the optimal control problem in combination with the circulant embedding method in order to sample the lognormal random field. The MLQMC part of this chapter is based on joint work with Andreas Van Barel and the corresponding article:

- Multilevel Quasi-Monte Carlo for Optimization under Uncertainty, 2021. Preprint at `https://arxiv.org/abs/2109.14367`.

Chapter 8

Novel results on the use of machine learning techniques motivated Chapter 8. In particular, we reformulate the PDE-constrained optimization problem under uncertainty as an unconstrained optimization problem by adding a quadratic penalty on the PDE residual to the objective function. We then replace the computational intense solution of the state of the system with a surrogate. The surrogate parameters are learned simultaneously during the optimization, and hence the surrogate only needs to be trained for the optimal control. Opposed to the simultaneous training, training of the surrogate parameters before the optimization must lead to a good surrogate for all feasible controls. Section 8.1 is based on the joint work with Claudia Schillings and Simon Weissmann and the corresponding article

- A General Framework for Machine Learning based Optimization Under Uncertainty, 2021. Preprint at `https://arxiv.org/abs/2112.11126`.

In the second part of this chapter we transfer the ideas of Section 8.1 to the setting of Bayesian inverse problems. Establishing a connection between the Bayesian approach and the one-shot formulation allows to interpret the penalization parameter as the level of model error in the forward problem, i.e., increasing the penalization parameter on the quadratic model residual corresponds to vanishing model noise in the Bayesian setting.

Furthermore, we show that the ensemble Kalman inversion is an efficient method to solve the resulting optimization problem. Section 8.2 is based on the joint work with Claudia Schillings and Simon Weissmann and the corresponding article

- 14 Ensemble Kalman filter for neural network based one-shot inversion. In *Optimization and Control for Partial Differential Equations: Uncertainty quantification, open and closed-loop control, and shape optimization* edited by R. Herzog, M. Heinkenschloss, D. Kalise, G. Stadler, E. Trélat, pp. 393–418. Berlin, Boston: De Gruyter, 2022. `https://doi.org/10.1515/9783110695984-014`.

2 Selected facts from functional analysis and measure theory

We will discuss, analyze, and solve constrained optimization problems with uncertainties entering the problem through the constraint. In this section we begin with the mathematical description of the problems studied in this manuscript.
To introduce the notation, we begin by recalling some basic results from functional analysis and measure theory. Most of the results presented in this chapter are well-known and can be found in textbooks like [13, 32, 42, 81, 90, 104, 135, 139, 140, 145].

2.1 Fundamental functional analysis

Linear space By $\mathbb{R}$ and $\mathbb{C}$ we will always denote the field of real numbers and the field of complex numbers respectively. An element in the scalar field $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ will be called scalar. A linear space (or vector space) over $\mathbb{K}$ is a set X, consisting of elements that are called vectors, and in which addition (A1-A4) and scalar multiplication (M1-M3) are defined by the following algebraic properties:

A1 For all vectors x, y, and $z \in X$, it holds $(x + y) + z = x + (y + z)$.

A2 For all vectors $x, y \in X$, it holds $x + y = y + x$.

A3 X contains a unique vector 0, such that $x + 0 = x$ for every $x \in X$.

A4 To each $x \in X$ corresponds a unique vector $-x$, such that $x + (-x) = 0$.

A1 and A2 are called associative property and commutative property, respectively. The vector 0 in $A3$ is called the neutral element of vector addition and the vector $-x$ in A4 is called the inverse element.

M1 For each $\alpha, \beta \in \mathbb{K}$ and $x, y \in X$ we have $\alpha(x+y) = \alpha x + \alpha y$ and $(\alpha + \beta)x = \alpha x + \beta x$.

M2 For each $\alpha, \beta \in \mathbb{K}$ and $x \in X$ we have $(\alpha\beta)x = \alpha(\beta x)$

M3 For each $x \in X$ we have for $1 \in \mathbb{K}$ that $1\,x = x$

M1 and M2 are called distributative property and associative property, respectively. The element 1 in M3 is called neutral element of scalar multiplication.

Normed space A linear space X is called a normed space, or normed linear space, if to every $x \in X$ there is assigned a nonnegative real number $\|x\|$, called the norm of x. A norm $\|\cdot\| : X \to [0, \infty)$ is a mapping with the properties

(i) For all $x \in X$ it holds that $\|x\| = 0$ implies $x = 0$.

(ii) For all $x \in X$ and $\alpha \in \mathbb{K}$ it holds that $\|\alpha x\| = |\alpha|\|x\|$.

(iii) For all $x, y \in X$ it holds that $\|x + y\| \leqslant \|x\| + \|y\|$.

The property (i) is called positive definiteness, (ii) is called absolute homogeneity and (iii) is the triangle inequality or subadditivity.
We will later often use the notation $\|\cdot\|_X$ if the space X is not clear from the context. Moreover by the absolute value (or modulus) $|x|$ of x is a norm on the one-dimensional linear space formed by the real or complex numbers.

Linear operator A mapping $A : X \to Y$ from a normed space X to a normed space Y is called operator. The operator A is called bounded if

$$\|A\|_{Y \leftarrow X} := \sup_{x \in X \colon \|x\| \leqslant 1} \|Tx\|_Y < \infty \,.$$

We call $\|A\|_{Y \leftarrow X}$ the operator norm of the operator A.
The set of all bounded linear operators $A : X \to Y$ will be denoted by $\mathcal{L}(X, Y)$. Together with the operator norm $\|\cdot\|_{Y \leftarrow X}$, the space of all bounded linear operators $\mathcal{L}(X, Y)$ is a normed space, see [135, Theorem 4.1]. If $X = Y$, we will use the abbreviation $\mathcal{L}(X)$ instead of $\mathcal{L}(X, X)$. Moreover, on a normed space X, we denote the identity by $\mathcal{I}_X \in \mathcal{L}(X)$. The mapping A^{-1} is called the inverse operator of the mapping $A \in \mathcal{L}(X, Y)$ if it holds that $AA^{-1} = \mathcal{I}_Y$ and $A^{-1}A = \mathcal{I}_X$.

Isomorphism Let X, Y be two linear spaces over the same field $\mathbb{K}$. A bijective mapping $A : X \to Y$, i.e., a mapping that is injective ($Ax_1 = Ax_2$ implies $x_1 = x_2$) and surjective (for all $y \in Y$ there exists an $x \in X$ with $Ax = y$), which preserves the algebraic properties of the linear space, is called an isomorphism. Moreover, if there is an isomorphism between two linear spaces X, Y, we say X and Y are isomorphic and write $X \cong Y$. If X, Y are normed spaces and it holds in addition that $\|Ax\|_Y = \|x\|_x$ for all $x \in X$, we call A an isometric isomorphism.

Metric space Every normed space X can be considered as a metric space, in which the distance, or metric, $d(x, y)$ between two elements x and y is defined by the norm $d(x, y) := \|x - y\|$. A metric $d : X \times X \to [0, \infty)$ is characterized by the following properties:[1]

(i) It holds that $d(x, y) = 0$ if and only if $x = y$.

(ii) For all x, y, and $z \in X$ it holds that $d(x, z) \leqslant d(x, y) + d(y, z)$.

(iii) For all $x, y \in X$ it holds that $d(x, y) = d(y, x)$.

The property (i) is called positive definiteness, (ii) is called triangle inequality and (iii) is called symmetry.

[1] The symbol $\times$ denotes the Cartesian product, i.e., $X \times Y$ is the set of all ordered pairs (x, y) with $x \in X$ and $y \in Y$.

Density A subset M of a metric space X is called a dense subset of X if the closure of M, $\overline{M}$, is equal to the superset X. That is every point in X either belongs to M or is a limit point of M, i.e., for every $x \in X$ there is a sequence $(x_n)_{n\in\mathbb{N}}$ in M such that the limit is also in M: $\lim_{n\to\infty} x_n = x$.
A metric space X is called separable, if it contains a countable and dense subset $E = \{e_n : n \in \mathbb{N}\}$.
In any metric space X one can define the open ball centered at $x \in X$, and with radius $r \geqslant 0$, as the set

$$B_r(x) := \{y : d(x, y) < r\}.$$

In particular, if X is a normed space, the sets

$$B_r(0) = \{x : \|x\| < 1\} \quad \text{and} \quad \bar{B}_r(0) = \{x : \|x\| \leqslant 1\}$$

are the open and closed unit ball in X, respectively.

Topology By declaring a subset of a metric space to be open if and only if it is a (possibly empty) union of open balls, we obtain a topology. A topological space is a set X containing a collection τ of subsets satisfying the following proerties

(i) The empty set $\varnothing$ and X itself belong to τ.

(ii) Any union of members of τ belongs to τ.

(iii) The intersection of any two members of τ belongs to τ.

Such a collection τ is called topology on X and the elements of τ are called open sets. Moreover a subset $A \subset X$ is called closed if and only if its complement $X\backslash A$ is open. If τ_1, τ_2 are two topologies on a common space X, we say that τ_1 is weaker than τ_2 (or equivalently τ_2 is stronger than τ_1) if $\tau_1 \subset \tau_2$.
Let $\mathcal{J}$ be an arbitrary index set, let X_j for $j \in \mathcal{J}$ be a topological space, and let $X = \prod_{j\in\mathcal{J}} X_j$ be the Cartesian product of the X_j. For each $j \in \mathcal{J}$ we call $P_j : X \to X_j$ the canonical projection. The weakest topology τ such that all canonical projections are continuous with respect to τ is called the product topology. The pair (X, τ) is called product space.
A topological space is said to be compact if each of its open covers has a finite subcover. That is, X is compact if for every collection A_1 of open sets of X with $X = \bigcup_{x\in A_1} x$, there is a finite subcollection $A_2 \subseteq A_1$ such that $X = \bigcup_{x\in A_2} x$.
By Tychonov's theorem [135, Theorem A3], the caresian product of any nonempty collection of compact spaces $(X_j)_{j\in\mathcal{J}}$ is compact. Here $\mathcal{J}$ is again an arbitrary index set.

Complete metric space A sequence $(x_n)_{n\in\mathbb{N}}$ of elements of a metric space X is called Cauchy sequence if for all $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $n, m \geqslant N$ it holds for the distance $d(x_n, x_m) < \epsilon$. Moreover the sequence $(x_n)_{n\in\mathbb{N}}$ converges to $x \in X$ if for all $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that for all $n \geqslant N$ it holds for the distance $d(x_n, x) < \epsilon$. A metric space is called complete if every Cauchy sequence converges.

Banach space A normed space which is complete in the metric induced by its norm is called a Banach space.

Linear isomorphism between Banach spaces Let X, Y be two Banach spaces. Then for any operator $A \in \mathcal{L}(X, Y)$ that is bijective, i.e., injective ($Ax_1 = Ax_2$ implies $x_1 = x_2$) and surjective (for all $y \in Y$ there exists an $x \in X$ with $Ax = y$), the operator $A^{-1} \in \mathcal{L}(Y, X)$ exists. This is a consequence of the open mapping theorem [135, Corollary 2.12].

Embeddings Let X, Y be two normed spaces with $X \subset Y$. Clearly the embedding (or injection) $\mathcal{I} : X \to Y$, defined by $\mathcal{I}x = x$ for all $x \in X$ is linear. We say X is continuously embedded in Y if $\mathcal{I}$ is also bounded, i.e., $\|x\|_Y = \|\mathcal{I}x\|_X \leqslant C\|x\|_X$ for all $x \in X$ and a constant $C > 0$. Moreover, if X is dense in Y, we say X is densely and continuously embedded in Y.

Dual space For a normed linear space over a field $\mathbb{K} \in \{\mathbb{R}, \mathbb{K}\}$, we call the space of all bounded linear mappings (or functionals)

$$X' = \mathcal{L}(X, \mathbb{K})$$

the dual space X' of X. Equipped with the norm

$$\|x'\|_{X'} := \sup_{x \in X:\, \|x\|_X \leqslant 1} |x'(x)| \tag{2.1}$$

the dual space X' of X is a Banach space.[2] For $x'(x)$ we will also write

$$\langle x, x' \rangle_{X,X'} = \langle x', x \rangle_{X',X} := x'(x)\,,$$

where we call $\langle \cdot, \cdot \rangle_{X,X'}$ and $\langle \cdot, \cdot \rangle_{X',X}$ dual forms or duality pairings.
If a normed linear space X is continuously embedded in a normed linear space Y, then $Y' \subset X'$ is continuously embedded, see [140, Lemma 2.2.11]

Dual Operator Let X, Y be two normed spaces. For each $A \in \mathcal{L}(X, Y)$ there is a unique dual operator $A' \in \mathcal{L}(Y', X')$ satisfying $\langle Ax, y \rangle_{Y,Y'} = \langle x, A'y' \rangle_{X,X'}$ for all $x \in X$ and $y' \in Y'$. Furthermore, their operator norms are identical: $\|A\|_{\mathcal{L}(X,Y)} = \|A'\|_{\mathcal{L}(Y',X')}$.
As a consequence of the Hahn–Banach theorem [135, Theorem 3.5], we can write by [135, Theorem 4.3] the following

$$\|x\|_X = \sup_{x' \in X':\, \|x'\| \leqslant 1} |\langle x, x' \rangle_{X,X'}|\,.$$

Using the dual space X' one can define a topology on X:

(i) The weak topology is the weakest topology on X that makes all maps $x'(\cdot) = \langle \cdot, x' \rangle_{X,X'} : X \to \mathbb{K}$ continuous, as x' ranges over X'.

(ii) The weak* topology is the weakest topology on X' that makes all maps $\langle x, \cdot \rangle_{X,X'} : X' \to \mathbb{K}$ continuous, as x ranges over X.

We say a sequence $(x_j)_{j \in \mathbb{N}}$ converges weakly (or in weak topology) to $x \in X$, denoted by $x_j \rightharpoonup x$, if $\phi(x_j) \to \phi(x)$ for all $\phi \in X'$. Clearly, convergence in the (stronger) norm topology implies convergence in the weak topology. The reverse holds for instance if

[2] In fact, for two normed spaces X and Y, the space of bounded linear operators $\mathcal{L}(X, Y)$ equipped with the operator norm $\|\cdot\|_{Y \leftarrow X}$ is a normed space. If Y is in addition a Banach space, then so is $\mathcal{L}(X, Y)$.

$\dim(X) < \infty$, but is not true in general. The weak* topology is important, since the Banach–Alaoglu theorem [135, Theorem 3.15] implies that the closed unit ball in the dual space X' of a normed space X is compact with respect to the weak* topology. Note that, if X' is infinite-dimensional, the closed unit ball cannot be compact with respect to the norm topology. This is a consequence of Riesz' lemma, which tells us that the unit ball in a normed linear space X' is compact if and only if X' is finite-dimensional, see [42, Theorem 4 in Chapter I].

Bidual space and reflexivity The dual space X' of a Banach space X over a field $\mathbb{K}$ is itself a Banach space and thus has its own dual space X'', called the bidual space of X. Hence, the bidual space of X is defined by $X'' = \mathcal{L}(X', \mathbb{K})$. Moreover, a consequence of [135, Theorem 4.3] is, that every $x \in X$ and $x' \in X'$ defines a unique $\phi x \in X''$ by the equation

$$\langle x, x' \rangle_{X,X'} = \langle x', \phi x \rangle_{X',X''} \tag{2.2}$$

and, for all $x \in X$, that

$$\|\phi x\| = \|x\| .$$

Hence, $\phi : X \to X''$ is a linear isometry. Since X is complete, $\phi(X)$ is closed in X''. Hence ϕ is an isometric isomorphism of X onto a closed subspace of X''. In this case X is usually identified with $\phi(X)$, a subspace of X''. Note that $\phi(X)$ contains the linear functionals on X' that are continuous relative to its weak* topology. It may therefore happen that $\phi(X)$ is a proper subspace of X'' as the norm topology of X' is stronger. The spaces for which the mapping ϕ is bijective with $\phi(X) = X''$ are called reflexive. Note that the existence of some isometric isomorphism ϕ is not sufficient for X to be reflexive, but it is essential that (2.2) is satisfied by ϕ.

Inner product space We call a linear space X an inner product space if to each ordered pair of vectors x and y in X, a scalar $\langle x, y \rangle_X$, called the inner product of x and y, is associated, which has the following properties:

(i) For all vectors $x, y \in X$ it holds that $\langle y, x \rangle_X = \overline{\langle x, y \rangle_X}$[3]

(ii) For all vectors x, y, and $z \in X$ it holds that $\langle x + y, z \rangle_X = \langle x, z \rangle + \langle y, z \rangle_X$.

(iii) For all scalar $\alpha \in \mathbb{K}$ and vectors $x, y \in X$ it holds that $\langle \alpha x, y \rangle_X = \alpha \langle x, y \rangle_X$.

(iv) For all $x \in X$ it holds $\langle x, x \rangle_X \geqslant 0$.

(v) It holds that $\langle x, x \rangle_X = 0$ if and only if $x = 0$.

Here the first property is called conjugate symmetry. The properties (ii) and (iii) describe linearity in the first argument and the last two properties are sometimes referred to as positive definiteness.
If $\langle x, y \rangle_X = 0$ we say that x is orthogonal to y and sometimes use the notation $x \perp y$. Moreover, we use the notation $E \perp F$ for $E, F \subset X$, to denote that $x \perp y$ whenever $x \in E$ and $y \in F$. The set of all $y \in X$ that are orthogonal to every $x \in E$ is denoted by $E^{\perp}$.

[3] Here $\bar{\alpha}$ denotes the complex conjugate of α, i.e., $\bar{\alpha} = a - b \cdot i$ for $\alpha = a + b \cdot i$, where $a, b \in \mathbb{R}$ and i denotes the imaginary unit.

Hilbert space Every inner product space X can be normed by

$$\|x\|_X = \sqrt{\langle x, x\rangle_X}$$

for all $x \in X$, see [135, Theorem 12.2]. In case the resulting normed space is complete, it is called a Hilbert space. Some important properties of Hilbert spaces are:

(i) The Cauchy–Schwarz inequality holds: $|\langle x, y\rangle_X| \leqslant \|x\|_X \|y\|_X$ for all $x, y \in X$.

(ii) For $E \subset X$, the orthogonal complement $E^\perp$ is a closed subspace of X.

(iii) Let E be a closed subspace of a Hilbert space X with orthogonal complement $E^\perp$. Then it holds that $X = E \oplus E^\perp$, that is any $x \in X$ can uniquely be decomposed by $x = u + v$ with $u \in E$ and $v \in E^\perp$, and it holds $\|x\|_X^2 = \|u\|_X^2 + \|v\|_X^2$.

(iv) Let Y be a closed subspace of the Hilbert space X. For $x \in X$ there exists a unique $\hat{y}(x) \in Y$ with

$$\|x - \hat{y}\|_X = \min_{y \in Y} \|x - y\|_X \,.$$

The mapping $Px := \hat{y}(x)$ is called an orthogonal projection. The projection operator P is linear, bounded, self-adjoint ($\langle Px, y\rangle_X = \langle x, Py\rangle_X$) and idempotent ($P^2 = P$).

Riesz' representation theorem Let X be a Hilbert space over a field $\mathbb{K}$. For any $y \in X$ the mapping

$$f_y(\cdot) := \langle \cdot, y\rangle_X : X \to \mathbb{K}$$

is a bounded linear functional. Hence, it holds that $f_y(\cdot) \in X'$ and $\|f_y\|_{X'} = \|y\|_X$. The converse result is known as Riesz' representation theorem: For all bounded linear functionals $f \in X'$ there exists a unique vector $y_f \in X$ that satisfies

$$f(x) = \langle x, y_f\rangle_X \text{ for all } x \in X \quad \text{and } \|f\|_{X'} = \|y_f\|_X \,.$$

Some important consequences of this result are:

(i) There exists a bounded, invertible conjugate linear[4] mapping $R_X : X \to X'$ with $R_X y = f_y$ and $R_X^{-1} f = y_f$. Moreover, the mapping R_X is an isometry: $\|R_X\|_{X \to X'} = \|R_X^{-1}\|_{X' \to X} = 1$.

(ii) The dual space X' is a Hilbert space with inner product $\langle x', y'\rangle_{X'} := \overline{\langle R_X^{-1} x', R_X^{-1} y'\rangle_X}$ and the norm in (2.1) equal to $\|x'\|_{X'} = \sqrt{\langle x', x'\rangle_{X'}}$.

(iii) $X \cong X''$ with $x(x') := x'(x)$, we can identify X with X'', and in particular any Hilbert space X is reflexive. Moreover, it holds that $R_{X'} = R_X^{-1}$, $R_X = (R_X)'$, and $A'' = A$ for $A \in \mathcal{L}(X, Y)$ if $Y = Y''$ and if both are Hilbert spaces.

[4] A mapping $f : X \to Y$ between to linear spaces over $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ is conjugate linear if $f(\alpha x + \beta y) = \overline{\alpha} f(x) + \overline{\beta} f(y)$ for all $x, y \in X$ and all $\alpha, \beta \in \mathbb{K}$.

(iv) If $\mathbb{K} = \mathbb{R}$ the spaces X and X' can be identified due to the isomorphism R_X, i.e., $X := X'$ implies $R_X = \mathcal{I}$. If $\mathbb{K} = \mathbb{C}$ one can choose an orthonormal basis[5] $(x_i)_{i\in\mathcal{S}}$ in X and define the complex conjugation by $Cx := \overline{x} := \sum_{j\in\mathcal{S}} \overline{\langle x, x_j\rangle_X} x_v$. Then $C = C^{-1}$ and C, C^{-1} are conjugate linear isometries. Hence $\overline{R}_X := R_X C$ is an isometric isomorphism and one can identify any Hilbert space over $\mathbb{K} = \mathbb{C}$ with its own dual.

(v) The adjoint operator A^* of an operator $A \in \mathcal{L}(X, Y)$ between two Hilbert spaces X, Y can be defined in terms of R_X as follows: $A^* := R_X^{-1} A' R_Y \in \mathcal{L}(Y, X)$. Moreover, we have $\|A\|_{\mathcal{L}(X,Y)} = \|A^*\|_{\mathcal{L}(Y,X)}$ and $\langle Ax, y\rangle_Y = \langle x, A^* y\rangle_X$ for all $x \in X$ and all $y \in Y$.

Gelfand triplet Let X, Y be two Hilbert spaces and let $X \subset Y$ be continuously and densely embedded. Then $Y' \subset X'$ is also continuously and densely embedded, see e.g. [140, Proposition 2.1.22]. One can identify Y with its own dual Y' and obtain the Gelfand triple

$$X \subset Y \subset X' ,$$

where both embeddings are continuous and dense. By this identification, the inner product $\langle x, y\rangle_Y$ can also be interpreted as the duality pairing $\langle x, y\rangle_{Y,Y'}$. For $x \in X \subset Y$, we have $y(x) = \langle x, y\rangle_{X,X'} = \langle x, y\rangle_Y$ for all $y \in Y \subset X'$. Since the embedding $Y \subset X'$ is continuous and dense, the inner product $\langle \cdot, \cdot\rangle_Y$ can be extended continuously to the duality pairing $\langle \cdot, \cdot\rangle_{X,X'}$.

Sesquilinear and bilinear form Let X, Y be linear spaces over $\mathbb{K}$. A mapping $a(\cdot, \cdot) : X \times Y \to \mathbb{K}$ is called sesquilinear form if for all $x_1, x_2 \in X$, for all $y_1, y_2 \in Y$ and all $\alpha \in \mathbb{K}$ it holds that

$$\begin{aligned} a(x_1 + \alpha x_2, y_1) &= a(x_1, y_1) + \alpha a(x_2, y_1) , \\ a(x_1, y_1 + \alpha y_2) &= a(x_1, y_1) + \overline{\alpha} a(x_1, y_2) . \end{aligned}$$

If $\mathbb{K} = \mathbb{R}$ we call $a(\cdot, \cdot) : X \times Y \to \mathbb{R}$ a bilinear form. In the case when X, Y are normed spaces, we say a sesquilinear form is continuous (or bounded) if there is a positive constant $C < \infty$ such that

$$|a(x, y)| \leqslant C \|x\|_X \|y\|_Y ,$$

for all $x \in X$, $y \in Y$. The norm of the sesquilinear form $a(\cdot, \cdot)$ is the smallest such constant:

$$\|a\| := \sup_{x\in X\setminus\{0\}} \sup_{y\in Y\setminus\{0\}} \frac{|a(x, y)|}{\|x\|_X \|y\|_Y} .$$

Bilinear form and linear operator To each continuous bilinear form $a(\cdot, \cdot) : X \times Y \to \mathbb{R}$ we can uniquely associate an operator $A \in \mathcal{L}(X, Y')$ such that

$$a(x, y) = \langle Ax, y\rangle_{Y',Y} \quad \forall x \in X, y \in Y ,$$

and

$$\|A\|_{\mathcal{L}(X,Y')} = \|a\| .$$

[5] A system of orthonormal vectors $(x_i)_{i\in\mathcal{S}}$ in a Hilbert space X is an orthonormal basis of X if, for every $x \in X$, the (Fourier) expansion $x = \sum_{i\in\mathcal{S}} \langle x, x_i\rangle_X x_i$ convergens.

Proof. From the continuity of $a(\cdot,\cdot) : X \times Y \to \mathbb{R}$ we get for arbitrary $x \in X$, that $\phi_x(y) := a(x,y)$ defines a bounded linear functional on Y, i.e., $\phi_x(\cdot) \in Y'$. In particular, we get $\|\phi_x\|_{Y'} \leqslant C\|x\|_X$ for all $x \in X$ and a constant $0 < C$. Defining $Ax := \phi_x$ for all $x \in X$, A is clearly linear, and we obtain $\|Ax\|_{Y'} \leqslant C\|x\|_X$ for all $x \in X$, and the same C, hence $A \in \mathcal{L}(X, Y')$. Conversely, let $A \in \mathcal{L}(X, Y')$. Then $a(x,y) := \langle Ax, y\rangle_{Y',Y}$ is a bilinear form with

$$|\langle Ax, y\rangle_{Y',Y}| \leqslant \|Ax\|_{Y'}\|y\|_Y \leqslant \|A\|_{\mathcal{L}(X,Y')}\|x\|_X\|y\|_Y\,.$$

The equality of the norms then follows per definition:

$$\|A\|_{\mathcal{L}(X,Y')} = \sup_{x\in X\setminus\{0\}} \frac{\|Ax\|_{Y'}}{\|x\|_X} = \sup_{x\in X\setminus\{0\}} \sup_{y\in Y\setminus\{0\}} \frac{|\langle Ax, y\rangle_{Y',Y}|}{\|x\|_X\|y\|_Y} = \|a\|\,.$$

□

Weak formulation of the problem and operator equation Let X, Y be normed spaces, $a(\cdot,\cdot) : X \times Y \to \mathbb{R}$ be a continuous bilinear form, and $f : Y \to \mathbb{K}$ a continuous linear functional. The weak problem is: Find $x \in X$ such that

$$a(x,y) = f(y) \quad \forall y \in Y\,.$$

Many differential and integral equations can be formulated as weak problems. By replacing the sesquilinear form with the associated operator we observe that the weak problem can be equivalently stated as an operator equation in Y':

$$\langle Ax, y\rangle_{Y',Y} = \langle f, y\rangle_{Y',Y} \quad \text{or equivalently} \quad Ax = f \ \text{ in } Y'\,. \tag{2.3}$$

Inf-sup-conditions We say the bilinear form $a(\cdot,\cdot)$ satisfies the inf-sup-conditions if there is a $\kappa > 0$ such that

$$\inf_{x\in X\setminus\{0\}} \sup_{y\in Y\setminus\{0\}} \frac{|a(x,y)|}{\|x\|_X\|y\|_Y} \geqslant \kappa\,, \tag{2.4a}$$

$$\inf_{y\in Y\setminus\{0\}} \sup_{x\in X\setminus\{0\}} \frac{|a(x,y)|}{\|x\|_X\|y\|_Y} \geqslant \kappa\,. \tag{2.4b}$$

For a bilinear form over two reflexive Banach spaces X and Y, we get the following relation between the inf-sup-conditions and the invertibility of the operator associated with the bilinear form:

Theorem 2.1.1 (Well-posed operator equation)**.** *Let X and Y be two reflexive Banach spaces and $A \in \mathcal{L}(X, Y')$ the bounded linear operator associated with the bilinear form in* (2.4)*. Then $A^{-1} \in \mathcal{L}(Y', X)$ with $\|A^{-1}\|_{\mathcal{L}(Y',X)} \leqslant \kappa^{-1}$ if and only if* (2.4a) *and* (2.4b) *hold. In this case, for any $f \in Y'$, the operator equation $Ax = f$ (or equivalently the weak problem, see* (2.3)*) has a unique solution $x \in X$, which satisfies the a-priori bound*

$$\|x\|_X \leqslant \frac{\|f\|_{Y'}}{\kappa}$$

2.2 Derivatives in function spaces

Let $F : X \subset \mathcal{X} \to \mathcal{Y}$ be an operator between Banach spaces $\mathcal{X}, \mathcal{Y}$, and $X \neq \varnothing$ open.

(i) F is (Gâteaux) directionally differentiable at $x \in X$ if the limit

$$\mathrm{d}F(x,h) = \lim_{t \searrow 0} \frac{F(x+th) - F(x)}{t} \in \mathcal{Y}$$

exists for all $h \in \mathcal{X}$. In this case, $\mathrm{d}F(x,h)$ is called directional derivative of F at x in the direction h.

(ii) F is Gâteaux differentiable at $x \in X$ if F is directionally differentiable at x and the directional derivative $F'(x) : \mathcal{X} \ni h \mapsto \mathrm{d}F(x,h) \in \mathcal{Y}$ is bounded and linear, i.e., $F'(x) \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$.

(iii) F is Hadamard directionally differentiable at $x \in X$ if the limit

$$F'(x,h) = \lim_{\substack{t \searrow 0 \\ h' \to h}} \frac{F(x+th') - F(x)}{t} \in \mathcal{Y}$$

exists for all $h \in \mathcal{X}$.

(iv) F is Fréchet differentiable at $x \in X$ if F is Gâteaux differentiable at x and if the following approximation condition holds:

$$\lim_{\|h\|_{\mathcal{X}} \to 0} \frac{1}{\|h\|_{\mathcal{X}}} \|F(x+h) - F(x) - F'(x)h\|_{\mathcal{Y}} = 0\,.$$

(v) If F is directionally/Gâteaux/Hadamard/Fréchet differentiable at every $x \in \widetilde{X}$, $\widetilde{X} \subset X$ open, then F is called directionally/Gâteaux/Hadamard/Fréchet differentiable on $\widetilde{X}$.

Higher derivatives are defined as follows: Let F be Gâteaux differentiable in a neighborhood X of x and $F' : X \to \mathcal{L}(\mathcal{X}, \mathcal{Y})$ is itself Gâteaux differentiable at x, then F is called twice Gâteaux differentiable at x. We denote the second Gâteaux derivative of F at x by $F''(x) \in \mathcal{L}(\mathcal{X}, \mathcal{L}(\mathcal{X}, \mathcal{Y}))$. Analogously, one defined the k-th order Gâteaux derivative, as well as the k-th order Fréchet derivative.

Proposition 2.2.1 (See [89] and [150])**.** *Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be Banach spaces and $F : X \subset \mathcal{X} \to \mathcal{Y}$, $G : Y \subset \mathcal{Y} \to \mathcal{Z}$, and $X, Y \neq \varnothing$ open.*

(i) For locally Lipschitz mappings in normed spaces, Hadamard and (Gâteaux) directional derivatives are equivalent.

(ii) Let F be Hadamard (Gâteaux) directionally differentiable at $x \in X$, and let G be Hadamard directionally differentiable at $y = F(x)$. Then, the composite mapping $G \circ F$ is Hadamard (Gâteaux) directionally differentiable at $x \in X$ and the chain rule holds:

$$\partial_x (G \circ F)(x,h) = \partial_y G \circ \partial_x F\,,$$

where ∂ denotes the Hadamard and (Gâteaux) directional derivatives, respectively.

(iii) *Let F be Fréchet (Gâteaux) differentiable at x and let G be Fréchet differentiable at $F(x)$. Then, $G \circ F$ is Fréchet (Gâteaux) differentiable at $x \in X$ and the chain rule holds:*

$$(G \circ F)'(x) = G'(F(x))F'(x)\,.$$

(iv) *If $H : X \times Y \to \mathcal{Z}$ is Fréchet differentiable at $(x, y) \in X \times Y$, then $F(\cdot, y)$ and $F(x, \cdot)$ are Fréchet differentiable at x and y, respectively. These derivatives are called partial derivatives of F with respect to x and y, and denoted by $\partial_x F(x, y)$ and $\partial_y F(x, y)$, respectively. Moreover, it holds that*

$$F'(x, y)(h_x, h_y) = \partial_x F(x, y)h_x + \partial_y F(x, y)h_y\,,$$

for $(h_x, h_y) \in \mathcal{X} \times \mathcal{Y}$.

(v) *If F is Gâteaux differentiable on a neighborhood of $x \in X$ and F' is continuous at x then F is Fréchet differentiable at x.*

(vi) *If F is Gâteaux differentiable in a neighborhood $\tilde{X}$ of x, then for all $h \in \mathcal{X}$ with $\{x + th : t \in [0, 1]\} \subset \tilde{X}$, the following holds:*

$$\|F(x + h) - F(x)\|_{\mathcal{Y}} \leqslant \sup_{t\in(0,1)} \|F'(x + th)h\|_{\mathcal{Y}}\,.$$

2.3 Parametric operator equations

As mentioned briefly in the preceding section, many differential equations can be formulated as variational problems, or equivalently as operator equations. The optimization problems that will be studied later in this manuscript, are subject to such operator equations. Uncertainties, which enter these problems through the operator equation constraints, are typically parameterized, resulting in so-called parametric operator equations as constraints. We start this section by introducing some multiindex notation.

Multiindex notation Here and in the following we will use the following multiindex notation: for a multiindex $\boldsymbol{\nu} = (\nu_j)_{j\in\mathbb{N}}$ with $\nu_j \in \mathbb{N}_0$, where $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$, we denote its order $|\boldsymbol{\nu}| := \sum_{j\in\mathbb{N}} \nu_j$ and its support as $\operatorname{supp}(\boldsymbol{\nu}) := \{j \in \mathbb{N} : \nu_j \geqslant 1\}$. Moreover, we denote the countable set of all finitely supported multiindices by

$$\mathcal{F} := \{\boldsymbol{\nu} \in \mathbb{N}_0^{\mathbb{N}} : |\operatorname{supp}(\boldsymbol{\nu})| < \infty\}\,.$$

Let $\boldsymbol{y} := (y_j)_{j\in\mathbb{N}}$ be a countably infinite sequence of real numbers taking values in a bounded domain $U \subset \mathbb{R}^{\mathbb{N}}$. Hereby we use the notational conventions

(i) For $\boldsymbol{m}, \boldsymbol{\nu} \in \mathcal{F}$ it holds $\boldsymbol{m} = \boldsymbol{\nu}$ if and only if $m_j = \nu_j$ for all $j \in \mathbb{N}$.

(ii) For $\boldsymbol{m}, \boldsymbol{\nu} \in \mathcal{F}$ it holds $\boldsymbol{m} \leqslant \boldsymbol{\nu}$ if and only if $m_j \leqslant \nu_j$ for all $j \in \mathbb{N}$.

(iii) For $\boldsymbol{m}, \boldsymbol{\nu} \in \mathcal{F}$ we define $\boldsymbol{m} + \boldsymbol{\nu} := (m_j + \nu_j)_{j\in\mathbb{N}}$ for all $j \in \mathbb{N}$.

(iv) For $\boldsymbol{m} \in \mathcal{F}$ we define $\boldsymbol{m}! := \prod_{j\in\mathbb{N}} m_j!$.

(v) For $\boldsymbol{m}, \boldsymbol{\nu} \in \mathcal{F}$ we define $\binom{\boldsymbol{\nu}}{\boldsymbol{m}} := \frac{\boldsymbol{\nu}!}{(\boldsymbol{\nu}-\boldsymbol{m})!\boldsymbol{m}!} = \prod_{j\in\mathbb{N}} \binom{\nu_j}{m_j}$.

(vi) $\boldsymbol{b}^{\boldsymbol{\nu}} := \prod_{j\in\mathbb{N}} b_j^{\nu_j}$, and $0^0 := 1$.

(vii) $\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} := \prod_{j\in\mathrm{supp}(\boldsymbol{\nu})} \frac{\partial^{\nu_j}}{\partial y_j^{\nu_j}}$.

In particular, in a Banach space X the ν_j-th partial (Gâteaux directional) derivative of an X-valued function $u(\boldsymbol{y}) : U \to X$, that depends on countably many parameters $\boldsymbol{y} \in U$, is defined as

$$\partial_{y_j}^{\nu_j} u(\boldsymbol{y}) = \lim_{t\searrow 0} \frac{u(\boldsymbol{y} + t\boldsymbol{e}_j) - u(\boldsymbol{y})}{t} \in X,$$

where $\boldsymbol{e}_j := (0, \ldots, 0, 1, 0, \ldots)$ has value 1 in the j-th component and 0 otherwise.
We are interested in parametric families of bounded linear operators $\{A(\boldsymbol{y}) \in \mathcal{L}(X, Y') : y \in U\}$. Later each y_j will be a realization of an independent random variable, cf., Example 2.4.1 and Chapter 3 – Section 7.2.
The precise dependence of the operators $A(\cdot)$ on the parameter sequence $\boldsymbol{y}$ is crucial for our regularity and approximation results later in this manuscript. Therefore we require $A(\boldsymbol{y})$ to be real analytic. Recall that a real analytic function is infinitely differentiable and coincides, in an open, nonempty neighborhood of each point, with its Taylor series about that point. This is detailed in the following, see also [111].

Assumption 2.3.1. *The parametric operator family $\{A(\boldsymbol{y}) \in \mathcal{L}(X, Y') : \boldsymbol{y} \in U\}$ is a regular p-analytic operator family for some $0 < p \leqslant 1$, that is*

(i) The operator $A(\boldsymbol{y})$ is invertible for every $\boldsymbol{y} \in U$ with uniformly bounded inverse $A^{-1}(\boldsymbol{y}) \in \mathcal{L}(Y', X)$, i.e., there exists $C > 0$ such that

$$\sup_{\boldsymbol{y}\in U} \|A(\boldsymbol{y})^{-1}\|_{\mathcal{L}(Y',X)} \leqslant C\,.$$

(ii) For each $\boldsymbol{y} \in U$, the operator $A(\boldsymbol{y})$ is a real analytic function with respect to $\boldsymbol{y}$. Precisely, this means there exists a nonnegative sequence $\boldsymbol{b} = (b_j)_{j\in\mathbb{N}} \in \ell^p(\mathbb{N})$[6] such that for all $\boldsymbol{\nu} \in \mathcal{F}\backslash\{0\}$ it holds that

$$\sup_{\boldsymbol{y}\in U} \|A(\boldsymbol{0})^{-1} \partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} A(\boldsymbol{y})\|_{\mathcal{L}(X,X)} \leqslant C\boldsymbol{b}^{\boldsymbol{\nu}} \tag{2.5}$$

for the same C as in (i).

Affine parameter dependence The case in which the operator depends affine on the parameters is well studied in the literature. This dependence structure arises for instance in diffusion problems with diffusion coefficients parameterized in terms of a Karhunen-Loève expansion. The operator $A(\boldsymbol{y})$ can then be written in terms of a family of operators $(A_j)_{j\in\mathbb{N}_0}$ such that

$$A(\boldsymbol{y}) = A_0 + \sum_{j\in\mathbb{N}} y_j A_j \quad \forall \boldsymbol{y} \in U\,. \tag{2.6}$$

We will now present conditions under which the operator family (2.6) satisfies Assumption 2.3.1.

[6] Here and in the following, for $0 < p < \infty$, we denote sequence spaces by $\ell^p(\mathbb{N}) := \{(x_j)_{j\in\mathbb{N}} : \sum_{j\geqslant 1} |x_j|^p < \infty\}$. Defining $\|(x_j)_{j\in\mathbb{N}}\|_{\ell^p(\mathbb{N})}^p = \sum_{j\geqslant 1} |x_j|^p$ for $p \in [1, \infty)$ and $\|(x_j)_{j\in\mathbb{N}}\|_{\ell^p(\mathbb{N})}^p = \sup_{j\in\mathbb{N}} |x_j|$ for $p = \infty$, the $\ell^p(\mathbb{N})$ spaces are Banach spaces which satify $\ell^p(\mathbb{N}) \subset \ell^q(\mathbb{N})$ for $1 \leqslant p < q \leqslant \infty$ and the duality $\ell^p(\mathbb{N})' = \ell^q(\mathbb{N})$ for $1 < p, q < \infty$ with $p^{-1} + q^{-1} = 1$ as well as $\ell^1(\mathbb{N})' = \ell^\infty(\mathbb{N})$. Moreover, $\ell^p(\mathbb{N})$ is reflexive for $1 < p < \infty$, separable for $1 \leqslant p < \infty$, and a Hilbert space for $p = 2$.

Assumption 2.3.2. *The operator family* $(A_j)_{j\in\mathbb{N}_0}$ *in* (2.6) *satisfies*

(i) The bilinear form associated with $A_0 \in \mathcal{L}(X, Y')$ *satisfies the inf-sup-conditions* (2.4) *with constant* $\gamma_0 > 0$.

(ii) The operators $(A_j)_{j\in\mathbb{N}}$ *are small with respect to* A_0 *in the following sense: There exists a constant* $0 < \kappa < 1$ *such that*

$$\sum_{j\in\mathbb{N}} \|A_0^{-1} A_j\|_{\mathcal{L}(X)} \leqslant \kappa\,.$$

Theorem 2.3.3. *Under Assumption 2.3.2, for each* $\boldsymbol{y} \in U$, *the parametric operator* $A(\boldsymbol{y})$ *satisfies the inf-sup-conditions* (2.4) *with* $\gamma = (1-\kappa)\gamma_0 > 0$.
In particular, for $f \in Y'$, *and for every* $\boldsymbol{y} \in U$, *the parametric operator equation*

$$A(\boldsymbol{y})u(\boldsymbol{y}) = f$$

admits a unique solution $u(\boldsymbol{y})$ *which satisfies the a-priori bound*

$$\sup_{\boldsymbol{y}\in U} \|u(\boldsymbol{y})\|_X \leqslant \frac{\|f\|_{Y'}}{\gamma}\,. \tag{2.7}$$

Proof. [111, Theorem 2] □

Corollary 2.3.4. *The affine parametric operator family* $(A_j)_{j\in\mathbb{N}_0}$ *in* (2.6) *satisfies Assumption 2.3.1 with* $p = 1$ *and*

$$C = \frac{1}{(1-\kappa)\gamma_0} \quad \text{and} \quad b_j = \frac{\|A_j\|_{\mathcal{L}(X,Y')}}{(1-\kappa)\gamma_0} \quad \text{for all } j \geqslant 1\,.$$

The solution $u \in X$ of the operator equation $A(\boldsymbol{y})u(\boldsymbol{y}) = f$ clearly depends on $\boldsymbol{y}$. The precise dependence is studied in the next paragraph.

Analytic dependence of solutions We will now present a result on the regularity of the solution u of the parametric operator equation with respect to the parameters, which later allows us to prove a-priori estimates for approximation and integration of the solution $u(\boldsymbol{y})$ with respect to the parameters $\boldsymbol{y} \in U$. In fact, it can be shown that the dependence of $u(\boldsymbol{y})$ on the parameter sequence is analytic.

Theorem 2.3.5. *Let the parametric family of operators* $\{A(\boldsymbol{y}) \in \mathcal{L}(X, Y') : \boldsymbol{y} \in U\}$ *satisfy Assumption 2.3.1 for some* $0 < p \leqslant 1$. *Then, for* $f \in Y'$, *and every* $\boldsymbol{y} \in U$ *there exists a unique solution* $u(\boldsymbol{y}) \in X$ *of the parametric operator equation*

$$A(\boldsymbol{y})u(\boldsymbol{y}) = f \tag{2.8}$$

and the parametric solution family $u(\boldsymbol{y})$ *depends analytically on the parameters* $\boldsymbol{y} \in U$, *with partial derivatives satisfying*

$$\sup_{\boldsymbol{y}\in U} \|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} u(\boldsymbol{y})\|_X \leqslant C\|f\|_{Y'} |\boldsymbol{\nu}|! \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}}\,,$$

where $\boldsymbol{b}$ *is defined in* (2.5).

Proof. [111, Theorem 4]. □

Corollary 2.3.6. *The affine parametric operator family* $(A_j)_{j\in\mathbb{N}_0}$ *in* (2.6) *satisfies*

$$\sup_{\boldsymbol{y}\in U} \|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} u(\boldsymbol{y})\|_X \leqslant C\|f\|_{Y'}|\boldsymbol{\nu}|!\boldsymbol{b}^{\boldsymbol{\nu}}\,, \tag{2.9}$$

where C *and* $\boldsymbol{b}$ *are defined in Corollary 2.3.4.*

Proof. We prove the result by induction with respect to $|\boldsymbol{\nu}|$. If $|\boldsymbol{\nu}| = 0$, then $\boldsymbol{\nu} = \mathbf{0}$ and the result follows from Corollary 2.3.4 and the a-priori bound (5.13). Given any multiindex $\boldsymbol{\nu} \in \mathcal{F}$ with $|\boldsymbol{\nu}| \geqslant 1$, suppose the result holds for any multiindex of order $|\boldsymbol{\nu}| - 1$. For $\mathbf{0} \neq \boldsymbol{\nu} \in \mathcal{F}$ we take the partial derivative $\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}$ of (2.8). By the Leibniz product rule we get

$$\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} (\partial_{\boldsymbol{y}}^{\boldsymbol{m}} A(\boldsymbol{y}))(\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}-\boldsymbol{m}} u(\boldsymbol{y})) = 0\,.$$

Separating out the $\boldsymbol{m} = \mathbf{0}$ term, we obtain

$$A(\boldsymbol{y})(\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} u(\boldsymbol{y})) = -\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},\boldsymbol{m}\neq\mathbf{0}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} (\partial_{\boldsymbol{y}}^{\boldsymbol{m}} A(\boldsymbol{y}))(\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}-\boldsymbol{m}} u(\boldsymbol{y}))\,.$$

By Corollary 2.3.4 and taking the norm we get

$$\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} u(\boldsymbol{y})\|_X \leqslant C \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},\boldsymbol{m}\neq\mathbf{0}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \|\partial_{\boldsymbol{y}}^{\boldsymbol{m}} A(\boldsymbol{y})\|_{\mathcal{L}(X)} \|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}-\boldsymbol{m}} u(\boldsymbol{y})\|_X\,.$$

From (2.6) we infer that

$$\partial^{\boldsymbol{m}} A(\boldsymbol{y}) = \begin{cases} A(\boldsymbol{y}) & \text{if } \boldsymbol{m} = \mathbf{0}, \\ A_j & \text{if } \boldsymbol{m} = \boldsymbol{e}_j, \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

Corollary 2.3.4 again leads to

$$\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} u(\boldsymbol{y})\|_X \leqslant \sum_{j\geqslant 1} \nu_j b_j \|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}-\boldsymbol{e}_j} u(\boldsymbol{y})\|_X\,.$$

The induction hypothesis gives

$$\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} u(\boldsymbol{y})\|_X \leqslant \sum_{j\in\geqslant 1} \nu_j b_j C\|f\|_{Y'}|\boldsymbol{\nu}-\boldsymbol{e}_j|\boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{e}_j} = C\|f\|_{Y'}|\boldsymbol{\nu}|!\boldsymbol{b}^{\boldsymbol{\nu}}\,.$$

□

2.4 Measure and integration theory

In the previous section we specified the dependence of the solution of the parameterized operator equation on the parameter sequence. In this section we will introduce measures of risk, that associate to each set of outcomes a real number, which quantifies the cost of this particular outcome.
We start this section with some facts from measure theory. Let Ω be a set. Let $\mathcal{P}(\Omega)$ be the power set of Ω, i.e., the set of all subsets of Ω, including the set itself and the empty set. A collection of subsets $\Sigma \subset \mathcal{P}(\Omega)$ is called σ-algebra on Ω if

(i) $\Omega \in \Sigma$.

(ii) $\mathcal{A} \in \Sigma$ implies $\mathcal{A}^{\complement} := \Omega \backslash \mathcal{A} \in \Sigma$.

(iii) $(\mathcal{A}_j)_{j\in\mathbb{N}} \in \Sigma$ implies $\cup_{j\in\mathbb{N}} \mathcal{A}_j \in \Sigma$.

The pair (Ω, Σ) is called a measurable space, and elements of Σ are called measurable sets. A subset $\Sigma' \subset \Sigma$ which is a σ-algebra is called a sub-σ-algebra of Σ. An important example is the Borel-σ-algebra $\mathfrak{B}(\Omega)$ on a topological space Ω, which is the smallest σ-algebra that is generated by all open subsets of Ω. The Borel-σ-algebra on $\mathbb{R}^d$ is generated by $\{(a_1, b_1) \times \cdots \times (a_d, b_d) : a_j < b_j \text{ for all } j \in \{1, \ldots, d\}\}$, and similar if the open intervals are replaced by half-open or closed intervals.

Measure space A measure on the measurable space (Ω, Σ) is a mapping $\mu : \Sigma \to [0, \infty]$ with $\mu(\varnothing) = 0$ and

$$\mu\Big(\bigcup_{j\in\mathbb{N}} \mathcal{A}_j\Big) = \sum_{j\in\mathbb{N}} \mu(\mathcal{A}_j)\,, \tag{2.10}$$

for all sequences $(\mathcal{A}_j)_{j\in\mathbb{N}} \subset \Sigma$ of mutually disjoint sets $\mathcal{A}_n \cap \mathcal{A}_m = \varnothing$ for $n \neq m$. The triplet (Ω, Σ, μ) is called measure space. The measure space and the measure μ are called finite if $\mu(\Omega) < \infty$ and σ-finite if $\Omega = \bigcup_{j\in\mathbb{N}} \mathcal{A}_j$ with $\mathcal{A}_j \in \Sigma$, and $\mu(\mathcal{A}_j) < \infty$ for all $j \in \mathbb{N}$. An important exampe is the Lebesgue measure λ on $\mathbb{R}^d$, which is the unique translation-invariant measure on the Borel-σ-algebra on $\mathbb{R}^d$ with $\lambda((a_1, b_1] \times \cdots \times (a_d, b_d]) = (b_1 - a_1) \cdots (b_d - a_d)$.

Null sets Let (Ω, Σ, μ) be a measure space. A measurable set $\mathcal{A} \in \Sigma$ is called a (μ-)null set if $\mu(\mathcal{A}) = 0$. A property depending on $x \in \Omega$ is said to hold μ-almost everywhere (μ-a.e.) or for μ-almost every $x \in \Omega$ (μ-a.e. $x \in \Omega$) if there is a null set $\mathcal{A} \in \Sigma$ such that the property holds for all $x \in \Omega \backslash \mathcal{A}$.
The measure space (Ω, Σ, μ) is called complete if every subset of any null set is measurable, i.e., if for all $\mathcal{A}' \in \mathcal{P}(\Omega)$ with $\mathcal{A}' \subset \mathcal{A}$ for $\mathcal{A} \in \Sigma$ with $\mu(\mathcal{A}) = 0$, it holds that $\mathcal{A}' \in \Sigma$.

Measurability For two measurable spaces $(\Omega, \Sigma), (\Omega', \Sigma')$, a function $f : \Omega \to \Omega'$ is called Σ-Σ'-measureable (or just measurable if the corresponding σ-algebras are clear from the context) if $f^{-1}(\mathcal{A}') \in \Sigma$ for all $\mathcal{A}' \in \Sigma'$.
Let (Ω, Σ, μ) be a measure space and (Ω', Σ') be a measurable space. A measurable function $f : \Omega \to \Omega'$ defines a measure

$$\mu_f(\mathcal{A}') := \mu(f^{-1}(\mathcal{A}'))\,, \quad \mathcal{A}' \in \Sigma'\,, \tag{2.11}$$

on (Ω', Σ'), which is called the image measure of μ under f.

Probability space Let (Ω, Σ, μ) be a measure space. If $\mu(\Omega) = 1$, then μ is called a probability measure, and (Ω, Σ, μ) is called a probability space.
For a probability space (Ω, Σ, μ), and a measurable space (Ω', Σ'), a Σ-Σ'-measureable function $f : \Omega \to \Omega'$ is called a (Ω', Σ')-valued random variable.
Let D be a set. A function $f : \Omega \times D \to \Omega'$, such that $f(\cdot, x)$ is a random variable for each $x \in D$, is called a random field. In this work the set D will be a domain, i.e., a nonempty, connected, and open set in $\mathbb{R}^d$.

The image measure μ_f, see (2.11), of a probability measure μ on (Ω, Σ) under a measureable function $f : \Omega \to \Omega'$ (or random variable) is a probability measure on the image space (Ω', Σ'). In this case μ_f is called the (probability) distribution of f under μ.

Stochastic independence Let (Ω, Σ, μ) be a probability space, and let $(\Omega'_j, \Sigma'_j)_{j\in\{1,2\}}$ be two measurable spaces. Two random variables $(y_j)_{j\in\{1,2\}}$, with $y_j : (\Omega, \Sigma, \mu) \to (\Omega'_j, \Sigma'_j)$ for each $j \in \{1, 2\}$, are called independent, if for every $\mathcal{A}_j \in \Sigma'_j$, $j \in \{1, 2\}$, we have

$$\mu(\{x \in \Omega : y_1(x) \in \mathcal{A}_1 \text{ and } y_2(x) \in \mathcal{A}_2\}) = \prod_{j\in\{1,2\}} \mu(\{x \in \Omega : y_j(x) \in \mathcal{A}_j\}) .$$

As shorthand notation, we say that two random variables $(y_j)_{j\in\{1,2\}}$ are i.i.d (for independent and identically distributed) if $(y_j)_{j\in\{1,2\}}$ are independent and if the have the same distribution $\mu_{y_1} = \mu_{y_2}$.
The concept of stochastic independence can be readily generalized to finite families of random variables. The existence of infinitely many independent random variables can be shown using the results in the paragraph "Countably infinite product of probability spaces" below.

Lebesgue space Integration of a measurable function with respect to a measure is called Lebesgue integration. Let (Ω, Σ, μ) be a measure space. A function $f : \Omega \to \mathbb{K}$ is called simple if $f = \sum_{j=1}^{n} \alpha_j \mathbb{I}_{\mathcal{A}_j}$ for some scalars $(\alpha_j)_{j\in\{1,\dots n\}} \in \mathbb{K}$ and mutually disjoint measurable sets $(\mathcal{A}_j)_{j\in\{1,\dots,n\}} \in \Sigma$ with $\mu(\mathcal{A}_j) < \infty$ for all $j \in \{1, \dots, n\}$, where $\mathbb{I}_{\mathcal{A}}(x)$ is the indicator function which is 1 if $x \in \mathcal{A}$ for a set $\mathcal{A} \in \Sigma$ and 0 otherwise. The Lebesgue integral of a simple function is defined as

$$\int_\Omega f \,\mathrm{d}\mu := \sum_{j=1}^{n} \alpha_j \mu(\mathcal{A}_j) .$$

Then we can define the integral of nonnegative measurable functions $f : \Omega \to [0, \infty]$ by

$$\sup \left\{ \int_\Omega \phi \,\mathrm{d}\mu \,:\, \phi : \Omega \to \mathbb{R} \text{ is a simple function, and } 0 \leqslant \phi(x) \leqslant f(x) \text{ for } \mu\text{-a.e. } x \in \Omega \right\} .$$

Finally, the integral of a function $f : \Omega \to \mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ is defined through integration of positive and negative parts, and real and imaginary parts, respectively:

$$\int_\Omega f \,\mathrm{d}\mu := \int_\Omega f^+ \,\mathrm{d}\mu - \int_\Omega f^- \,\mathrm{d}\mu ,$$

where $f^+ := \max\{f, 0\}$ and $f^- := \max\{-f, 0\}$ if $\mathbb{K} = \mathbb{R}$ and

$$\int_\Omega f \,\mathrm{d}\mu := \int_\Omega \mathrm{Re}(f) \,\mathrm{d}\mu + i \int_\Omega \mathrm{Im}(f) \,\mathrm{d}\mu ,$$

if $\mathbb{K} = \mathbb{R}$. We say the function f is μ-integrable if $\int_\Omega |f| \,\mathrm{d}\mu < \infty$.
If two functions $f, g : \Omega \to \mathbb{K}$ are μ-almost everywhere identical, i.e., $f(x) = g(x)$ for μ-a.e. $x \in \Omega$, then integration with respect to the measure μ cannot distinguish between the functions f and g. (μ-)Almost everywhere equality hence defines an equivalence relation

among μ-measurable functions and we locally denote the equivalence class of f by $[f]_\mu$ to clarify the distinction between functions and equivalence classes.[7]
Let (Ω, Σ, μ) be a measure space. For $1 \leqslant p \leqslant \infty$ we define the Lebesgue space by

$$L^p(\Omega, \Sigma, \mu, \mathbb{K}) := \{[f]_\mu : \Omega \to \mathbb{K} : f \text{ is measurable and } \|f\|_{L^p} < \infty\}, \tag{2.12}$$

where $\|f\|_{L^p}$ is defined by

$$\|f\|_{L^p} := \left(\int_\Omega |f(x)|^p \, \mathrm{d}\mu \right)^{\frac{1}{p}}, \tag{2.13}$$

for $1 \leqslant p < \infty$, and by

$$\|f\|_{L^\infty} := \operatorname*{ess\,sup}_{x \in \Omega} |f(x)|. \tag{2.14}$$

The values of (2.13) and (2.14) coincide for functions which are equal μ-a.e., i.e., for all representatives of the equivalence class $[f]_\mu$. Hence, $L^p(\Omega, \Sigma, \mu, \mathbb{K})$ is well-defined by (2.12). Moreover, by the identification of μ-a.e. identical functions, the expressions (2.13) and (2.14) become norms. In the following we will always make this identification, and not make the distinction between functions f and equivalence classes explicit. Moreover, we will usually use the abbreviations $L^p(\Omega) = L^p_\mu(\Omega) = L^p(\Omega, \Sigma, \mu, \mathbb{K})$ when the omitted notation is clear from the context.
Some properties of the Lebesgue spaces are the following:

(i) For $1 \leqslant p \leqslant \infty$, the L^p spaces are Banach spaces.

(ii) For $1 < p < \infty$ the L^p spaces are reflexive, with $(L^p)' \cong L^q$ for $q^{-1} = 1 - p^{-1}$. Moreover, $(L^1)' \cong L^\infty$.

(iii) For $p = 2$ the L^p space is a Hilbert space.

(iv) Using Hölder's inequality one can show that on finite measure spaces (Ω, Σ, μ), $1 \leqslant p < q \leqslant \infty$ implies that $L^p \subset L^q$.

(v) Let (Ω, Σ) be separable, then L^p is separable for $1 \leqslant p < \infty$.

Lebesgue–Bochner space The solution of a parametric operator equation in general takes values in a Banach space. A generalization of the Lebesgue integral to integrals over Banach space-valued functions is the Bochner-integral, which is defined as follows:
Let (Ω, Σ, μ) be a σ-finite measure space and X a Banach space. A function $f : \Omega \to X$ is called simple if $f = \sum_{j=1}^n x_j \mathbb{I}_{\mathcal{A}_j}$ for some $(x_j)_{j \in \{1,\dots,n\}} \in X$ and mutually disjoint μ-measurable sets $(\mathcal{A}_j)_{j \in \{1,\dots,n\}} \in \Sigma$ with $\mu(\mathcal{A}_j) < \infty$ for all $j \in \{1, \dots, n\}$, where $\mathbb{I}_{\mathcal{A}}(x)$ is the indicator function which is 1 if $x \in \mathcal{A}$ for a set $\mathcal{A} \in \Sigma$ and 0 otherwise. The X-valued Bochner integral with respect to μ over measurable sets $\mathcal{A} \in \Sigma$ of the simple function f is defined as

$$\int_{\mathcal{A}} f \, \mathrm{d}\mu := \sum_{j=1}^n x_j \mu(\mathcal{A} \cap \mathcal{A}_j). \tag{2.15}$$

[7] A binary relation $\sim$ on a nonempty set A, which associates two elements $a \in A$ and $b \in A$, is a set of ordered pairs (a, b) and hence a subset of the Cartesian product $A \times A$. A binary relation is called equivalence relation if $(a, a) \in \sim$ for all $a \in A$, and $(a, b) \in \sim$ implies $(b, a) \in \sim$, and $(a, b) \in \sim$ and $(b, c) \in \sim$ implies $(a, c) \in \sim$ for all $a, b, c \in A$. We define the subset $[a]_\sim := \{b \in A : (b, a) \in \sim\}$ and call it the equivalence class of a.

A function $f : \Omega \to X$ is called Bochner-measurable (or strongly μ-measurable) if there exists a sequence $(f_j)_{j\in\mathbb{N}}$ of simple functions such that $\lim_{j\to\infty} f_j(x) = f(x)$ for μ-a.e. $x \in \Omega$. If the approximating sequence $(f_j)_{j\in\mathbb{N}}$ of simple functions satisfies

$$\lim_{j\to\infty} \int_\Omega \|f - f_j\|_X \,\mathrm{d}\mu = 0\,,$$

then f is called Bocher-integrable and the Bochner integral is defined by

$$\int_\Omega f \,\mathrm{d}\mu = \lim_{j\to\infty} \int_\Omega f_j \,\mathrm{d}\mu\,.$$

Note that the limit is independent of the approximating sequence.
In order to ensure the ability to approximate every element $f \in X$ of a Banach space X by a countable family $(f_j)_{j\in\mathbb{N}}$, we will in the following chapters mostly assume that the Banach space X is separable.
A Bochner measurable function $f : \Omega \to X$ is Bochner integrable if and only if the function $\|f\|_X : \Omega \to \mathbb{R}$ is μ-integrable and it holds, [90, Proposition 1.2.2]

$$\left\| \int_\Omega f \,\mathrm{d}\mu \right\|_X \leqslant \int_\Omega \|f\|_X \,\mathrm{d}\mu\,.$$

Moreover, for an operator $A \in \mathcal{L}(X, Y)$ between two Banach spaces X, Y, and a Bochner integrable function $f : \Omega \to X$, Af is a Y-valued Bochner integrable function, and

$$\int_\Omega Af \,\mathrm{d}\mu = A \int_\Omega f \,\mathrm{d}\mu\,. \tag{2.16}$$

Almost everywhere equality again defines an equivalence relation among strongly μ-measurable functions. In the following definition, let $[f]_\mu$ denote again the equivalence class of the function f.
For $0 < p \leqslant \infty$ the Lebesgue–Bochner space is

$$L^p(\Omega, \Sigma, \mu, X) := \{[f]_\mu : \Omega \to \mathbb{K} \,:\, f \text{ is strongly measurable and } \|f\|_{L^p(\Omega,X)} < \infty\}\,,$$

where

$$\|f\|_{L^p(\Omega,X)} := \left(\int_\Omega \|f\|_X^p \,\mathrm{d}\mu \right)^{\frac{1}{p}}$$

for $0 < p < \infty$ and

$$\|f\|_{L^\infty(\Omega,X)} := \operatorname*{ess\,sup}_{x\in\Omega} \|f(x)\|_X\,.$$

We will usually use the abbreviations $L^p(\Omega, X) = L^p_\mu(\Omega, X) = L^p(\Omega, \Sigma, \mu, X)$ when the omitted notation is clear from the context.
Some properties of the Lebesgue spaces are the following:

(i) For $1 \leqslant p \leqslant \infty$, the $L^p(\Omega, X)$ spaces are Banach spaces.

(ii) For $1 < p < \infty$ the $L^p(\Omega, X)$ spaces are reflexive if the underlying Banach spaces X are reflexive or X' is separable, with $L^p(\Omega, X)' \cong L^q(\Omega, X')$ for $q^{-1} = 1 - p^{-1}$. Moreover, $L^1(\Omega, X)' \cong L^\infty(\Omega, X')$ if μ is σ-finite, see [90, Corollary 1.3.22].

(iii) For $p = 2$ the $L^p(\Omega, X)$ space is a Hilbert space.

(iv) Using Hölder's inequality one can show that on finite measure spaces (Ω, Σ, μ), for $1 \leqslant p < q \leqslant \infty$ implies that $L^p(\Omega, X) \subset L^q(\Omega, X)$.

Let (Ω, Σ, μ) be a measure space, let $1 \leqslant p < \infty$, and let X be a Banach space. If $\dim(L^p(\Omega, X)) \geqslant 1$, the following assertions are equivalent, see [90, Proposition 1.2.29]:

(i) $L^p(\Omega, X)$ is separable.

(ii) The space X is separable and there is a disjoint decomposition $\Omega = \Omega_1 \cup \Omega_2$ in Σ such that $\mu|_{\Omega_1}(\mathcal{A}) \in \{0, \infty\}$ for all $\mathcal{A} \in \Sigma|_{\Omega_1}$ and $(\Omega_2, \Sigma|_{\Omega_2}, \mu|_{\Omega_2})$ is μ-countably generated. That is, there is a sequence $(\Omega_j)_{j\geqslant 1}$ in Σ, consisting of sets with finite measure, such that for all $\mathcal{A} \in \Sigma$ there is a set $\mathcal{A}'$ in the σ-algebra that is generated by $(\Omega_j)_{j\geqslant 1}$ such that $\mu((\mathcal{A}\backslash\mathcal{A}') \cup (\mathcal{A}'\backslash\mathcal{A})) = 0$.[8]

Pettis integral and weak measurability In spaces that are not separable, like the Banach space of bounded, measurable functions $L^\infty(D)$, the notion of strong measurability and Bochner integrability is not immediately available. Let (Ω, Σ, μ) be a measure space and X a Banach space. A function $f : \Omega \to X$ is called weakly Σ-measurable if, for every $\phi \in X'$, the mapping $\phi(f) : \Omega \to \mathbb{R}$ is Σ-measurable. The mapping f is weakly μ-integrable if for all $\phi \in X'$, the mapping $\phi(f) : \Omega \to \mathbb{R}$ is μ-integrable. For a separable[9] Banach space X, the notions of strong μ-measurability, weak Σ-measurability coincide.
For a fixed, weakly μ-integrable $f : \Omega \to X$, we define $S_f : X' \to L^1_\mu(\Omega) : \phi \mapsto \phi(f)$. Using the closed graph theorem it can be shown that S_f is bounded, see [90, Lemma 1.2.18]. The dual of S_f is called the Dunford operator $T_f := S'_f : (L^1_\mu(\Omega))' \to X''$. By identifying $(L^1_\mu(\Omega))'$ with $L^\infty_\mu(\Omega)$, we get for every $g \in L^\infty_\mu(\Omega)$ and every $\phi \in X'$ that

$$\langle \phi, T_f g\rangle = \langle S_f \phi, g\rangle = \int_\Omega (\phi(f))g \,\mathrm{d}\mu\,.$$

Applying T_f to the indicator function $\mathbb{I}_\mathcal{A}$, we can define the so-called Dunford Integral of f over $\mathcal{A} \in \Sigma$ as

$$\int_\mathcal{A} f \,\mathrm{d}\mu := T_f \mathbb{I}_\mathcal{A} \in X''\,, \quad \mathcal{A} \in \Sigma\,, \tag{2.17}$$

Note that the Dunford integral is in general an element of X''. In view of the canonical embedding $X \subset X''$, we say that a weakly μ-integrable function f is Pettis integrable if for every $\mathcal{A} \in \Sigma$ its Dunford integral $\int_\mathcal{A} f \,\mathrm{d}\mu$ belongs to X. In this case (2.17) is called Pettis integral of f over $\mathcal{A} \in \Sigma$ with the interpretation of elements of X'' as elements of X characterized by

$$\left\langle \phi, \int_\mathcal{A} f \mathrm{d}\mu \right\rangle = \int_\mathcal{A} \phi(f) \mathrm{d}\mu \quad \forall \phi \in X'\,. \tag{2.18}$$

[8] For $\mathcal{A} \in \Sigma$ we denote $\mathcal{A}|_\Sigma = \{\mathcal{A} \cap \mathcal{A}' : \mathcal{A}' \in \Sigma\} = \{\mathcal{A}' \in \Sigma : \mathcal{A}' \subseteq \Sigma\}$. The restriction of μ to $\Sigma_\mathcal{A}$ is denoted by $\mu|_\mathcal{A}$.

[9] The result can be stated more generally by replacing the assumption of a separable Banach space with the condition that there is a μ-null set $N \subset \Omega$ such that the image $f(\Omega\backslash N) \subset X$ is separable, since this is always true if X is a separable Banach space, see [90, Theorem 1.1.20]

Clearly the Dunford and Pettis integrals coincide if X is reflexive. Moreover, for a Bochner integrable $f : \Omega \to X$, and for an approximating sequence $(f_j)_{j\in\mathbb{N}} \subset X$ of simple functions, by (2.18) each f_j is Pettis integrable with Pettis integral of f_j over $\mathcal{A}$ given by (2.15). Taking the limit $j \to \infty$, we obtain that f is also Pettis integrable. Moreover, by the definition of Bochner integral, for this f the Bochner and the Pettis integral coincide. Hence the Pettis integral is a consistent extension of the Bochner integral.

Product measures Let (Ω_1, Σ_1) and (Ω_2, Σ_2) be two measurable spaces. The so-called product σ-algebra $\Sigma_1 \otimes \Sigma_2$ on $\Omega_1 \times \Omega_2$ is the σ-algebra generated by the boxed $\mathcal{A}_1 \times \mathcal{A}_2$ with $\mathcal{A}_1 \in \Sigma_1$ and $\mathcal{A}_2 \in \Sigma_2$. For two measures μ_1 on (Ω_1, Σ_1) and μ_2 on (Ω_2, Σ_2) there exists a unique measure $\mu_1 \otimes \mu_2$ on $(\Omega_1 \times \Omega_2, \Sigma_1 \otimes \Sigma_2)$, called the product measure, such that

$$\mu_1 \otimes \mu_2(\mathcal{A}_1 \times \mathcal{A}_2) = \mu_1(\mathcal{A}_1)\mu_2(\mathcal{A}_2) \quad \forall \mathcal{A}_1 \in \Sigma_1, \forall \mathcal{A}_2 \in \Sigma_2 \,.$$

Fubini's theorem An important result about the product of two probability spaces is Fubini's theorem: Let $(\Omega_1, \Sigma_1, \mu_1)$ and $(\Omega_2, \Sigma_2, \mu_2)$ be two σ-finite measure spaces. Let X be a Banach space, and let $f : \Omega_1 \times \Omega_2 \to X$ be $\mu_1 \otimes \mu_2$-Bochner integrable. Then $f(x_1, \cdot)$ is μ_2-Bochner integrable for μ_1-a.e. $x_1 \in \Omega_1$ and $x_1 \mapsto \int_{\Omega_2} f(x_1, \cdot)\,\mathrm{d}\mu_2$ is μ_1-Bochner integrable, see [90, Proposition 1.2.7]. Analogously, $f(\cdot, x_2)$ is μ_1-Bochner integrable for μ_2-a.e. $x_2 \in \Omega_2$ and $x_2 \mapsto \int_{\Omega_1} f(\cdot, x_2)\,\mathrm{d}\mu_1$ is μ_2-Bochner integrable. In particular we have

$$\int_{\Omega_1\times\Omega_2} f \,\mathrm{d}\mu_1 \otimes \mu_2 = \int_{\Omega_1} \left(\int_{\Omega_2} f(x_1, \cdot)\,\mathrm{d}\mu_2 \right) \mathrm{d}\mu_1(x_1) = \int_{\Omega_2} \left(\int_{\Omega_1} f(\cdot, x_2)\,\mathrm{d}\mu_1 \right) \mathrm{d}\mu_2(x_2) \,.$$

Fubini's theorem holds also for Lebesgue integrable functions in the case $X = \mathbb{K}$.

Countably infinite product of probability spaces Let $(\Omega_j, \Sigma_j, \mu_j)_{j\in\mathbb{N}}$ be a countably infinite family of probability spaces, and let (Ω, Σ) be the product of the measurable spaces $(\Omega_j, \Sigma_j)_{j\in\mathbb{N}}$. That is, Ω is the set of all sequences $(x_j)_{j\in\mathbb{N}}$ such that $x_j \in \Omega_j$ for each $j \in \mathbb{N}$.[10] Moreover, for each j we define the coordinate projection $y_j : \Omega \to \Omega_j$ by $y_j(x) := x_j$, then the product σ-algebra Σ is the smallest σ-algebra on Ω that makes all these coordinate projections measurable, i.e., $\Sigma = \sigma(\{y_j^{-1}(\mathcal{A}_j), j \in \mathbb{N}\})$. Moreover, there is a unique probability measure μ on (Ω, Σ) such that

$$\mu\Big(\mathcal{A}_1 \times \cdots \mathcal{A}_n \times \bigtimes_{j=n+1}^{\infty} \Omega_j\Big) = \prod_{k=1}^{n} \mu_k(\mathcal{A}_k)$$

for $\mathcal{A}_k \in \Sigma_k$, $k = 1, \ldots, n$ and $n \in \mathbb{N}_0$. The probability measure μ is called the product of the measures $(\mu_j)_{j\in\mathbb{N}}$ and we write $\mu = \otimes_{j\in\mathbb{N}}\mu_j$. Furthermore, for each $j \in \mathbb{N}$ the distribution of y_j is μ_j and the random variables $(y_j)_{j\in\mathbb{N}}$ are independent. A proof can be found, e.g., in [32, Proposition 10.6.1], or [81, Sec. 38, Theorem B]. This result is a corollary to Ionescu-Tulcea's theorem, see [104, Corollary 14.33]. Generalizations to uncountable index sets are Kolmogorov's extension theorem [104, Theorem 14.13] and the

[10] For general index sets $\mathfrak{J}$ one defines the Cartesian product $\Omega = \times_{j\in\mathfrak{J}} \Omega_j$ as the set of maps $x : \mathfrak{J} \to \bigcup_{j\in\mathfrak{J}} \Omega_j$ such that $x(j) \in \Omega_j$ for all $j \in \mathfrak{J}$.

more general Andersen–Jessen theorem, see [145, Chapter 10.6]. Note that these results prove the existence of infinitely many independent random variables.
It can then be readily verified that if an arbitrary set of indices $\mathfrak{J}$ is split into two disjoint parts $\mathfrak{J}_1$ and $\mathfrak{J}_2$ giving $\mu_1 = \otimes_{j\in\mathfrak{J}_1}\mu_j$ and $\mu_2 = \otimes_{j\in\mathfrak{J}_2}\mu_j$, then $\mu_1 \otimes \mu_2 = \otimes_{j\in\mathfrak{J}}\mu_j$, see [13, Chapter 3.5].

Example 2.4.1. *Let $\Omega_j = [-1,1]$, and let μ_j be a probability measure on $(\Omega_j, \mathfrak{B}(\Omega_j))$. Clearly $([-1,1], |\cdot|)$ is a separable[11] metric space. Hence, by the more general result [13, Lemma 6.4.2.] we have $\mathfrak{B}(\times_{j\geqslant 1}\Omega_j) = \otimes_{j\geqslant 1}\mathfrak{B}(\Omega_j)$ and the countable product of the measure spaces $(\Omega_j, \mathfrak{B}(\Omega_j))_{j\in\mathbb{N}}$ is given by $([-1,1]^{\mathbb{N}}, \mathfrak{B}([-1,1]^{\mathbb{N}}))$, where $[-1,1]^{\mathbb{N}} = \times_{j\in\mathbb{N}}\Omega_j$, see also [13, Example 7.6.1.]. Together with the product measure $\mu = \otimes_{j\in\mathbb{N}}\mu_j$ we define the product probability space*

$$([-1,1]^{\mathbb{N}}, \mathfrak{B}([-1,1]^{\mathbb{N}}), \mu)\,. \tag{2.19}$$

We usually deal with continuous functions that have compact domain. Such functions are Bochner-integrable, see [165, Lemma A.1.5]. Let X be a Banach space, and let $(\Omega, \mathfrak{B}, \mu)$ be a finite measure space, where Ω is a compact topological space. Then any continuous $f : (\Omega, \mathfrak{B}) \to (X, \mathfrak{B}(X))$ is μ-Bochner-integrable.

Example 2.4.2. *Let X be a Banach space. Observe that $[-1,1]$ is compact and thus, by Tychonov's theorem [135, Theorem A3], the Cartesian product $[-1,1]^{\mathbb{N}} = \times_{j\in\mathbb{N}}[-1,1]$ is compact with respect to the product topology. By [165, Lemma A.1.5], any continuous function $f : ([-1,1]^{\mathbb{N}}, \mathfrak{B}([-1,1]^{\mathbb{N}})) \to (X, \mathfrak{B}(X))$ is μ-Bochner integrable, where $\mu = \otimes_{j\in\mathbb{N}}$ is the probability measure in* (2.19).

Vector measure A vector measure is the natural extension to Banach space-valued measures: Let (Ω, Σ) be a measurable space and let X be a Banach space. Then a vector measure is a mapping $\lambda : \Sigma \to X$, that satisfies $\lambda(\varnothing) = 0$ and

$$\lambda(\cup_{j\in\mathbb{N}}\mathcal{A}_n) = \sum_{j\in\mathbb{N}} \lambda(\mathcal{A}_j)\,,$$

for all sequences $(\mathcal{A}_j)_{j\in\mathbb{N}} \subset \Sigma$ of mutually disjoint sets $\mathcal{A}_n \cap \mathcal{A}_m = \varnothing$ for $n \neq m$. Note that the convergence of the series on the right-hand side is understood to be in the norm of the Banach space X.
A vector measure λ is said to have bounded variation if

$$|\lambda|(\mathcal{A}) = \sup\Big\{ \sum_{i=1}^{n} \|\lambda(\mathcal{A}_i)\|_X : \{\mathcal{A}_1, \ldots, \mathcal{A}_n\} \text{ a partition of } \mathcal{A}\Big\} < \infty$$

for each $\mathcal{A} \in \Sigma$, where a partition of $\mathcal{A}$ is a finite set of mutually disjoint measurable sets whose union is $\mathcal{A}$.
The vector measure $\lambda : \Sigma \to X$ is said to be absolutely continuous with respect to the finite measure μ on Σ if and only if $\lambda(\mathcal{E})$ implies $\mu(\mathcal{A}) = 0$ for all $\mathcal{A} \in \Sigma$.
A Banach space X is said to have the Radon–Nikodým property if the following result is valid in the Banach space X:

[11] Indeed the countable set $\mathbb{Q} \cap [0,1]$ is a dense subset of $[0,1]$.

Radon–Nikodým property Let X be a Banach space, (Ω, Σ) a measurable space equipped with a finite measure μ, i.e., (Ω, Σ, μ) is a finite measure space. Let λ be a vector measure of bounded variation, that is absolutely continuous with respect to the measure μ. Then there exists a μ-Bochner integrable function $f : \Omega \to X$ such that

$$\lambda(\mathcal{A}) = \int_{\mathcal{A}} f \, \mathrm{d}\mu$$

for every $\mathcal{A} \in \Sigma$.
This result is not true in arbitrary Banach spaces, for a counterexample see [139, Example 5.15]. Using the notion of Radon–Nikodým property one can characterize reflexivity of the Lebegue–Bochner spaces. Let (Ω, Σ, μ) be a σ-finite measure space, let X be a Banach space, and let $1 \leqslant p < \infty$ with $p^{-1} + q^{-1} = 1$. The following assertions are equivalent, see [90, Theorem 1.3.10]:

- X' has the Radon–Nikodým property.
- The mapping $g \mapsto \phi_g$ from $L^q(\Omega, X') \to (L^p(\Omega, X))'$ defined by

$$\langle f, g \rangle := \int_\Omega \langle f, g \rangle \, \mathrm{d}\mu \quad f \in L^p(\Omega, X)$$

  establishes an isometric isomorphism between the Banach spaces $L^q(\Omega, X') \cong (L^p(\Omega, X))'$.[12]

Note that a Banach space X has the Radon–Nikodým property with respect to every σ-finite measure space, [90, Theorem 1.3.21] if X is reflexive or X is separable dual space.

Expected value Let (Ω, Σ, μ) be a probability space and X be a Banach space. The mean or expectation of a random variable $y : \Omega \to X$ is $m_y \in X$ such that

$$\phi(m_y) = \mathbb{E}[\phi(y)] = \int_\Omega \phi(y) \, \mathrm{d}\mu \quad \forall \phi \in X' \, .$$

If $\mathbb{E}[\|y\|_X] = \int_\Omega \|y\|_X \, \mathrm{d}\mu < \infty$, we conclude from $\phi(m_y) = \mathbb{E}[\phi(y)] = \phi(\mathbb{E}[y])$ and the Hahn–Banach theorem, that

$$m_y = \mathbb{E}[y] = \int_\Omega y \, \mathrm{d}\mu \in X \, .$$

Note that in Banach spaces the mean m_y of y is in general an element of the bidual space X'' of X, and given by $m_y(\phi) := \mathbb{E}[\phi(y)]$, $\phi \in X'$. The integrability assumption $\mathbb{E}[\|y\|_X] < \infty$ is sufficient but not necessary for the existence of m_y in X. For instance, if X is reflexive and $\mathbb{E}[\phi(y)] < \infty$ for all $\phi \in X'$, the Gelfand–Pettis theorem implies continuity of the linear map $\phi \mapsto \mathbb{E}[\phi(y)]$ on X', and hence $m_y \in X'' = X$, see [156, Chapter II.3].

[12] The assumption of σ-finite μ is only necessary for $p = 1$, i.e., can be dropped for $1 < p < \infty$, see [90, Corollary 1.3.22].

Moments Let (Ω, Σ, μ) be a probability space and let X be a Banach space. For $k \in \mathbb{N}$ the k-th moment of the random variable $y : \Omega \to X$ is the map $\mathcal{M}^k : \Omega^k \to \mathbb{R}$,

$$\mathcal{M}_y^k(\phi_1, \dots, \phi_k) := \mathbb{E}[\phi_1(y), \dots, \phi_k(y)] = \int_\Omega \phi_1(y) \cdots \phi_k(y) \, \mathrm{d}\mu \,.$$

If $\mathbb{E}[\|y\|_X^k] = \int_\Omega \|y\|_X^k \, \mathrm{d}\mu < \infty$, then μ^k is a conitinuous symmetric k-linear form[13] on X'. In particular, the covariance of y is $\mathrm{Cov}_y := \mathcal{M}_{y-m_y}^2$, i.e., for $\phi \in X'$

$$\mathrm{Cov}_y(\phi_1, \phi_2) = \mathbb{E}[\phi_1(y - m_y)\phi_2(y - m_y)] = \int_\Omega \phi_1(y - m_y)\phi_2(y - m_y) \, \mathrm{d}\mu \,.$$

If $\mathbb{E}[\|y\|_X^2] = \int_\Omega \|y\|_X^2 \, \mathrm{d}\mu < \infty$, then $\mathcal{M}_y^2$ and Cov_y are continuous, symmetric, positive bilinear forms on X', and

$$\mathrm{Cov}_y(\phi_1, \phi_2) = \mathcal{M}_y^2(\phi_1, \phi_2) - \phi_1(m_y)\phi_2(m_y) \quad \forall \phi_1, \phi_2 \in X' \,.$$

Proof. Continuity of $\mathcal{M}_y^k$ follows from

$$\begin{aligned} |\mathcal{M}_y^k(\phi_1, \dots, \phi_k)| &\leqslant \mathbb{E}[|\phi_1(y)| \cdots |\phi_k(y)|] \\ &\leqslant \mathbb{E}[\|y\|_X^k] \|\phi_1(y)\|_{X'} \cdots \|\phi_k(y)\|_{X'} \quad \forall \phi_1, \dots, \phi_k \in X' \,. \end{aligned}$$

Symmetry follows from $\mathcal{M}_y^k(\phi_{\pi(1)}, \dots, \phi_{\pi(k)}) = \mathcal{M}_y^k(\phi_1, \dots, \phi_k)$ for all permutations π of $\{1, \dots, k\}$. Moreover, for $k = 2$, positivity $\mathcal{M}_k^2(\phi, \phi) = \mathbb{E}[\phi(y)^2] \geqslant 0$ follows by definition. Since $\mathrm{Cov}_y := \mathcal{M}_{y-m_y}^2$ these properties hold also for the covariance. Moreover, for $\phi_1, \phi_2 \in X'$ we have

$$\begin{aligned} &\int_\Omega \phi_1(y - m_y)\phi_2(y - m_y) \, \mathrm{d}\mu \\ &= \int_\Omega \big(\phi_1(y) - \phi_1(m_y)\big)\big(\phi_2(y) - \phi_2(m_y)\big) \, \mathrm{d}\mu \\ &= \int_\Omega \phi_1(y)\big(\phi_2(y) - \phi_2(m_y)\big) \, \mathrm{d}\mu - \int_\Omega \phi_1(m_y)\big(\phi_2(y) - \phi_2(m_y)\big) \, \mathrm{d}\mu \\ &= \int_\Omega \phi_1(y)\phi_2(y) \, \mathrm{d}\mu - \int_\Omega \phi_1(y) \, \mathrm{d}\mu \, \phi_2(m_y) - \phi_1(m_y) \int_\Omega \phi_2(y) \, \mathrm{d}\mu + \phi_1(m_y)\phi_2(m_y) \\ &= \int_\Omega \phi_1(y)\phi_2(y) \, \mathrm{d}\mu - \phi_1(m_y)\phi_2(m_y) \,. \end{aligned}$$

□

Change of variables Let (Ω, Σ, μ) be a measure space, let (Ω', Σ') be a measurable space, and let X be a Banach space. Let $Y : \Omega \to \Omega'$ be measurable and let $u : \Omega' \to X$ be strongly measurable. Let $\mu_Y = \mu(Y^{-1})$ be the image measure of μ under Y. Then, see [90, Proposition 1.2.6], $f(\phi)$ is Bochner integrable with respect to μ if and only if f is Bochner integrable with respect to μ_Y, and then

$$\int_\Omega f(Y) \, \mathrm{d}\mu = \int_{\Omega'} f \, \mathrm{d}\mu_Y \,. \tag{2.20}$$

[13] A mapping $\mathcal{M}^k : \times_{j=1}^k \Omega_j \to \mathbb{K}$ is called k-linear form if it is linear in each argument, i.e., for all $\lambda \in \mathbb{K}$, for all $n \in \Omega_i$, and for all $i = 1, \dots, k$, it holds that $\mathcal{M}^k(m_1, \dots, \lambda m_i, \dots, m_k) = \lambda \mathcal{M}^k(m_1, \dots, m_i, \dots, m_k)$ and $\mathcal{M}^k(m_1, \dots, m_i + n, \dots, m_k) = \mathcal{M}^k(m_1, \dots, m_i, \dots, m_k) + \mathcal{M}^k(m_1, \dots, n, \dots, m_k)$.

3 A general formulation of optimal control problems under uncertainty

In this section we formulate the optimal control problem in a very general form and embed existing results from the literature into the general framework. We analyze the general problem under different sets of assumptions on the risk measure, the random variable cost functional, the constraint and the uncertainty.

The structure of this chapter is as follows. We discuss several risk measures and classify them according to their properties. Afterwards, we focus on the cost functional, which in the context of optimal control problems under uncertainty, is a random variable. In particular, we present conditions which ensure differentiability of the risk measure composed with the random variable cost functional.

Based on the results for the risk measure and the cost functional, we derive results about the existence and uniqueness of solutions to the optimal control problem under a set of assumptions about the constraint. Similarly, we can reformulate the problem in its so-called reduced form provided that the forward operator fulfills certain conditions. Finally, we are able to present optimality conditions for the general optimal control problem under uncertainty.

At the end of this chapter we briefly focus on parametric linear forward operators, which will play a central role in the remainder of this thesis. In particular, we establish a setting which allows us to replace the almost surely formulation of the constraint by the weak formulation in the uncertain parameters. This equivalence is used in Chapter 8 to derive one-shot methods based on a penalization of the model residual.

In the following sections we develop efficient methods for optimal control problems that have sufficiently high regularity with respect to the uncertain variables. We analyze the different error contributions of our methods and verify our theoretical findings in numerical experiments. In the last chapters we discuss further improvement of the solvers in terms of efficiency. To do so, we impose appropriate assumptions on the risk measure, the cost functional, the constraint, the uncertainty, and the regularization throughout this thesis.

However, most of the results presented in this section are not restricted to problems with high regulartiy with respect to the uncertain variables. In fact, we only make few assumptions about the structure or distribution of the uncertainty in this chapter. Moreover, we make few assumptions about the constraints, allowing for nonlinear constraints in many results. Furthermore, only a few assumptions are made on the regularization and we employ moderate assumptions on the random variable cost functional. Hence, the presented formulation covers a wide range of optimal control problems.

We start this chapter with the problem formulation and a detailed list of the most important components of the optimal control problems under uncertainty.

3.1 Problem formulation

Let $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ be Banach spaces, let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, let U be a compact topological space, and let $(U, \mathfrak{B}(U))$ be a separable metric space.
The optimal control problem under uncertainty has the following ingredients:

Uncertainty. The inherent randomness of the problem is described by a random variable $Y : \Omega \to U$. In many cases the random influence is parameterized, which leads to product probability spaces of the form $(\Omega, \Sigma, \mathbb{P}) = (\times_{j\in\mathbb{N}}\Omega_j, \times_{j\in\mathbb{N}}\Sigma_j, \otimes_{j\in\mathbb{N}}\mathbb{P}_j)$ and $(U, \mathfrak{B}(U), \mu) = (\times_{j\in\mathbb{N}}U_j, \mathfrak{B}(\times_{j\in\mathbb{N}}U_j), \otimes_{j\in\mathbb{N}}\mu_j)$. In this case the randomness is described by the countably infinite sequence of i.i.d. random variables $\boldsymbol{Y} = (Y_j)_{j\in\mathbb{N}} : \Omega_j \to U_j$ with realization $\boldsymbol{y} = (y_j)_{j\in\mathbb{N}} \in U$. Note that the realization of the random variable is usually denoted by $\boldsymbol{y} \in U$. The change of variables formula (2.20) allows us to work in the image space $(U, \mathfrak{B}(U))$ of the random variable Y equipped with the image measure μ. When dealing with the realizations $\boldsymbol{y} \in U$, we will oftentimes call them stochastic variables, parametric variables, or parameters.

Control. A fundamental component of an optimal control problem is the so-called control or control variable $z \in \mathcal{Z}$. The space $\mathcal{Z}$ is called the control space. The control z is a deterministic quantity, chosen by a controller and hence does not depend on the stochastic variables $\boldsymbol{y} \in U$.

State. The state or state variable u can be steered using the control $z \in \mathcal{Z}$. For a fixed control $z \in \mathcal{Z}$, the state maps the uncertainty $\boldsymbol{y} \in U$ into the state space $\mathcal{X}$. Hence the state is typically a random field in some Lebesgue–Bochner space $L^q(U, \mathfrak{B}(U), \mu, \mathcal{X})$ for $q \in [1, \infty]$. The dependence of the state $u \in \mathcal{X}$ on the uncertainty $\boldsymbol{y} \in U$ and the control $z \in \mathcal{Z}$ is described by the model.

Model. The underlying model is mathematically described by the map $\mathcal{M} : U \times \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}'$. In particular, the so-called model equation or constraint $\mathcal{M}(\boldsymbol{y}, u(\boldsymbol{y}), z) = 0$ relates the control to the state of the system. Moreover, for fixed control $z \in \mathcal{Z}$, the state of the system is subject to the randomness. In most of our applications the state $u \in \mathcal{X}$ depends linearly on the control $z \in \mathcal{Z}$ and analytically on the stochastic variables $\boldsymbol{y} \in U$.

Cost functional. The cost functional $J : U \times \mathcal{X} \to \mathbb{R}$ associates a nonnegative cost to each pair of $\boldsymbol{y} \in U$ and $u(\boldsymbol{y}) \in \mathcal{X}$. For $u : U \to \mathcal{X}$, we refer to the Nemytskii (superposition) operator

$$\mathcal{J}(u)(\boldsymbol{y}) = J(\boldsymbol{y}, u(\boldsymbol{y}))\,, \quad \boldsymbol{y} \in U \tag{3.1}$$

as the random variable objective function. The Nemytskii operator assigns a function $\mathcal{J}(u) : U \to \mathbb{R}$ to the state $u(\cdot) : U \to \mathcal{X}$. In order to account for the randomness the objective function is composed with a risk measure.

Risk measure. For a given state $u : U \to \mathcal{X}$, for $p \in [1, \infty)$, the risk measure $\mathcal{R} : L^p(U, \mathfrak{B}(U), \mu, \mathbb{R}) \to [0, \infty]$ associates a nonnegative real number to the random variable objective function $\mathcal{J}(u)(\cdot)$.

Regularization. For stability reasons we introduce a regularization $R(z) : \mathcal{Z} \to [0, \infty]$ on the control $z \in \mathcal{Z}$. In general a regularization can be viewed as a penalization of the control variable in order to ensure desirable properties of the control.

Control constraint. We also may impose additional direct constraints on the control $z \in \mathcal{Z}$, i.e., the controls are elements of the admissible set $\mathcal{Z}_{ad} \subseteq \mathcal{Z}$.

For $\alpha > 0$, we consider optimization problems of the general form

$$\min_{u \in L^q(\mathcal{X}), z \in \mathcal{Z}_{ad}} \mathcal{R}(\mathcal{J}(u)) + \alpha R(z) \tag{3.2}$$

subject to

$$M(\cdot, u(\cdot), z) = 0 \quad \text{in } L^q(\mathcal{Y}')\,, \tag{3.3}$$

for $1 = \frac{1}{q} + \frac{1}{r}$. We abbreviated $L^q(\mathcal{X}) = L^q(U, \mathfrak{B}(U), \mu, \mathcal{X})$ for $q \in [1, \infty]$, and analogously for r. Typically, it will be the case that $\mathcal{X} \subset \mathcal{Y}$ are reflexive Banach spaces.
This includes a wide range of optimal control problems, particularly very general constraints, various cost functionals, several risk measures, different input uncertainties, and various regularizations. We do not cover constraints on the states, which will be subject to future work.
Other possible formulations include the replacement of the random input quantity, e.g., with its mean. As a result, only one deterministic optimal control problem needs to be solved. However, the solution obtained with this approach is in general not robust with respect to the randomness of the problem. A further possible formulation of the problem is the path-wise solution of many optimal control problems, and then computing statistics of the obtained result afterwards. In general the obtained solution is not a solution of an optimal control problem and has limited meaning. Moreover, in some applications it can make sense to have a stochastic control, i.e., a control z which depends directly on $\boldsymbol{y} \in U$, see, e.g., [4, 24, 108, 111]. We do not consider such cases and rather suppose a controller is interested in a single deterministic control of the system. Thus we focus on the approach that leads to a robust solution of the optimal control problem (3.2) – (3.3) with respect to variations of the uncertain variables, cf., e.g., [75, 76, 77, 105, 106, 157, 158] and many others.

3.2 Risk measures

The presence of uncertainty or randomness in the optimization problem leads to a random variable objective function. Since there is no total order on random variables, one cannot minimize a random variable directly, but needs to define a meaningful order, as for example, by the introduction of a risk measure $\mathcal{R} : L^p(U, \mathfrak{B}(U), \mu, \mathbb{R}) \to \mathbb{R} \cup \{\infty\}$, that maps the random variable objective function to the extended real numbers. While the possible choices for $\mathcal{R}$ are virtually limitless, in this section we discuss sensible properties for measures of risk and derive optimality conditions. The proper choice of the risk measure and its desired properties may depend on the problem at hand.
Perhaps the most straightforward way to account for the risk in the optimization problem is considering the expected value $\mathcal{R} = \mathbb{E}$ as a risk measure, which is oftentimes referred to as the risk-neutral case. A solution, if it exists, optimizes the random outcome $J(u(\boldsymbol{y}), \boldsymbol{y})$ on average. This is justified, for instance, when one is interested in the long-term performance, irrespective of the fluctuations of specific outcome realizations. Based on this definition we characterize risk-aversion as follows: we say a risk measure $\mathcal{R} : L^p(U, \mathfrak{B}(U), \mu, \mathcal{R}) \to \mathbb{R} \cup \{\infty\}$ is called risk-averse (see, [136]) if

$$\mathcal{R}(X) > \mathbb{E}[X] \tag{3.4}$$

for all nonconstant $X \in L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$.
Since the concept of risk measures is not limited to the application in optimal control problems, in this section we use the generic notation $X \in L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$ for a random variable. However, one can think of X being the random variable objective function (3.1). A risk measure $\mathcal{R}$ is called regular if it satisfies (3.4) and if it

(D1) is proper, that is $\mathcal{R}(X) > -\infty$ for all $X \in L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$ and

$$\operatorname{dom}(\mathcal{R}) := \{X \in L^p(U, \mathfrak{B}(U), \mu, \mathbb{R}) \,:\, \mathcal{R}(X) < \infty\} \neq \varnothing.$$

(D2) is lower semicontinuous or closed, that is its epigraph

$$\operatorname{epi}(\mathcal{R}) := \{(X, \eta) \in L^p(U, \mathfrak{B}(U), \mu, \mathbb{R}) \times \mathbb{R} \,:\, \mathcal{R}(X) \leqslant \eta\}$$

is closed in the product topology on $L^p(U, \mathfrak{B}(U), \mu, \mathbb{R}) \times \mathbb{R}$,

(D3) is convex, that is $\mathcal{R}(\lambda X + (1-\lambda)\tilde{X}) \leqslant \lambda \mathcal{R}(X) + (1-\lambda)\mathcal{R}(\tilde{X})$, for all $\lambda \in [0,1]$ and all $X, \tilde{X} \in L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$.

(D4) satisfies $\mathcal{R}(C) = C$ for all constant random variables $C \in \mathbb{R}$.

As we will see in the following subsection, regular risk measures provide a minimal set of assumptions to ensure that many essential properties hold. Note, that (3.4) suggests that the expected value is not capable of adequately representing risk in many applications. To this end we introduce risk-averse measures.

3.2.1 Mean based risk measures

The idea of mean based risk models is to characterize the uncertain outcome by the mean $\mathbb{E}$, and a second scalar characteristic $\mathbb{D}$ describing the risk or dispersion. This approach allows to formulate the problem as a parametric optimization problem, and it facilitates the trade-off analysis between mean and risk. For $c \geqslant 0$, we consider

$$\mathcal{R}(X) = \mathbb{E}[X] + c\,\mathbb{D}(X)\,,$$

for a proper random variable X. In the case of the variance $\mathbb{D} := \mathbb{V}$ for example, we chose $X \in L^2(U, \mathfrak{B}(U), \mu, \mathbb{R})$. The variance treats the excess over the mean as the shortfall, i.e., positive and negative deviations from the mean equally. In minimization problems, we are not concerned if a particular realization of the random variable objective J is significantly below the mean of J, however we do not want it to be large. Two particular classes of risk functionals, which we discuss next, play an important role in the theory of mean-risk models.

Semideviations

An important group of risk functionals (representing dispersion measures) are central semideviations. The upper semideviation of order p is defined as

$$\sigma_p^+(X) := \left(\mathbb{E}\left[(X - \mathbb{E}[X])_+^p\right]\right)^{\frac{1}{p}}$$

where $p \in [1, \infty)$ is a fixed parameter, and $(x)_+ := \max\{0, x\}$. The upper semideviation is well defined and finite valued for all $X \in L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$.

The upper semideviation measure is appropriate for minimization problems as it penalizes excess of X over its mean. For maximization problems one would consider

$$\sigma_p^-(X) := -\left(\mathbb{E}\left[(\mathbb{E}[X]-X)_+^p\right]\right)^{\frac{1}{p}}.$$

In the special case of $p=1$, first order semideviations are related to the mean absolute deviation, in particular it holds $\sigma_1^+(X)=\sigma_1^-(X)=\frac{1}{2}\sigma_1(X)$ for all $X\in L^1(U,\mathfrak{B}(U),\mu,\mathbb{R})$, where

$$\sigma_1(X) := \mathbb{E}\left|X-\mathbb{E}[X]\right|.$$

Quantiles and Value-at-Risk

Let $F_X(x)=\mu(X\leqslant x)$ be the cumulative distribution function of the real valued random variable X and let $\alpha\in(0,1)$. We define the left-side α-quantile of F_X by

$$F_X^{-1}(\alpha) := \inf_{t\in\mathbb{R}}\{t\,:\,F_X(t)\geqslant\alpha\},$$

and the right-side α-quantile by

$$\sup_{t\in\mathbb{R}}\{t\,:\,F_X(t)\leqslant\alpha\}.$$

If X represents losses, as is the case in the minimization problems in this work, the left-side quantile $F_X^{-1}(\alpha)$ is also called Value-at-Risk with confidence level α and denoted by

$$\begin{aligned}\mathrm{V@R}_\alpha(X) := F_X^{-1}(\alpha) &= \inf_{t\in\mathbb{R}}\{t\,:\,\mu(X\leqslant t)\geqslant\alpha\}\\ &= \inf_{t\in\mathbb{R}}\{t\,:\,\mu(X>t)\leqslant 1-\alpha\}\end{aligned} \tag{3.5}$$

It has the following interpretation: losses larger than $V@R_\alpha(X)$ occur with probability not exceeding $1-\alpha$.

Weighted Mean Deviations from Quantiles

For $X\in L^1(U,\mathfrak{B}(U),\mu,\mathbb{R})$ we define the weighted mean deviation from a quantile as

$$\mathbb{W}_\alpha := \mathbb{E}\left[\max\{(1-\alpha)(F_X^{-1}(\alpha)-X),\alpha(X-F_X^{-1}(\alpha))\}\right].$$

Defining $\phi(t):=\mathbb{E}[\max\{(1-\alpha)(t-X),\alpha(X-t)\}]$ one gets the alternative representation

$$\mathbb{W}_\alpha = \min_{t\in\mathbb{R}}\phi(t).$$

Indeed, the right- and left-side derivatives[14] $\phi'_+(t)$ and $\phi'_-(t)$ of ϕ at a minimizer t are

$$\begin{aligned}\phi'_+(t) &= (1-\alpha)\mu(X\leqslant t)-\alpha\mu(X>t)\geqslant 0,\\ \phi'_-(t) &= (1-\alpha)\mu(X<t)-\alpha\mu(X\geqslant t)\leqslant 0,\end{aligned}$$

and thus $\mu(X<t)\leqslant\alpha\leqslant\mu(X\leqslant t)$, i.e., every α-quantile is a minimizer of ϕ. As we shall see in the next paragraph, the mean deviation from a quantile is related to the Average Value-at-Risk.

[14] Let $f: I\to\mathbb{R}$ be a real-valued function defined on a subset $I\subset\mathbb{R}$. Let a be a limit point of the set $\{x\in D: x>a\}$. Then $f'_+(a)$ is the right-side derivative of f at a if it is the right-sided limit $f'_+(a):=\lim_{x\searrow a}\frac{f(x)-f(a)}{x-a}$. A point x in a subset S of a topological space X is called limit point of S, if every neighborhood (a set including an open set that contains x) of x contains at least one point of S different from x itself. The right-sided limit is defined as the real number L, if for all $\epsilon>0$ there is a $\delta>0$ such that for all $x\in I$ with $0<x-a<\delta$ it holds that $|f(x)-L|<\epsilon$. The left-sided limit and derivative is defined analogously.

Average Value-at-Risk

Suppose one wants to satisfy the (chance) constraint

$$\mathrm{V@R}_{1-\alpha}(X) \leqslant 0\,. \tag{3.6}$$

Recalling (3.5), this is equivalent to the constraint $\mu(X \leqslant 0) \geqslant 1-\alpha$ and hence can be written as

$$\mathbb{E}\left[\mathbb{1}_{(0,\infty)}(X)\right] \leqslant \alpha\,,$$

where $\mathbb{1}_{(0,\infty)}(x) = 0$ if $x \leqslant 0$, and $\mathbb{1}_{(0,\infty)}(x) = 1$ if $x > 0$.
Such constraints are difficult to handle, since the step function $\mathbb{1}_{(0,\infty)}(X)$ is not convex, and discontinuous at zero. As a result, chance constraints are often nonconvex, even if the function $z \mapsto J(S(\boldsymbol{y})z) = X$ is convex almost surely. To avoid these difficulties such problems are often approached by constructing a convex approximation of $\mathbb{E}\left[\mathbb{1}_{(0,\infty)}(X)\right]$. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a nonnegative valued, nondecreasing, convex function such that $\phi(x) \geqslant \mathbb{1}_{(0,\infty)}(x)$ for all $x \in \mathbb{R}$. Noting that $\mathbb{1}_{(0,\infty)}(xt) = \mathbb{1}_{(0,\infty)}(x)$ for any $t > 0$ and $x \in \mathbb{R}$, we have that $\phi(tx) \geqslant \mathbb{1}_{(0,\infty)}(x)$, and hence the following inequality holds:

$$\inf_{t>0} \mathbb{E}[\phi(tX)] \geqslant \mathbb{E}[\mathbb{1}_{(0,\infty)}(X)]\,.$$

We obsere that the constraint

$$\inf_{t>0} \mathbb{E}[\phi(tX)] \leqslant \alpha \tag{3.7}$$

is a conservative approximation of (3.6) in the sense that the feasible set defined by (3.7) is contained in the feasible set defined by (3.6).
Aiming to make the upper bound as tight as possible, we take a piecewise linear function $\phi(x) = \max\{1+\gamma x, 0\}$ for some $\gamma > 0$. Since (3.7) is invariant with respect to change of $\phi(\gamma x)$ to $\phi(x)$, we have that $\phi(x) := \max\{1+x, 0\}$ gives the best choice of such a function. For this choice of ϕ, we have that (3.7) is equivalent to

$$\inf_{t>0}\{t\,\mathbb{E}[\max\{t^{-1}+X, 0\}] - \alpha\} \leqslant 0\,,$$

or equivalently after replacing t with $-t^{-1}$

$$\inf_{t<0}\{t + \alpha^{-1}\,\mathbb{E}[\max\{X-t, 0\}]\} \leqslant 0\,. \tag{3.8}$$

For $X \in L^1(U, \mathfrak{B}(U), \mu, \mathbb{R})$ we define the Average Value-at-Risk of X at level $1-\alpha$ by

$$\mathrm{AV@R}_{1-\alpha}(X) := \inf_{t\in\mathbb{R}}\{t + \alpha^{-1}\,\mathbb{E}[\max\{X-t, 0\}]\}\,.$$

Observe that $\psi(t) := t + \alpha^{-1}\,\mathbb{E}[\max\{X-t, 0\}]$ is a convex function, with derivative $\psi'(t) = 1 + \alpha^{-1}(F_X(t) - 1)$ if $F_X(t)$ is continuous at t. In case F_X is not continuous at t, the right- and left-sided derivatives are given by the same formula with $F_X(t)$ replaced by the right- and left-sided limits, respectively. Thus ψ attains a minimum on the interval $[\inf\{t : F_X(t) \geqslant 1-\alpha\}, \sup\{t : F_X(t) \leqslant 1-\alpha\}]$. For any minimizer t^* of ψ we have

$$\mathrm{AV@R}_{1-\alpha}(X) = t^* + \alpha^{-1}\mathbb{E}[\max\{X-t^*, 0\}]\,, \tag{3.9}$$

where the second summand is clearly nonnegative. Thus $\mathrm{AV@R}_{1-\alpha}(X) \leqslant 0$ implies $t^* \leqslant 0$, and hence (3.8) is equivalent to $\mathrm{AV@R}_{1-\alpha}(X) \leqslant 0$, and provides a conservative approximation of (3.6). Note that (3.9) holds for $t^* = \mathrm{V@R}_{1-\alpha}(X)$.

Lemma 3.2.1. *Let $F(X)$ be convex and monotone, i.e., $F(x) \geqslant F(y)$ for $x \geqslant y$, and let $G(x)$ be convex. Then $F(G(X))$ is convex.*

Proof. By convexity of G and monotonicity of F we have

$$F(G(\lambda x + (1-\lambda)y)) \leqslant F(\lambda G(x) + (1-\lambda)G(x))\,,$$

and thus by convexity of F

$$F(G(\lambda x + (1-\lambda)y)) \leqslant \lambda F(G(x)) + (1-\lambda)F(G(x))\,.$$

□

Since $\mathrm{AV@R}_{1-\alpha}(X)$ is convex and monotone, the mapping $z \mapsto \mathrm{AV@R}_{1-\alpha}(J(S(\boldsymbol{y})z))$ is convex provided $z \mapsto J(S(\boldsymbol{y})z)$ is convex, and hence $\mathrm{AV@R}_{1-\alpha}(X) \leqslant 0$ is a convex, conservative approximation of the chance constraint (3.6).
Moreover, there is a relation between the Average Value-at-Risk and the weighted mean deviations from quantiles.

Theorem 3.2.2. *Let $X \in L^1(U, \mathfrak{B}(U), \mu, \mathbb{R})$ with cumulative distribution function F_X. Then it holds*

$$\mathrm{AV@R}_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{V@R}_\tau(X)\,\mathrm{d}\tau = \mathbb{E}[X] + \frac{1}{1-\alpha}\mathbb{W}_\alpha(X)\,.$$

Moreover, if $F_X(x)$ is continuous at $x = \mathrm{V@R}_\alpha(X)$, then

$$\mathrm{AV@R}_\alpha(X) = \frac{1}{1-\alpha}\int_{\mathrm{V@R}_\alpha(X)}^\infty x\,\mathrm{d}F_X(x) = \mathbb{E}[X | X \geqslant \mathrm{V@R}_\alpha(X)]\,.$$

Proof. See [137, Theorem 6.2]. □

Thus we have For $\alpha = 0$ we get

$$\mathrm{AV@R}_\alpha(X) = \int_0^1 \mathrm{V@R}_\tau(X)\,\mathrm{d}\tau = \int_0^1 F_X^{-1}(\tau)\,\mathrm{d}\tau = \int_{\mathbb{R}} x\,\mathrm{d}F_X(x) = \mathbb{E}[X]\,,$$

and

$$\lim_{\alpha \nearrow 1} \frac{1}{1-\alpha}\int_{1-\alpha}^1 F_X^{-1}(\tau)\,\mathrm{d}\tau = \operatorname{ess\,sup}(X)\,.$$

Moreover, $\mathrm{AV@R}_\alpha(X)$ is continuous and monotonically increasing in $\alpha \in [0,1)$. Thus we have

$$\mathbb{E}[X] \leqslant \mathrm{AV@R}_\alpha(X) \leqslant \operatorname{ess\,sup}(X)\,.$$

and

$$\mathrm{V@R}_\alpha(X) \leqslant \mathrm{AV@R}_\alpha(X)$$

for $\alpha \in [0,1)$ and $X \in L^1(U, \mathfrak{B}(U), \mu, \mathbb{R})$.
Figure 3.1 illustrates the relationship between the $\mathrm{V@R}_\alpha(X)$ and the $\mathrm{V@R}_\alpha(X)$ for $X \sim \mathcal{U}([0,1])$.
One can show, see, e.g., [133] that the Average Value-at-Risk has certain properties that characterize the class of coherent risk measures.

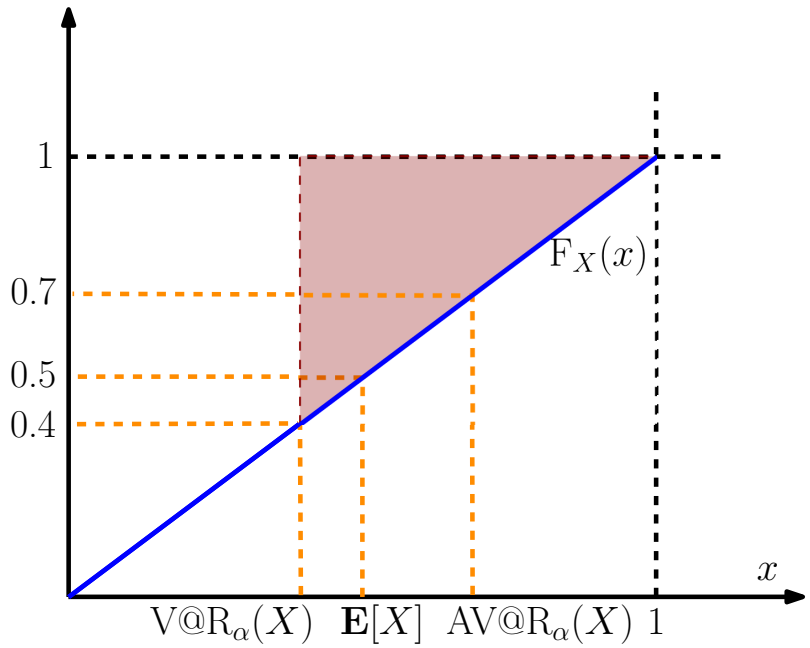


Figure 3.1: Let $X \sim \mathcal{U}([0,1])$ be uniformly distributed in the unit interval and $\alpha = 0.4$. Then we have $\mathrm{V@R}_\alpha(X) = 0.4$, $\mathbb{E}[X] = 0.5$, and $\mathrm{AV@R}_\alpha(X) = 0.7$. In general the $\mathrm{AV@R}_\alpha(X)$ is the average of the (here red) area between the cumulative distribution function $y = F_X(x)$ (here blue) and $y = 1$, to the right of the $\mathrm{V@R}_\alpha(X)$.

3.2.2 Coherent risk measures

A popular class of risk measures is the class of coherent risk measures, which is characterized by the following conditions (see [7]). A functional $\mathcal{R} : L^p(U, \mathfrak{B}(U), \mu, \mathbb{R}) \to \mathbb{R} \cup \{\infty\}$, for $p \in [1, \infty)$, is said to be a coherent measure of risk if the following conditions hold: For $X, \tilde{X} \in L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$ it holds

(R1) (Convexity) $\mathcal{R}(\lambda X + (1-\lambda)\tilde{X}) \leqslant \lambda\mathcal{R}(X) + (1-\lambda)\mathcal{R}(\tilde{X})$, for all $\lambda \in [0,1]$.

(R2) (Translation equivariance) $\mathcal{R}(X + c) = \mathcal{R}(X) + c$, for all $c \in \mathbb{R}$.

(R3) (Monotonicity) If $X \leqslant \tilde{X}$ μ-a.e., then $\mathcal{R}(X) \leqslant \mathcal{R}(\tilde{X})$.

(R4) (Positive homogeneity) $\mathcal{R}(tX) = t\mathcal{R}(X)$ for all $t \geqslant 0$.

If a risk measure $\mathcal{R}$ satisfies only conditions (R1) - (R3) it is called convex risk measure. Note that the condition (R1) implies $\mathcal{R}(\frac{X_1+X_2}{2}) \leqslant \frac{\mathcal{R}(X_1)+\mathcal{R}(X_2)}{2}$, and thus with (R4) it follows that

$$\mathcal{R}(X + \tilde{X}) \leqslant \mathcal{R}(X) + \mathcal{R}(\tilde{X})\,, \tag{3.10}$$

for $X, \tilde{X} \in L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$. The property (3.10) is called subadditivity. Subadditivity and positive homogeneity (R4) imply in turn the convexity (R1) of a risk measure $\mathbb{R}$. Thus the convexity condition (R1) can be replaced by 3.10 in the characterization of coherent risk measures.

Moreover, subadditivity has the interpretation that the risk of the sum of two quantities of interest X and $\tilde{X}$ is less or equal to the sum of the risk of X_1, X_2. By adding uncertainty $\tilde{X}$ to X the total risk increases at most by the risk of $\tilde{X}$, i.e., the effect of diversification on risk is considered. The translation equivariance (R2) means that the addition of a risk-free quantity c to the loss X, changes the risk of X exactly by c. For condition ((R3)) it is assumed that $\tilde{X}$ is larger than X for any possible outcome, that is $\tilde{X}$ is always a higher loss than X. The monotonicity of a risk measure ensures that in this case $\tilde{X}$ has a higher risk than X. Positive homogeneity ensures that a loss $\tilde{X} = tX$ for $t \geqslant 0$ taking t times the value of X has t times the risk of X.

Remark 3.2.3. *In the case when the random objective function represents a reward that is to be maximized instead of a loss that is to be minimized, large realizations are preferred, and we can define a risk measure $\tilde{\mathcal{R}}(X) = \mathcal{R}(-X)$, where $\mathcal{R}$ satisfies (R1) - (R4). In this case $\tilde{\mathcal{R}}$ satisfies (R1) and (R4) and*

(R1') (Translation equivariance) $\mathcal{R}(X + c) = \mathcal{R}(X) - c$, for all $c \in \mathbb{R}$.

(R2') (Monotonicity) If $X \leqslant \tilde{X}$ μ-a.e., then $\mathcal{R}(X) \leqslant \mathcal{R}(\tilde{X})$.

Hence all statements regarding risk measure satisfying (R1) - (R4) have their trivial counterparts for risk measures satisfying (R1), (R1'), (R2'), (R4).

The axioms characterizing coherent risk measures guarantee a number of desirable properties to told:

Lemma 3.2.4 ([137, Proposition 6.5])**.** *Let $\mathcal{R} : L^p(U, \mathfrak{B}(U), \mu, \mathbb{R}) \to \mathbb{R}$ with $p \in [1, \infty]$, satisfy (R1) and (R3). Then $\mathcal{R}$ is continuous and subdifferentiable[15] on $L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$.*

Lemma 3.2.5 ([137, Proposition 6.7])**.** *Let $\mathcal{R} : L^p(U, \mathfrak{B}(U), \mu, \mathbb{R}) \to \mathbb{R} \cup \{\pm\infty\}$ with $p \in [1, \infty)$ be a proper risk measure satisfying (R1), (R2) and (R3), with $\mathrm{dom}(\mathcal{R})$ having nonempty interior.[16] Then $\mathcal{R}$ is finite valued and continuous on $L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$.*

Besides the above continuity properties, the Fenchel–Moreau theorem, see, e.g., [137, Theorem 7.82] allows the identification $\mathcal{R} = \mathcal{R}^{**}$ of $\mathcal{R}$ with its biconjugate $\mathcal{R}^{**}$, defined by

$$\mathcal{R}^{**}(X) = \sup_{\xi \in L^q(U, \mathfrak{B}(U), \mu, \mathbb{R})} \{\langle \xi, X \rangle_{L^q(U, \mathfrak{B}(U), \mu, \mathbb{R}), L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})} - \mathcal{R}^*(\xi)\}$$

where the so-called conjugate $\mathcal{R}^*$ of $\mathcal{R}$ is defined as

$$\mathcal{R}^*(\xi) := \sup_{X \in L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})} \{\langle \xi, X \rangle_{L^q(U, \mathfrak{B}(U), \mu, \mathbb{R}), L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})} - \mathcal{R}(X)\}$$

and $\frac{1}{p} + \frac{1}{q} = 1$. Particularly we have for $\mathcal{R}$ proper, convex, and lower semicontinuous, that it has the representation

$$\mathcal{R}(X) = \sup_{\xi \in \mathrm{dom}(\mathcal{R}^*)} \{\langle \xi, X \rangle_{L^q(U, \mathfrak{B}(U), \mu, \mathbb{R}), L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})} - \mathcal{R}^*(\xi)\}.$$

[15] A vector x^* is said to be the subgradient of a convex function f at a point x if $f(y) \geqslant f(x) + \langle x^*, y - x \rangle$ for all $y \in \mathbb{R}^n$. The set of all subgradients of f at x is called subdifferential of f at x and denoted by $\partial f(x)$. If $\partial f(x)$ is nonempty, f is called subdifferentiable at x.

[16] see also (D1).

Moreover, we have that the condition (R3) is equivalent to $\xi \geqslant 0$ μ-a.e. for all $\xi \in L^q(U, \mathfrak{B}(U), \mu, \mathbb{R})$, and that (R2) is equivalent to $\mathbb{E}[\xi] = 1$ for all $\xi \in \mathrm{dom}(\mathcal{R}^*)$, and that (R4) holds if and only if $\mathcal{R}$ can be represented for all $X \in L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$ by $\mathcal{R}(X) = \sup_{\xi \in \mathrm{dom}(\mathcal{R}^*)} \langle \xi, X \rangle_{L^q(U,\mathfrak{B}(U),\mu,\mathbb{R}), L^p(U,\mathfrak{B}(U),\mu,\mathbb{R})}$, see [137, Theorem 6.5]. In the case $p = \infty$ the space $L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$ is equipped with the weak*-topology and paired with $L^1(U, \mathfrak{B}(U), \mu, \mathbb{R})$. For an analogous dual representation one needs the additional assumption that $\mathcal{R}$ is lower semicontinuous in the weak*-topology. For this special case we refer to [137, Chapter 6.3].

3.2.3 Entropic risk measure and entropic Value-at-Risk

For $\theta > 0$ the entropic risk measure is defined by

$$\mathcal{R}(X) = \frac{1}{\theta} \ln \left(\mathbb{E}\left[\exp\left(\theta X \right) \right] \right).$$

Observe that $\mathcal{R}(X) < \infty$ can be ensured for $X \in L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$ with $p = \infty$. It is shown next, that $\mathcal{R}$ satisfies (R1) - (R3), and thus by Lemma 3.2.4 is continuous and subdifferentiable on $L^\infty(U, \mathfrak{B}(U), \mu, \mathbb{R})$. Using Hölder's inequality one gets

$$\mathbb{E}[XY] \leqslant \left(\mathbb{E}[|X|^r] \right)^{\frac{1}{r}} \left(\mathbb{E}[|Y|^s] \right)^{\frac{1}{s}},$$

for any $1 \leqslant r, s \leqslant \infty$ satisfying $\frac{1}{r} + \frac{1}{s} = 1$. Setting $X = \exp\left((1-\lambda)\theta X \right)$ and $Y = \exp\left(\lambda \theta Y \right)$, $r = \frac{1}{1-\lambda}$, $s = \frac{1}{\lambda}$ for any $\lambda \in [0, 1]$, gives

$$\mathbb{E}[\exp\left((1-\lambda)\theta X + \lambda \theta Y \right)] \leqslant \left(\mathbb{E}\left[\exp\left(\theta X \right) \right] \right)^{1-\lambda} \left(\mathbb{E}\left[\exp\left(\theta Y \right) \right] \right)^{\lambda}.$$

Taking the ln and dividing by θ on both sides gives

$$\frac{1}{\theta} \ln(\mathbb{E}[\exp((1-\lambda)\theta X + \lambda \theta Y)]) \leqslant (1-\lambda) \frac{1}{\theta} \ln(\mathbb{E}\left[\exp(\theta X) \right]) + \lambda \frac{1}{\theta} \ln(\mathbb{E}\left[\exp(\theta Y) \right]),$$

and thus the convexity (R1) of $\mathcal{R}$.
The entropic risk measures are translation equivariant (R2), since for $c \in \mathbb{R}$ it holds that

$$\mathcal{R}(X + c) = \frac{1}{\theta} \ln \left(\mathbb{E}\left[\exp\left(\theta (X + c) \right) \right] \right) = \frac{1}{\theta} \ln \left(\mathbb{E}\left[\exp\left(\theta X \right) \right] \exp\left(\theta c \right) \right) = \mathcal{R}(X) + c.$$

From the monotonicity of ln and exp it follows that the entropic risk measures $\mathcal{R}$ are monotonic, i.e., satisfy (R3). Moreover, we have

$$\mathbb{E}[X] \leqslant \frac{1}{\theta_1} \ln \left(\mathbb{E}\left[\exp\left(\theta_1 X \right) \right] \right) \leqslant \frac{1}{\theta_2} \ln \left(\mathbb{E}\left[\exp\left(\theta_2 X \right) \right] \right) \leqslant \operatorname{ess\,sup}(X), \tag{3.11}$$

for $0 \leqslant \theta_1 \leqslant \theta_2$. For nonconstant $X \in L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$ and $0 < \theta_1 < \theta_2$, (3.11) holds with strict inequality, see, e.g., [50].
The entropic risk measures are not positively homogeneous, i.e., do not satisfy (R4). This deficiency is overcome by the so-called entropic Value-at-Risk, at level α defined by

$$\mathrm{EV@R}_\alpha(X) = \inf_{\theta > 0} \frac{1}{\theta} \ln \left(\frac{1}{1-\alpha} \mathbb{E}\left[\exp\left(\theta X \right) \right] \right),$$

for $\alpha \in [0, 1)$ and $\mathrm{EV@R}_\alpha(X) = \operatorname{ess\,sup}(X)$ for $\alpha = 1$.

The $\mathrm{EV@R}_\alpha$ satisfies (R1) - (R4) and is thus a coherent risk measure. Moreover it is an upper bound on the $\mathrm{AV@R}_\alpha$, and it holds

$$\mathrm{V@R}_\alpha(X) \leqslant \mathrm{AV@R}_\alpha(X) \leqslant \mathrm{EV@R}_\alpha \leqslant \operatorname{ess\,sup}(X)$$

and

$$\mathbb{E}[X] \leqslant \mathrm{AV@R}_\alpha(X) \leqslant \mathrm{EV@R}_\alpha \leqslant \operatorname{ess\,sup}(X)\,.$$

Figure 3.2 illustrates the behaviour of $\mathrm{V@R}_\alpha(X)$, $\mathrm{AV@R}_\alpha(X)$, and $\mathrm{EV@R}_\alpha(X)$ for different values of α and $X \sim \mathcal{U}([0,1])$ being uniformly distributed in the unit interval, cf., [3].

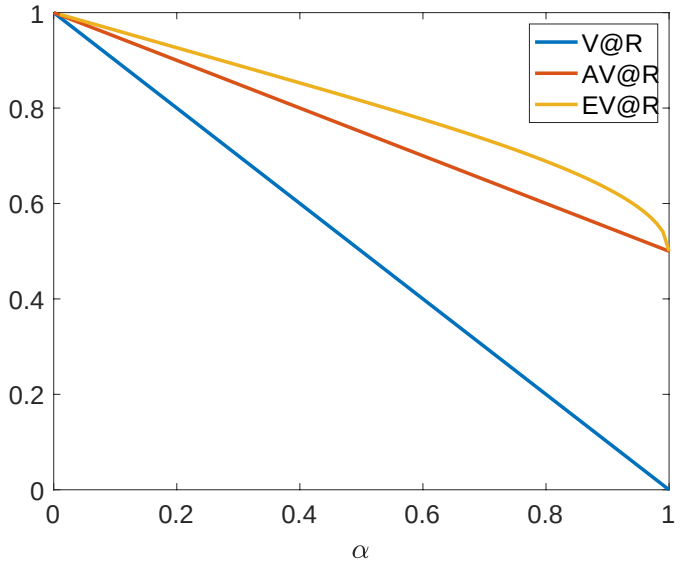


Figure 3.2: Comparison of $\mathrm{V@R}_\alpha(X)$, $\mathrm{AV@R}_\alpha(X)$, and $\mathrm{EV@R}_\alpha(X)$ for different values of α and $X \sim \mathcal{U}([0,1])$ uniformly distributed in the unit interval.

We will later see in Section 4.6.3 that the entropic risk measures have certain regularity properties, and hence they are well suited for the application of higher-order cubature rules.

	(R1)	(R2)	(R3)	(R4)
$\mathbb{E}[X]$	Yes	Yes	Yes	Yes
$\mathbb{E}[X] + c\mathbb{E}[(X - \mathbb{E}[X])_+^p]^{\frac{1}{p}}$	Yes	Yes	If $c \in [0,1]$	Yes
$\mathrm{V@R}_\alpha(X)$	No	Yes	Yes	Yes
$\mathrm{AV@R}_\alpha(X)$	Yes	Yes	Yes	Yes
$\mathrm{EV@R}_\alpha(X)$	Yes	Yes	Yes	Yes
$\frac{1}{\theta}\ln\left(\mathbb{E}[\exp(\theta X)]\right)$	Yes	Yes	Yes	No
$\operatorname{ess\,sup}(X)$	Yes	Yes	Yes	Yes

Table 3.1: For $\alpha \in (0,1), \theta > 0$, and $c \geqslant 0$.

Remark 3.2.6. *It is easy to see that all risk measures in Table 3.1, except the* $\mathrm{V@R}_\alpha$ *and the expected value,* $\mathbb{E}$*, are risk-averse. Hence* $\mathbb{E}[X] > -\infty$ *for all* $X \in L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$*, and* $\{X \in L^p(U, \mathfrak{B}(U), \mu, \mathbb{R}) : \operatorname{ess\,sup}(X) < \infty\} \neq \varnothing$ *are sufficient conditions ensuring that all risk measures in Table 3.1 are proper, i.e., satisfy (D1). Furthermore, all risk measures in Table 3.1 satisfy (D4). For risk measures satisfying (R1) and (R3) continuity (and hence lower semicontinuity (D2)) follows from Lemma 3.2.4. We conclude that the risk measures in Table 3.1 are regular ones, except for the expected value (not risk-averse) and the* $\mathrm{V@R}_\alpha$ *(not risk-averse, not convex).*

A major focus of this work is the design and application of efficient numerical methods for the approximation of high-dimensional integrals introduced by the risk measures to the optimal control problems. These methods heavily rely on the parametric regularity of the integrands. For this reason, we focus on smooth risk measures, that inherit the parametric regularity, such as the expected value and the entropic risk measures. The expected value is in addition a coherent measure of risk and the coherent entropic Value-at-Risk can easily be recovered from the entropic risk measures in an optimization problem by the additional minimization with respect to the risk-aversion parameter θ, see Section 4.2.5. Approximations of other risk-measures typically rely on Monte Carlo methods, as for instance in [157], where the authors consider a combination of the expected value and the variance in conjunction with a multilevel Monte-Carlo method.
We note that recent work on smoothing by preintegration [61] might enable the application of efficient methods, that exploit the parametric regularity of the integrands, to nonsmooth risk measures. While smoothing of risk measures, such as the $\mathrm{AV@R}_\alpha$ in [105], has successfully been applied to PDE-constrained optimal control problems under uncertainty, smoothing by preintegration of risk measures in optimal control problems under uncertainty remains for future research.

3.3 Random variable objective function

The parameterized cost functional $J : U \times \mathcal{X} \to \mathbb{R}$ associates a nonnegative cost to each pair of $\boldsymbol{y} \in U$ and $u(\boldsymbol{y}) \in \mathcal{X}$. Recall from (3.1), that for $u : U \to \mathcal{X}$ we refer to the Nemytskii (superposition) operator

$$\mathcal{J}(u)(\boldsymbol{y}) = J(u(\boldsymbol{y}), \boldsymbol{y}), \quad \boldsymbol{y} \in U$$

as the random variable objective function. The Nemytskii operator assigns a function $\mathcal{J}(u) : U \to \mathbb{R}$ to the state $u(\cdot) : U \to \mathcal{X}$. Recalling that

$$u \in L^q(U, \mathfrak{B}, \mu, \mathcal{X}) \quad \text{and} \quad \mathcal{J} \in L^p(U, \mathfrak{B}, \mu, \mathbb{R}), \tag{3.12}$$

we conclude that the Nemytskii operator maps from $L^q(U, \mathfrak{B}, \mu, \mathcal{X})$ into $L^p(U, \mathfrak{B}, \mu, \mathbb{R})$. In this section, we show that from the following assumptions on the objective function various desirable properties of the objective function follow.

Assumption 3.3.1.

(i) The mapping $J : U \times \mathcal{X} \to \mathbb{R}$ *is Carathéodory, that is* $J(\cdot, \boldsymbol{y})$ *is continuous for* μ*-a.e.* $\boldsymbol{y} \in U$ *and* $J(u, \cdot)$ *is measurable for all* $u \in \mathcal{X}$*.*

(ii) *For* $1 \leqslant p, q < \infty$ *in* (3.12)*, there is* $v \in L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$ *with* $v \geqslant 0$ μ*-a.e. and a constant* $C > 0$ *such that*

$$|J(u, \boldsymbol{y})| \leqslant v(\boldsymbol{y}) + C\|u\|_{\mathcal{X}}^{\frac{q}{p}}. \tag{3.13}$$

For $q = \infty$ *and* $1 \leqslant p < \infty$ *in* (3.12)*, it holds that for all* $C > 0$ *there exists* $b = b(C) \in L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$ *such that*

$$|J(u, \boldsymbol{y})| \leqslant b(\boldsymbol{y}) \quad \mu\text{-a.e.} \quad \forall u \in \mathcal{X}\,, \|u\|_{\mathcal{X}} \leqslant C\,.$$

For $q = p = \infty$ *in* (3.12)*, it holds that for all* $C > 0$ *there is* $d(C)$ *such that*

$$|J(u, \boldsymbol{y})| \leqslant d \quad \text{for almost all } \boldsymbol{y} \in U \text{ and for all } u \in \mathcal{X},\ \|u\|_{\mathcal{X}} \leqslant C. \tag{3.14}$$

(iii) *For* μ*-a.e.* $\boldsymbol{y} \in U$*,* $J(\cdot, \boldsymbol{y})$ *is convex.*

Common objective functions are so-called tracking-type objective functionals.

Example 3.3.2. *Consider a tracking-type functional based on* $L^q(U, \mathfrak{B}(U), \mu, \mathcal{X})$ *for* $q = 2$ *and let* $\mathcal{X} \hookrightarrow \mathcal{Y}$ *in a Hilbert space* $\mathcal{Y}$ *with* $\widehat{u} \in \mathcal{Y}$*. Then there is a constant* $C > 0$ *such that for all* $u \in \mathcal{X}$ *it holds that*

$$0 \leqslant J(\boldsymbol{y}, u) := \frac{1}{2}\|u(\boldsymbol{y}) - \widehat{u}\|_{\mathcal{Y}}^2 \leqslant \|u(\boldsymbol{y})\|_{\mathcal{Y}}^2 + \|\widehat{u}\|_{\mathcal{Y}}^2 \leqslant \|u(\boldsymbol{y})\|_{\mathcal{X}}^2 + \|\widehat{u}\|_{\mathcal{Y}}^2\,. \tag{3.15}$$

By setting $v(\boldsymbol{y}) = \|\widehat{u}\|_{\mathcal{Y}}^2 \in L^\infty(U, \mathfrak{B}(U), \mu, \mathbb{R}) \subset L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$ *for* $p = 1$*, the condition* (3.13) *is satisfied. Moreover, for given* $\boldsymbol{y} \in U$ *the mapping* $J(\cdot, \boldsymbol{y}) : \mathcal{X} \to \mathbb{R}$ *is continuous as the composition of continuous mappings. If* $u(\boldsymbol{y})$ *is Bochner integrable,* $\|u(\boldsymbol{y})\|_{\mathcal{X}}$ *is Lebesgue-integrable, and hence it is in particular measurable. Thus* $J(u, \cdot)$ *is measurable for all* $u \in \mathcal{X}$*. For given* $\boldsymbol{y} \in U$*, it is easy to verify the convexity of* $J(\cdot, \boldsymbol{y})$*. For given* $\boldsymbol{y} \in U$*, the norm* $\|\cdot - \widehat{u}\|_{\mathcal{X}} : \mathcal{X} \to \mathbb{R}$ *is convex, because of the absolute homogeneity and the triangle inequality of the norm, see Section 2.1. Clearly the function* $x \mapsto x^2$ *is convex and monotonically increasing for* $x \geqslant 0$*, and hence by Lemma 3.2.1* $J(\cdot, \boldsymbol{y})$ *is convex for given* $\boldsymbol{y} \in U$*, which in turn implies Assumption 3.3.1 (iii).*

Theorem 3.3.3 ([106, Theorem 3.5])**.** *Let Assumption 3.3.1 (i) - (ii) hold. Then the Nemytskii operator* $\mathcal{J} : L^q(U, \mathfrak{B}(U), \mu, \mathcal{X}) \to L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$ *is continuous.*

Corollary 3.3.4. *Let* $\mathcal{R} : L^p(U, \mathfrak{B}(U), \mu, \mathbb{R}) \to \mathbb{R}$ *with* $p \in [1, \infty]$ *satisfy (R1) and (R3), and let Assumption 3.3.1 (i) - (ii). Then the composite functional* $\mathcal{R} \circ \mathcal{J}$ *is continuous. If in addition Assumption 3.3.1 (iii) holds, then* $\mathcal{R} \circ \mathcal{J}$ *is convex and weakly lower semicontinuous.*

Proof. Continuity of $\mathcal{R} \circ \mathcal{J}$ follows from the preceeding theorem and Lemma 3.2.4. Convexity of $\mathcal{R} \circ \mathcal{J}$ follows from Lemma 3.2.1. Thus the composite functional is convex and continuous on a Banach space and hence it is weakly lower semicontinuous (see, e.g., [155, Theorem 2.12]). □

Theorem 3.3.5. *Let the conditions of the preceeding corollary hold. In addition, let* $\mathcal{R} \circ \mathcal{J} : L^q(U, \mathfrak{B}(U), \mu, \mathcal{X}) \to L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$ *be coercive, as well as* $1 < q < \infty$ *and* $\mathcal{X}$ *be a reflexive Banach space. Then there exists* $u^*(\boldsymbol{y}) \in L^q(U, \mathfrak{B}(U), \mu, \mathcal{X})$ *with*

$$\mathcal{R}(\mathcal{J}(u^*)(\boldsymbol{y})) = \inf_{u \in L^q(U, \mathfrak{B}(U), \mu, \mathcal{X})} \mathcal{R}(\mathcal{J}(u)(\boldsymbol{y}))$$

Proof. The composite objective functional $\mathcal{R} \circ \mathcal{J}$ is weakly lower semicontinuous and coercive on a reflexive Banach space. It is well-known (see, e.g., [9, Theorem 1.5.6]) that there exists a minimizer $u^*(\boldsymbol{y}) \in L^q(U, \mathfrak{B}(U), \mu, \mathcal{X})$ of $\mathcal{R} \circ \mathcal{J}$. □

Example 3.3.6. *Let $\mathcal{R} = \mathbb{E}$ and let J be a tracking-type functional, see* (3.15)*. Then there exists $u^* \in L^q(U, \mathfrak{B}(U), \mu, \mathcal{X})$, with $q = 2$, such that*

$$\mathbb{E}\left[\frac{1}{2}\|u^*(\boldsymbol{y}) - \widehat{u}\|_{\mathcal{Y}}^2\right] = \inf_{u \in L^q(U, \mathfrak{B}(U), \mu, \mathcal{X})} \mathbb{E}\left[\frac{1}{2}\|u(\boldsymbol{y}) - \widehat{u}\|_{\mathcal{Y}}^2\right].$$

We know that $\mathbb{E}$ is coherent and the tracking-type functional satisfies Assumption 3.3.1 (i) - (ii). Thus $\mathcal{R} \circ \mathcal{J}$ is weakly lower semicontinuous. Moreover, the space $\mathcal{X}$ is reflexive since it is a Hilbert space. Thus $L^q(U, \mathfrak{B}(U), \mu, \mathcal{X})$ is reflexive for $1 < q < \infty$, i.e., in particular for $q = 2$. Finally, it is easy to see, that $\mathcal{R}(\mathcal{J})$ is coercive, i.e., $\|u\|_{L^q(U, \mathfrak{B}(U), \mu, \mathcal{X})} \to \infty$ implies $|\mathcal{R}(\mathcal{J}(u))| \to \infty$.

Theorem 3.3.7 ([106, Theorem 3.9])**.** *Let Assumption 3.3.1 (i) - (iii) hold. Then $\mathcal{J}$ is Gâteaux directionally differentiable. If $\mathcal{J}$ is in addition locally Lipschitz[17] continuous, then $\mathcal{J}$ is Hadamard directionally differentiable.*

Theorem 3.3.8 ([106, Theorem 3.11])**.** *Suppose $J(\boldsymbol{y}, \cdot)$ is continuously Fréchet differentiable with respect to $u \in \mathcal{X}$ for μ-a.e. $\boldsymbol{y} \in U$, and there exists an $\alpha > 0$ and $K \in L^s(U, \mathfrak{B}(U), \mu, \mathbb{R})$ with*

$$s = \begin{cases} pq/(q - (1+\alpha)p) & \text{if } q > (1+\alpha)p\,, \\ \infty & \text{if } q = (1+\alpha)p \end{cases}$$

such that

$$|\partial_u J(\boldsymbol{y}, u) - \partial_u J(\boldsymbol{y}, v)| \leqslant K(\boldsymbol{y})\|u - v\|_U^\alpha$$

for μ-a.e. $\boldsymbol{y} \in U$, where $\partial_u J$ denotes the partial derivative with respect to u.
Then $\mathcal{J}$ is Fréchet differentiable from $L^q(U, \mathfrak{B}(U), \mu, \mathcal{X})$ into $L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$. Moreover, the derivative of $\mathcal{J}$ is

$$\mathcal{J}'(u) = \partial_u J(\cdot, u(\cdot)) \quad \text{for } u \in L^q(U, \mathfrak{B}(U), \mu, \mathcal{X})\,.$$

In this section we discuss which of the discussed risk measures are Fréchet differentiable. In this regard, an important result is the following.

Theorem 3.3.9 ([107, Theorem 7])**.** *Let $\mathcal{X}$ be a real Banach space and $F : \mathcal{X} \to \mathbb{R}$ be proper, convex, closed and positively homogeneous satisfying $F(0) < \infty$. Then the following statements are equivalent:*

(i) F is Gâteaux differentiable at zero.

(ii) F is Gâteaux differentiable everywhere on $\mathcal{X}$.

[17] Let (X, d_x) and (Y, d_y) be two metric spaces. A function $f : X \to Y$ is called Lipschitz (continuous) if there exists a real constant $L \geqslant 0$ such that $d_Y(f(x_1), f(x_2)) \leqslant L\, d_X(x_1, x_2)$ for all $x_1, x_2 \in X$. A function f is locally Lipschitz (continuous) if for every $x \in X$ there exists a neighborhood U of x such that f restricted to U is Lipschitz.

(iii) F is a bounded linear functional on $\mathcal{X}$.

From this theorem we directly obtain the following result.

Theorem 3.3.10 ([107, Theorem 1])**.** *Let $(U, \mathfrak{B}(U), \mu)$ be a probability space and let $L^p(U, \mathfrak{B}(U), \mu, \mathbb{R})$ with $p \in [1, \infty)$. If $\mathcal{R} : L^p(U, \mathfrak{B}(U), \mu, \mathbb{R}) \to \mathbb{R}$ is a coherent risk measure, then $\mathcal{R}$ is Fréchet differentiable if and only if there exists $\vartheta \in L^p(U, \mathfrak{B}, \mu, \mathbb{R})'$ with $\vartheta \geqslant 0$ μ-a.e., $\mathbb{E}[\vartheta] = 1$, and $\mathcal{R}(X) = \mathbb{E}[\vartheta X]$ for all $X \in L^p(U, \mathfrak{B}, \mu, \mathbb{R})$.*

Thus the only Fréchet differentiable coherent risk measures are linear functionals. This includes the expected value.

Theorem 3.3.11 ([56, Lemma C3])**.** *Let $\mathcal{Y}$ be an open subset of a Banach space $\mathcal{X}$ and let $J : U \times \mathcal{X} \to \mathbb{R}$ be a parameterized random variable functional with expected value $j : \mathcal{X} \to \mathbb{R}$, given by*

$$j(u) = \int_U J(\boldsymbol{y}, u) \, \mathrm{d}\mu \, .$$

Suppose that $J(\boldsymbol{y}, u) \in L^1(U, \mathfrak{B}(U), \mu, \mathcal{Y})$, and for a.e. $\boldsymbol{y} \in U$ let $J(\boldsymbol{y}, \cdot)$ be Fréchet differentiable at u, with derivative $\partial_u J(\boldsymbol{y}, u)$. Moreover, let $C(\cdot) \in L^1(U, \mathfrak{B}(U), \mu, \mathbb{R})$ such that for all $v \in \mathcal{Y}$ and almost every $\boldsymbol{y} \in U$ it holds that $\|\partial_u J(\boldsymbol{y}, u)\|_{\mathcal{X}'} \leqslant C(\boldsymbol{y})$. Then j is Fréchet differentiable at u and

$$j'(u) = \mathbb{E}[\partial_u J(\boldsymbol{y}, u)] \, .$$

Example 3.3.12. *The tracking-type objective functional considered in* (3.15) *is Fréchet differentiable at u for every $\boldsymbol{y} \in U$, with Fréchet derivative*

$$\partial_u J(\boldsymbol{y}, u)h := \langle u(\boldsymbol{y}) - \widehat{u}, h \rangle_{\mathcal{Y}} \, .$$

This is true since

$$\frac{1}{2}\|u(\boldsymbol{y}) + h - \widehat{u}\|_{\mathcal{Y}}^2 = \frac{1}{2}\|u(\boldsymbol{y}) - \widehat{u}\|_{\mathcal{Y}}^2 + \langle u(\boldsymbol{y}) - \widehat{u}, h \rangle_{\mathcal{Y}} + \frac{1}{2}\|h\|_{\mathcal{Y}}^2$$

and thus

$$J(\boldsymbol{y}, u + h) - J(\boldsymbol{y}, u) - J'(\boldsymbol{y}, u)h = \frac{1}{2}\|h\|_{\mathcal{Y}}^2 =: r(h, u)$$

satisfies

$$\frac{r(h, u)}{\|h\|_{\mathcal{Y}}} = \frac{\frac{1}{2}\|h\|_{\mathcal{Y}}^2}{\|h\|_{\mathcal{Y}}} = \frac{1}{2}\|h\|_{\mathcal{Y}} \to 0 \quad \text{as } \|h\|_{\mathcal{Y}} \to 0 \, .$$

Theorem 3.3.13. *Let $J(\boldsymbol{y}, u)$ be the tracking-type objective functional considered in* (3.15) *and let $\mathcal{R}(J(\cdot, u)) = \frac{1}{\theta} \ln(\mathbb{E}[\exp(\theta J(\cdot, u)]$ be the entropic risk measure for some $\theta \in (0, \infty)$. Furthermore, assume $\exp(\theta J(\cdot, u)) \in L^1(U, \mathfrak{B}(U), \mu, \mathcal{Y})$ and $\exp(\theta J(\boldsymbol{y}, u))\partial_u J(\boldsymbol{y}, u) \leqslant C(\boldsymbol{y})$ for some $C \in L^1(U, \mathfrak{B}(U), \mu, \mathcal{Y})$, where for a.e. $\boldsymbol{y} \in U$, $\partial_u J(\boldsymbol{y}, u)$ denotes the Fréchet derivative of $J(\boldsymbol{y}, u)$ at u. Then the Fréchet derivative at u of the composite functional $\mathcal{R}(J(\cdot, u))$ is given by*

$$\partial_u \mathcal{R}(J(\cdot, u)) = \frac{1}{\mathbb{E}\left[\exp\left(\theta \, J(\cdot, u)\right)\right]} \mathbb{E}[\exp(\theta \, J(\cdot, u))\partial_u J(\cdot, u)] \, . \tag{3.16}$$

Proof. The application of the chain rule gives

$$\partial_u \mathcal{R}(J(\cdot,u)) = \frac{1}{\mathbb{E}\left[\exp\left(\theta\, J(\cdot,u)\right)\right]} \partial_u \big(\mathbb{E}[\exp(\theta\, J(\cdot,u))]\big)\,.$$

From Example 3.3.12 and the chain rule we obtain the Fréchet derivative of the integrand $\partial_u \exp(\theta\, J(\boldsymbol{y},u)) = \exp(\theta\, J(\boldsymbol{y},u))\partial_u J(\boldsymbol{y},u)$ for a.e. $\boldsymbol{y} \in U$. Moreover, by assumption we have $\mathbb{E}[\exp(\theta\, J(\cdot,u))] \in L^1(U,\mathfrak{B}(U),\mu,\mathcal{Y})$ and $\exp\left(\theta J(\boldsymbol{y},u)\right)\partial_u J(\boldsymbol{y},u) \leqslant C(\boldsymbol{y})$ for some $C \in L^1(U,\mathfrak{B}(U),\mu,\mathcal{Y})$. Thus, from Theorem 3.3.11 we conclude that

$$\partial_u \big(\mathbb{E}[\exp(\theta\, J(\cdot,u))]\big) = \mathbb{E}[\exp(\theta\, J(\cdot,u))\partial_u J(\cdot,u)]\,,$$

as required. □

3.4 Existence and uniqueness of solutions

In this section we will discuss conditions that guarantee existence and uniqueness of solutions of the problem (3.2) - (3.3).
Let $\mathcal{Z},\mathcal{X}$ and $\mathcal{Y}$ be reflexive Banach spaces and $j : L^q(\mathcal{X}) \times \mathcal{Z} \to \mathbb{R}$, with $j(u,z) := \mathcal{R}(\mathcal{J}(u)(\boldsymbol{y})) + \alpha R(z)$ and $M : U \times \mathcal{X} \times \mathcal{Z}$ be continuous for almost all $\boldsymbol{y} \in U$.

Assumption 3.4.1.

(i) $\mathcal{Z}_{ad} \subset \mathcal{Z}$ *is convex and closed.*

(ii) $\mathcal{X}_{ad} \subset L^q(\mathcal{X})$ *is convex and closed, such that the feasible set is nonempty.*

$$F_{\mathrm{ad}} := \{(u,z) \in \mathcal{Z}_{\mathrm{ad}} \times \mathcal{X}_{\mathrm{ad}} \,:\, M(\cdot,u(\cdot),z) = 0 \in L^q(\mathcal{Y}')\} \neq \varnothing\,.$$

(iii) The model equation $M(\cdot,u(\cdot),z) = 0 \in L^q(\mathcal{Y}')$ *has a bounded solution operator* $\mathcal{Z}_{ad} \ni z \mapsto u \in L^q(\mathcal{X})$.

(iv) $L^q(\mathcal{X}) \times \mathcal{Z} \ni (u,z) \mapsto M(\cdot,u(\cdot),z) \in L^q(\mathcal{Y}')$ *is continuous under weak convergence.*

(v) Assume that $j(u,z) := \mathcal{R}(\mathcal{J}(u)) + \alpha R(z)$ *is weakly sequentially*[18] *lower semicontinuous, i.e.,* $(u_k,z_k) \rightharpoonup (u,z)$ *implies* $j(u,z) \leqslant \liminf_{k\to\infty} j(u_k,z_k)$.

Remark 3.4.2. *In view of Corollary 3.3.4, Assumption 3.4.1 (v) is true if Assumption 3.3.1 holds for* J*, if* $\mathcal{R}$ *satisfies (R1) and (R3), and if* $R(z)$ *is weakly (sequentially) lower semicontinuous. This follows from the superadditivity of the limit inferior:*

$$(f+g)(u,z) \leqslant \liminf_{k\to\infty} f(u_k,z_k) + \liminf_{k\to\infty} g(u_k,z_k) \leqslant \liminf_{k\to\infty} (f+g)(u_k,z_k)\,,$$

for $(u_k,z_k) \rightharpoonup (u,z)$ *and weakly (sequentially) lower semicontinuous functionals* f,g.

Theorem 3.4.3. *Let Assumption 3.4.1 hold and let* $\mathcal{Z}_{\mathrm{ad}}$ *be bounded or* j *be coercive. Then* (3.2) – (3.3) *has a solution.*

[18] On a metric space X, (weak) sequential lower semicontinuity at $x \in X$, defined by $f(x) \leqslant \liminf_{k\to\infty} f(x_k)$ for $x_k \to x$ (weakly, i.e., $x_k \rightharpoonup x$), is equivalent to (weak) lower semicontinuity, defined by $\{x \in X : f(x) \leqslant c\}$ being closed (in the weak topology) for all $c \in \mathbb{R}$. For convex functions, a well-known consequence of the Hahn–Banach theorem (see, e.g., [135, Theorem 3.2]) is that, the lower semicontinuity with respect to the strong topology of X is equivalent to the weak (or weak sequential) lower semicontinuity.

Proof. Let $(u_k, z_k) \subset F_{\text{ad}}$ be a minimizing sequence, i.e.,

$$\lim_{k\to\infty} j(u_k, z_k) = j^* := \inf_{(u,z)\in F_{\text{ad}}} j(u, z) \geqslant -\infty\,,$$

where j is defined in Assumption 3.4.1 (v). The minimizing sequence (of controls) is bounded, since either $\mathcal{Z}_{\text{ad}}$ or $j(u, z)$ is coercive. The boundedness of the control-to-state mapping (see, (iii)) ensures the boundedness of the state sequence. From the reflexivity of $\mathcal{Z} \times L^q(\mathcal{X})$ we conclude that there is a weakly convergent subsequence $(u_{k_i}, z_{k_i}) \subset (u_k, z_k)$ and (u^*, z^*) with $(u_{k_i}, z_{k_i}) \rightharpoonup (u^*, z^*)$ as $i \to \infty$.[19] By Assumption 3.4.1 (i) and (ii), the sets $\mathcal{Z}_{\text{ad}}$ and $\mathcal{X}_{\text{ad}}$ are weakly sequentially[20] closed as they are convex and closed sets in Banach spaces, see, e.g., [5, Theorem 3.7]. This implies together with the continuity under weak convergence of the model, Assumption 3.4.1 (iv), that the feasible set F_{ad} is weakly sequentially closed and thus $(u^*, z^*) \in F_{\text{ad}}$. By Assumption 3.4.1 (v) we obtain

$$j^* = \lim_{i\to\infty} j(u_{k_i}, z_{k_i}) \geqslant j(u^*, z^*) \geqslant j^*\,,$$

where the last inequality follows from the feasibility of (u^*, z^*). In particular, we have $j^* > -\infty$. □

Theorem 3.4.4. *Let the assumptions of the preceeding theorem hold. The solution is unique if $\alpha > 0$ and $R(z)$ is strictly convex and if the state-to-control mapping $A : u \mapsto z$ is injective or $\mathcal{R}(\mathcal{J}(u))$ is strictly convex.*

Proof. We know that $\mathcal{R}(\mathcal{J}(u))$ is convex in u and $R(z)$ is strictly convex in z. Assume there are two minimizers (u_1, z_1) and (u_2, z_2), i.e., $j(u_1, z_1) = j(u_2, z_2) \leqslant j(u, z)\,\forall (u, z) \in \mathcal{X}_{\text{ad}} \times \mathcal{Z}_{\text{ad}}$, where j is defined in Assumption 3.4.1 (v). Let $(\tilde{u}, \tilde{z}) := (\frac{u_1+u_2}{2}, \frac{z_1+z_2}{2})$ and $\mathcal{R}(\mathcal{J}(u))$ be strictly convex, then

$$\begin{aligned} j(\tilde{u}, \tilde{z}) = \mathcal{R}(\mathcal{J}(\tilde{u})) + \alpha R(\tilde{z}) &< \frac{1}{2}(\mathcal{R}(\mathcal{J}(u_1)) + \alpha R(z_1)) + \frac{1}{2}(\mathcal{R}(\mathcal{J}(u_2)) + \alpha R(z_2)) \\ &= \frac{1}{2} j(u_1, z_1) + \frac{1}{2} j(u_2, z_2) \\ &= j(u_1, z_1)\,, \end{aligned}$$

which contradicts the assumption of two minimizers. If $\mathcal{R}(\mathcal{J}(u))$ is only convex and $z_1 \neq z_2$, then the above inequality remains true. If $z_1 = z_2$ the injectivity of A implies $u_1 = u_2$ and thus gives uniqueness. □

3.5 Reduced formulation of the optimization problem

Assume that our mathematical model $\mathcal{M}$ is well-posed. That is, there exists a unique solution $u \in L^q(\mathcal{X})$ such that for $z \in \mathcal{Z}$ the constraint holds with equality, i.e., $\mathcal{M}(\cdot, u(\cdot), z) = 0$ in $L^q(\mathcal{Y}')$. We call the operator $S(z) : U \to L^q(\mathcal{X})$ defined by $u(\boldsymbol{y}) = S(z)(\boldsymbol{y})$ for each $z \in \mathcal{Z}$, the solution operator of the model $\mathcal{M}$. The formulation of the so-called reduced

[19] Every bounded sequence in a reflexive Banach space has a weakly convergent subsequence, see, e.g., [89, Theorem 1.17]

[20] Weakly sequentially closed and weakly closed in the sense of weak topology are equivalent in reflexive Banach spaces

problem is based on the substitution of the state $u \in L^q(\mathcal{X})$ by the solution operator for given control $S(z)$. For $\alpha > 0$, the reduced optimization problem is of the general form

$$\min_{z \in \mathcal{Z}_{\mathrm{ad}}} \mathcal{R}(\mathcal{J}(Sz)) + \alpha R(z)\,. \tag{3.17}$$

In order to ensure well-posedness we make the following assumptions.

Assumption 3.5.1.

(i) The mapping $S(z) : U \to \mathcal{X}$ is strongly μ-measurable for all $z \in \mathcal{Z}_{\mathrm{ad}}$.

(ii) There exists a nonnegative increasing function $\rho : [0,\infty) \to [0,\infty)$, and a nonnegative random variable $C \in L^p(U, \mathfrak{B}(U), \mu, \mathcal{X})$ with $p \in [1,\infty]$ satisfying

$$\|S(z)\|_{\mathcal{X}} \leqslant C\rho(\|z\|_{\mathcal{Z}}) \quad \mu\text{-a.s.}\,,$$

for all $z \in \mathcal{Z}_{ad}$.

(iii) Weak convergence of control sequences $z_j \rightharpoonup z \in \mathcal{Z}_{ad}$ implies weak convergence of the soutions $S(z_j) \rightharpoonup S(z)$ in X μ-a.s.

(iv) There exists an open set $E \subset \mathcal{Z}$ with $\mathcal{Z}_{ad} \subset E$ such that the solution map $z \mapsto S(z) : E \to L^q(U, \mathfrak{B}(U), \mu, \mathcal{X})$ is continuously Fréchet differentiable.

By Assumption 3.5.1 (i) – (ii) it holds that $S(z) \in L^q(\mathcal{X})$ is bounded for $z \in \mathcal{Z}_{ad}$. If in addition, Assumption 3.5.1 (iii) holds, we have weak convergence $S(z_n) \rightharpoonup S(z)$ in $L^q(\mathcal{X})$.

Theorem 3.5.2 ([106, Proposition 3.8])**.** *Let Assumption 3.5.1 (i) – (iii) and Assumption 3.3.1 hold. If $\mathcal{R}$ is proper, closed, monotonic, convex, and subdifferentiable at $\mathcal{J}(S(z))$ for some $z \in \mathcal{Z}_{\mathrm{ad}}$, then the composite functional $(R \circ J \circ S) : \mathcal{Z}_{\mathrm{ad}} \to \mathcal{R}$ is weakly lower semicontinuous at z.*

Assumption 3.5.3.

(i) $\mathcal{Z}_{ad} \subset \mathcal{Z}$ is convex and closed.

(ii) Let the feasible set

$$\widetilde{F}_{\mathrm{ad}} := \{z \in \mathcal{Z}_{\mathrm{ad}} \,:\, M(\cdot, S(z)(\cdot), z) = 0 \in L^q(\mathcal{Y}')\,, S(z) \in \mathcal{X}_{\mathrm{ad}}\} \neq \varnothing$$

be nonempty.

(iii) $\mathcal{Z} \ni z \mapsto M(\cdot, S(z)(\cdot), z) \in L^q(\mathcal{Y}')$ is continuous under weak convergence.

Theorem 3.5.4 ([106, Proposition 3.12])**.** *Let Assumption 3.5.3, Assumption 3.5.1 (i) – (iii) and Assumption 3.3.1 hold. Let $\mathcal{R} : L^p(\mathcal{X}) \to \mathcal{R}$ be a proper, closed, convex, and monotonic risk measure, taking values in $\mathbb{R}$ and let $R : \mathcal{Z} \to \mathbb{R}$ be proper, closed, and convex. Suppose that $\mathcal{Z}_{\mathrm{ad}}$ is bounded or $\mathcal{R}(\mathcal{J}(Sz)) + \alpha R(Z)$ is coercive, then* (3.17) *has a solution.*

Proof. Since $\mathcal{R}$ takes values in $\mathbb{R}$ and fulfills the axioms (R1) and (R3), it is continuous and subdifferentiable by Lemma 3.2.4. In particular, $\mathcal{R}$ is subdifferentiable at $\mathcal{J}(S(z))$ for

any $z \in \mathcal{Z}_{\text{ad}}$. By Theorem 3.5.2, the composite functional $(\mathcal{R} \circ \mathcal{J} \circ S) : \mathcal{Z}_{\text{ad}} \to \mathbb{R}$ is weakly (sequentially) lower semicontinuous. Let $z_k \subset \widetilde{F}_{\text{ad}}$ be a minimizing sequence with

$$\lim_{k\to\infty} \mathcal{R}(\mathcal{J}(S(z_k))) + \alpha R(z_k) = j^* := \inf_{z\in\widetilde{F}_{\text{ad}}} \mathcal{R}(\mathcal{J}(S(z))) + \alpha R(z)\,.$$

The minimizing sequence is bounded, since either $\mathcal{Z}_{\text{ad}}$ is bounded or $\mathcal{R}(\mathcal{J}(Sz)) + \alpha R(Z)$ is coercive. From the reflexivity of $\mathcal{Z}$ we conclude that there is a weakly convergent subsequence $z_{k_i} \subset z_k$ and z^* with $z_{k_i} \rightharpoonup z^*$ as $i \to \infty$. By Assumption 3.5.3 the feasible set $\widetilde{F}_{\text{ad}}$ is weakly sequentially closed and thus $z^* \in \widetilde{F}_{\text{ad}}$. Since $\mathcal{R}(\mathcal{J}(S(z))) + \alpha R(z)$ is weakly sequentially lower continuous, we obtain

$$j^* = \lim_{i\to\infty} \mathcal{R}(\mathcal{J}(S(z_{k_i}))) + \alpha R(z_{k_i}) \geqslant \mathcal{R}(\mathcal{J}(S(z^*))) + \alpha R(z^*) \geqslant j^*\,,$$

where the last inequality follows from the feasibility of z^*. In particular $j^* > -\infty$. □

Remark 3.5.5. *Let z^* be a minimizer of the reduced problem* (3.17)*, then $(S(z^*), z^*)$ is a solution of* (3.2) – (3.3)*. Thus, solving the reduced problem is equivalent to solving* (3.2) – (3.3)*. In the following discussions we will hence only consider the reduced problem.*

3.6 Optimality conditions

Recall that the objective functional in our optimization problem (3.17) has the form $(\mathcal{R} \circ \mathcal{J} \circ S) : \mathcal{Z}_{\text{ad}} \to \mathbb{R}$. In order to apply the chain rule, the outer functional $\mathcal{R}$ must be at least Hadamard directionally differentiable, whereas the inner mapping $(\mathcal{J} \circ S)$ must be at least Gâteaux directionally differentiable. Therefore, under Assumption 3.5.1 (iv) and Assumption 3.3.1, the composite functional will be Gâteaux directionally differentiable provided $\mathcal{J}$ is locally Lipschitz.

Theorem 3.6.1 ([106, Proposition 3.13])**.** *Let $\mathcal{Z}_{\text{ad}} \subset \mathcal{Z}$ be a nonempty and convex subset of a Banach space and $z^* \in \mathcal{Z}_{\text{ad}}$ be a solution of* (3.17)*. Let $\mathcal{R}$ be Hadamard directionally differentiable at $\mathcal{J}(S(z^*))$ and $\mathcal{J}$ be Gâteaux directionally differentiable at $S(z^*)$ and locally Lipschitz. Moreover let Assumption 3.5.1 (iv) and Assumption 3.3.1 hold, and let $R(z)$ be Gâteaux directionally differentiable. Then the following optimality condition holds:*

$$\sup_{\xi\in\partial\mathcal{R}(\mathcal{J}(S(z^*)))} \mathbb{E}[\langle \xi, \mathcal{J}'(S(z^*); S(z^*)'\delta z\rangle] + \alpha R'(z^*;\delta z) \geqslant 0 \quad \forall \delta z \in T_{\mathcal{Z}_{ad}}(z^*)\,,$$

where $T_{\mathcal{Z}_{\text{ad}}}(z^*)$ denotes the tangent cone of $\mathcal{Z}_{\text{ad}}$ at z^*, defined by

$$T_{\mathcal{Z}_{ad}}(z^*) := \{d \in \mathcal{Z} \,:\, \exists \tau_k \searrow 0\,, \exists d_k \to d \text{ in } \mathcal{Z} \,:\, z^* + \tau_k d_k \in \mathcal{Z}_{\text{ad}}\, \forall k\}\,.$$

If composite functional is Gâteaux differentiable, then the variational inequality holds.

Theorem 3.6.2 ([89, Theorem 1.46])**.** *Let $\mathcal{Z}$ be a Banach space and $\mathcal{Z}_{\text{ad}}$ be nonempty and convex. Moreover, let $J : \widetilde{\mathcal{Z}} \to \mathbb{R}$ be defined on an open neighborhood of $\mathcal{Z}_{\text{ad}}$. Let z^* be a local solution of*

$$\min_{z\in\mathcal{Z}} J(z) \quad \text{s.t. } z \in \mathcal{Z}_{\text{ad}}\,,$$

and let J be Gâteaux differentiable at z^*. Then it holds for $z^* \in \mathcal{Z}_{\text{ad}}$

$$\langle J'(z^*), z - z^*\rangle_{\mathcal{Z}',\mathcal{Z}} \geqslant 0 \quad \forall z \in \mathcal{Z}_{\text{ad}}\,.$$

Furthermore, if in addition

- *J is convex on $\mathcal{Z}_{\mathrm{ad}}$, then this condition is necessary and sufficient for global optimality.*
- *J is strictly convex on $\mathcal{Z}_{\mathrm{ad}}$, then there is at most one solution z^*.*
- *$\mathcal{Z}$ is reflexive, $\mathcal{Z}_{\mathrm{ad}}$ is closed and convex and J is convex and continuous with $\lim_{z\in\mathcal{Z}_{\mathrm{ad}},\|z\|_{\mathcal{Z}}\to\infty} J(z) = \infty$, then there exists a (global) solution.*
- *$\mathcal{Z}$ is a Hilbert space, $\mathcal{Z}_{\mathrm{ad}}$ is closed and convex, then denoting the orthogonal projection P onto $\mathcal{Z}_{\mathrm{ad}}$[21] the following conditions are equivalent for $z^*, y \in \mathcal{Z}_{\mathrm{ad}}$ and $\gamma > 0$*

$$\begin{aligned} \langle y, z - z^*\rangle_{\mathcal{Z}} &\geqslant 0 \quad \forall z \in \mathcal{Z}_{\mathrm{ad}} \\ z^* - P(z^* - \gamma y) &= 0\,. \end{aligned}$$

3.7 Parametric linear forward operators

In this section we consider linear model constraints, i.e., constraints of the form

$$\mathcal{M}(\cdot, u(\cdot), z) = \mathcal{A}(\cdot)u(\cdot) - \mathcal{B}z = 0 \quad \text{in } L^q_\mu(\mathcal{Y}')\,, \tag{3.18}$$

for bounded and linear operators $\mathcal{A} : L^q_\mu(\mathcal{X}) \to L^q_\mu(\mathcal{Y}')$ and $\mathcal{B} : \mathcal{Z} \to \mathcal{Y}'$. Note that we assume that z is constant in $\boldsymbol{y}$, i.e., $\boldsymbol{y} \mapsto z(\boldsymbol{y}) = z$ for all $\boldsymbol{y} \in U$. Oftentimes the solution operator of the model equation is more regular with respect to the parameters $\boldsymbol{y} \in U$, e.g., it is continuous. To this end, let U be a nonempty topological space and let the parametric linear operator from $\mathcal{X}$ to $\mathcal{Y}'$ with parameter domain U be a continuous map

$$A : U \to \mathcal{L}(\mathcal{X}, \mathcal{Y}')\,, \quad \boldsymbol{y} \mapsto A(\boldsymbol{y})\,.$$

In this case one can impose the model constraint pointwise for all $\boldsymbol{y} \in U$:

$$\langle A(\boldsymbol{y})u(\boldsymbol{y}), v\rangle_{\mathcal{Y}',\mathcal{Y}} = \langle \mathcal{B}z, v\rangle_{\mathcal{Y}',\mathcal{Y}} \quad \forall v \in \mathcal{Y}\,, \forall \boldsymbol{y} \in U\,. \tag{3.19}$$

or equivalently

$$\mathcal{M}(\boldsymbol{y}, u(\boldsymbol{y}), z) = A(\boldsymbol{y})u(\boldsymbol{y}) - \mathcal{B}z = 0 \quad \forall \boldsymbol{y} \in U\,,$$

Theorem 3.7.1 ([64, Theorem 1.1.1 and Lemma 1.1.3])**.** *Let $A(\boldsymbol{y})$ be bijective for all $\boldsymbol{y} \in U$. Then (3.19) has a unique solution $u : U \to \mathcal{X}$. The solution $\boldsymbol{y} \mapsto u(\boldsymbol{y})$ is continuous if $\boldsymbol{y} \mapsto z(\boldsymbol{y})$ is continuous. Moreover, if U is a compact Hausdorff space, then there exists $a_{\min}, a_{\max} > 0$ such that*

$$\|A(\boldsymbol{y})\|_{\mathcal{L}(\mathcal{X},\mathcal{Y}')} \leqslant a_{\max} \quad \textit{and} \quad \|A(\boldsymbol{y})^{-1}\|_{\mathcal{L}(\mathcal{Y}',\mathcal{X})} \leqslant 1/a_{\min} \quad \forall \boldsymbol{y} \in U\,.$$

One can show that under the measure μ the parametric model constraint (3.19) is equivalent to the weak parameter form of the model constraint (3.18). This is discussed in more detail in the following subsection.

[21] see Chapter 2

3.7.1 Equivalence between parametric and weak parameter formulation

We will extend the parametric linear operators to operators between Lebesgue spaces of vector-valued functions. To this end, not that the operators

$$\begin{aligned} \mathcal{A}: &\quad \mathcal{C}(U,\mathcal{X}) \to \mathcal{C}(U,\mathcal{Y}'), \quad v \mapsto [\boldsymbol{y} \mapsto A(\boldsymbol{y})v] \\ \mathcal{A}^{-1}: &\quad \mathcal{C}(U,\mathcal{Y}') \to \mathcal{C}(U,\mathcal{X}), \quad w \mapsto [\boldsymbol{y} \mapsto (A(\boldsymbol{y}))^{-1}w] \end{aligned} \tag{3.20}$$

are well-defined, inverse to eachother with norms $\|\mathcal{A}\| \leqslant a_{\max}$ and $\|\mathcal{A}^{-1}\| \leqslant 1/a_{\min}$. This result can be extended to Lebesgue spaces of vector-valued functions. To this end, let $\mathfrak{B}(U)$ be the Borel σ-algebra on U, and let μ be a finite measure on $(U,\mathfrak{B}(U))$.

Theorem 3.7.2 ([64, Corollary 1.1.6])**.** *For all $1 \leqslant q < \infty$, the operator $\mathcal{A}$ in* (3.20) *extends uniquely to a boundedly invertible operator on the Lebesgue–Bochner spaces*

$$\mathcal{A}: L^q_\mu(U,\mathcal{X}) \to L^q_\mu(U,\mathcal{Y}')\,. \tag{3.21}$$

The norms of $\mathcal{A}$ and $\mathcal{A}^{-1}$ are bounded by $a_{\max}$ and $1/a_{\min}$, respectively.

The applications of the operators $\mathcal{A}$ on $L^q(U,\mathcal{X})$ and $\mathcal{A}^{-1}$ on $L^q(U,\mathcal{Y})$ is equal to pointwise application of $A(\boldsymbol{y})$ on $\mathcal{X}$ and $A(\boldsymbol{y})^{-1}$ on $\mathcal{Y}'$ up to μ-equivalence. As a corollary to this result we get:

Corollary 3.7.3 ([64, Corollary 1.1.8])**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be separable Banach spaces, $1 \leqslant r < \infty$, and let q be the Hölder conjugate of r. If $\mathcal{B}z \in L^q_\mu(U,\mathcal{Y}')$, then there is a unique $\tilde{u} \in L^q_\mu(U,\mathcal{X})$ such that*

$$\int_U \langle A(\boldsymbol{y})\tilde{u}(\boldsymbol{y}), w(\boldsymbol{y})\rangle_{\mathcal{Y}',\mathcal{Y}}\,\mathrm{d}\mu(\boldsymbol{y}) = \int_U \langle \mathcal{B}z, w(\boldsymbol{y})\rangle_{\mathcal{Y}',\mathcal{Y}}\,\mathrm{d}\mu(\boldsymbol{y})\,, \quad \forall\, w(\boldsymbol{y}) \in L^r_\mu(U,\mathcal{Y})\,. \tag{3.22}$$

Moreover, the solution $u^{\boldsymbol{y}}$ of (3.19) *is a version of $\tilde{u}^{\boldsymbol{y}}$, i.e., $\|u - \tilde{u}\|_{L^q_\mu(U,\mathcal{X})} = 0$, or equivalently $u = \tilde{u}$ μ-a.e. in $\mathcal{X}$.*

This result shows the equivalence between (3.18) and (3.19).
Since the measure μ cannot distinguish between the states obtained for both formulations of the constraint, in many cases the deterministic solution of the optimal control problem under uncertainty is equal for both formulations. We illustrate this using a linear quadratic optimal control problem.

3.7.2 Linear quadratic optimal control

We are interested in solving an optimal control problem in the presence of uncertainty by minimizing the averaged least square difference of the state u and a desired target state $\widehat{u}$. The state u is the solution of a linear operator equation, steered by a control function, and depends on a parameter vector. The parameter vector is in principle infinite-dimensional, and in practice might need a large finite number of terms for accurate approximation.
Our goal of computation is the following optimal control problem

$$\min_{z\in\mathcal{Z}_{\mathrm{ad}},u\in\mathcal{X}_{\mathrm{ad}}} J(u,z)\,, \quad J(u,z) := \frac{1}{2}\int_U \|\mathcal{Q}u - \widehat{u}\|^2_{\mathcal{Y}}\,\mathrm{d}\mu(\boldsymbol{y}) + \frac{\alpha}{2}\|z\|^2_{\mathcal{Z}}\,, \tag{3.23}$$

subject to the linear operator equation in $L^q_\mu(U,\mathcal{Y}')$

$$\mathcal{A}u = \mathcal{B}z\,, \tag{3.24}$$

for $1 \leqslant q < \infty$, a Hilbert space $\mathcal{Z}$ with $\mathcal{Z}_{\text{ad}} \subset \mathcal{Z}$, $\mathcal{X}_{\text{ad}} \subset L^q_\mu(U, \mathcal{X})$, and a Hilbert space $\mathfrak{J}$, $\widehat{u} \in \mathfrak{J}$, $\mathcal{Q} \in \mathcal{L}(\mathcal{X}, \mathfrak{J})$, $\mathcal{B} \in \mathcal{L}(\mathcal{Z}, \mathcal{Y}')$. In particular, the operators $\mathcal{B}$ and $\mathcal{Q}$ are not dependent on $\boldsymbol{y}$ and thus can be uniformly bounded for all $\boldsymbol{y}$, i.e., $\|\mathcal{B}\|_{\mathcal{L}(\mathcal{Z},\mathcal{Y}')} \leqslant C_1$ and $\|\mathcal{Q}\|_{\mathcal{L}(\mathcal{X},\mathfrak{J})} \leqslant C_2$ for some $C_1, C_2 > 0$ and all $\boldsymbol{y} \in U$. This implies in particular, that $\mathcal{B}z \in L^p_\mu(U, \mathcal{Y}')$ for all p and all deterministic controls $z \in \mathcal{Z}$ and $\mathcal{Q}u \in L^p_\mu(U, \mathfrak{J})$ for all $u \in L^p_\mu(U, \mathcal{X})$.

Theorem 3.7.4. *Let $\alpha \geqslant 0$, $\mathcal{Z}_{\text{ad}} \subset \mathcal{Z}$ convex, closed and in the case $\alpha = 0$ bounded. Let $\mathcal{X}_{\text{ad}} \subset L^q_\mu(U, \mathcal{X})$ with $q = 2$ be convex and closed, such that* (3.24) *has a feasible point. Then problem* (3.23) – (3.24) *has a solution (z^*, u^*). If $\alpha > 0$ then the solution is unique.*

Proof. Observe that the objective function can be written as $J(u, z) = \frac{1}{2}\|\mathcal{Q}u - \widehat{u}\|^2_{L^2(U,\mathfrak{J})} + \frac{\alpha}{2}\|z\|^2_{\mathcal{Z}}$, and $L^2(U, \mathfrak{J})$ is a Hilbert space. Since $\mathcal{A}$ has a bounded inverse, the result follows from [89, Theorem 1.43]. □

Substituting $u = \mathcal{A}^{-1}\mathcal{B}z$ into J gives $\hat{J}(z) := J(\mathcal{A}^{-1}\mathcal{B}z, z)$ and leads to the equivalent formulation of problem (3.23) – (3.24)

$$\min_{z \in \hat{\mathcal{Z}}_{\text{ad}}} \hat{J}(z)\,, \quad \hat{J}(z) := \frac{1}{2}\int_U \|\mathcal{Q}\mathcal{A}^{-1}\mathcal{B}z - \widehat{u}\|^2_{\mathfrak{J}}\,\mathrm{d}\mu(\boldsymbol{y}) + \frac{\alpha}{2}\|z\|^2_{\mathcal{Z}}\,, \tag{3.25}$$

where $\hat{\mathcal{Z}}_{\text{ad}} := \{z \in \mathcal{Z} : z \in \mathcal{Z}_{ad}, \mathcal{A}^{-1}\mathcal{B}z \in \mathcal{X}_{\text{ad}}\}$.

Remark 3.7.5. *From Corollary 3.7.3 we know that the solutions of* (3.19) *and the solution of* (3.24) *are μ-almost everywhere identical. In consequence, the problem* (3.23) *subject to* (3.24) *is equivalent to* (3.23) *subject to* (3.19) *in the sense that they have the same solution.*

4 Examples of optimal control problems

In this chapter we present three optimal control problems that fit into the framework of the previous chapter. The first example is an optimal control problem with an elliptic PDE constraint, with quadratic tracking-type objective functional, and with expected value as a risk measure. This example is based on [76]. The second example is an optimal control problem that is subject to a parabolic PDE constraint. The objective function is a tracking-type functional composed with the expected value or the more conservative entropic risk measure. This example is based on [77]. In both problems, we employ additional constraints on the control. The objective function of the third problem is again a tracking-type functional composed with the expected value or the entropic risk measure, and the constraint is an abstract parametric linear operator equation.

The novelty of this chapter lies in the generalization of the results we derived for elliptic and parabolic problems in [76], [77] to abstract analytic parametric linear operator equations.

Optimal control problems subject to parametric linear operator equations have been considered in [111] in conjunction with the expected value as a risk measure and without additional constraints on the control. Hence, the problem in [111] can be formulated as a saddle-point problem and be solved efficiently. Additional control constraints lead to a projection operator in the optimality conditions. Due to the projection, the resulting system of equations is no longer linear and hence the results obtained in [111] do not apply directly. Neither cover the results therein nonlinear risk measures, such as the entropic risk measure, which is considered in this chapter.

We restrict the analysis in this chapter to these two risk measures, since they are smooth and inherit the parametric regularity of the integrands, as we will see. Moreover, the expected value is a coherent measure of risk and the coherent entropic Value-at-Risk can be recovered from the entropic risk measure by the additional minimization with respect to the risk aversion parameters.

We emphasize that the theory about the existence and uniqueness of solutions developed in the previous chapter applies to all three example problems, and the elliptic and parabolic problems are special cases of the problem with abstract linear operator equation constraints. However, for better illustration, we provide the complete analysis for all examples.

4.1 Elliptic PDE constraint

Let $D \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$ denote a bounded domain with Lipschitz boundary ∂D and let $U = [-\frac{1}{2}, \frac{1}{2}]^{\mathbb{N}}$ denote a space of parameters. Let $\alpha \geqslant 0$ and $Z_{ad} \subset \mathcal{Z} = L^2(D)$ be closed, convex and in the case $\alpha = 0$ bounded. Let $\widehat{u} \in L^2(D)$ and consider the optimal control

problem

$$\min_{z\in L^2(D),u\in L^2_\mu(U,V)} J(u,z)\,,\quad J(u,z) := \frac{1}{2}\int_U \|u(\cdot,\boldsymbol{y})-\widehat{u}\|^2_{L^2(D)}\,\mathrm{d}\mu(\boldsymbol{y}) + \frac{\alpha}{2}\|z\|^2_{L^2(D)}\,, \quad (4.1)$$

subject to

$$-\nabla\cdot(a(\boldsymbol{x},\boldsymbol{y})\nabla u(\boldsymbol{x},\boldsymbol{y})) = z(\boldsymbol{x}) \qquad \boldsymbol{x}\in D,\quad \boldsymbol{y}\in U, \quad (4.2)$$

$$u(\boldsymbol{x},\boldsymbol{y}) = 0 \qquad \boldsymbol{x}\in \partial D,\quad \boldsymbol{y}\in U, \quad (4.3)$$

$$z_{\min}(\boldsymbol{x}) \leqslant z(\boldsymbol{x}) \leqslant z_{\max}(\boldsymbol{x}) \qquad \text{a.e. in } D. \quad (4.4)$$

To ensure wellposedness of (4.1)-(4.3) we make the following assumtions:

(AE1) Let $z_{\min}, z_{\max}\in L^2(D)$ with $z_{\min}\leqslant z_{\max}$ a.e. in D. Then the feasible set of controls

$$\mathcal{Z}_{\mathrm{ad}} := \{z\in L^2(D) : z_{\min}\leqslant z\leqslant z_{\max} \text{ a.e. in } D\}$$

(AE2) The sequence of parameters $\boldsymbol{y} = (y_j)_{j\geqslant 1}$ is independently and identically distributed (i.i.d.) uniformly in $[-\frac{1}{2},\frac{1}{2}]$ for each $j\in\mathbb{N}$, i.e., $\boldsymbol{y}$ is distributed on U with probability measure μ, where $\mu(\mathrm{d}\boldsymbol{y}) = \bigotimes_{j\geqslant 1}\mathrm{d}y_j = \mathrm{d}\boldsymbol{y}$.

(AE3) The input uncertainty is described by the diffusion coefficient $a(\boldsymbol{x},\boldsymbol{y})$ in (4.2), which is assumed to depend linearly on the parameters y_j, i.e.

$$a(\boldsymbol{x},\boldsymbol{y}) := a_0(\boldsymbol{x}) + \sum_{j\geqslant 1} y_j\psi_j(\boldsymbol{x})\,,\quad \boldsymbol{x}\in D\,,\quad \boldsymbol{y}\in U\,. \quad (4.5)$$

(AE4) Let $a_0(\cdot)\in L^\infty(D)$, $\psi_j(\cdot)\in L^\infty(D)$ for all $j\geqslant 1$, and $(\|\psi_j\|_{L^\infty})_{j\geqslant 1}\in \ell^1(\mathbb{N})$.

(AE5) The uniform ellipticity assumption holds, i.e.

$$0 < a_{\min}\leqslant a(\boldsymbol{x},\boldsymbol{y})\leqslant a_{\max} < \infty\,,\quad \boldsymbol{x}\in D\,,\quad \boldsymbol{y}\in U\,,$$

for some positive real numbers $a_{\min}$ and $a_{\max}$.

4.1.1 Weak formulation

We define $V := H^1_0(D)$ and its (topological) dual space $V' := H^{-1}(D)$, and identify $L^2(D)$ with its own dual. Let $\langle\cdot,\cdot\rangle_{V',V}$ denote the duality pairing between V' and V. The norm and inner product in V are defined as usual by

$$\|v\|_V := \|\nabla v\|_{L^2(D)},\quad \langle v_1,v_2\rangle_V := \langle\nabla v_1,\nabla v_2\rangle_{L^2(D)}.$$

We introduce the continuous embedding operators $E_1 : L_2(D)\to V'$ and $E_2 : V\to L_2(D)$, with the embedding constants $c_1,c_2>0$ for the norms

$$\|E_1 v\|_{V'}\leqslant c_1\|v\|_{L^2(D)} \quad (4.6)$$

$$\|E_2 v\|_{L^2(D)}\leqslant c_2\|v\|_V. \quad (4.7)$$

Based on this function space setting, the PDE (4.2) and (4.3) can be stated in the parametric variational form: For fixed $\boldsymbol{y}\in U$ and given $E_1 z\in V'$, find $u(\boldsymbol{y})\in V$ such that

$$\int_D a(\boldsymbol{x},\boldsymbol{y})\nabla u(\boldsymbol{x},\boldsymbol{y})\cdot\nabla v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \langle E_1 z, v\rangle_{V',V}\quad \forall v\in V\,. \quad (4.8)$$

Note that by the identification of $L^2(D)$ with its dual $(L^2(D))'$, we have $\langle E_1 z, v\rangle_{V',V} = \int_D z\, E_2 v \,\mathrm{d}\boldsymbol{x}$. Moreover, by (AE5), the parametric bilinear form, defined as

$$b(\boldsymbol{y}, w, v) := \int_D a(\boldsymbol{x}, \boldsymbol{y}) \nabla w(\boldsymbol{x}, \boldsymbol{y}) \cdot \nabla v(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}, \quad \forall\, w, v \in V, \tag{4.9}$$

is continuous and coercive on $V \times V$, i.e.,

$$b(\boldsymbol{y}, v, v) \geqslant a_{\min} \|v\|_V, \quad \text{and} \quad b(\boldsymbol{y}, u, w) \leqslant a_{\max} \|w\|_V \|v\|_V,$$

for all $\boldsymbol{y} \in U$ and all $w, v \in V$. Hence, by the Lax–Milgram lemma (see, e.g., [47, Theorem 1 in Chapter 6]) the parametric variational problem (4.8) admits a unique solution $u(\boldsymbol{y}) \in V$ for each $z \in V'$ and fixed $\boldsymbol{y} \in U$, which satisfies the a-priori bound

$$\|u(\boldsymbol{y})\|_{H_0^1(D)} \leqslant \frac{\|E_1 z\|_{H^{-1}(D)}}{a_{\min}} \leqslant \frac{c_1 \|z\|_{H^{-1}(D)}}{a_{\min}}\,. \tag{4.10}$$

In particular, the operator $A(\boldsymbol{y}) \in \mathcal{L}(V, V')$ that can be associated with the bilinear form in (4.8), i.e.

$$\langle A(\boldsymbol{y}) w, v\rangle_{V',V} = b(\boldsymbol{y}, w, v), \quad \forall w, v \in V, \tag{4.11}$$

satisfies $\|A(\boldsymbol{y})\|_{\mathcal{L}(V,V')} \leqslant a_{\max}$ and $\|A(\boldsymbol{y})^{-1}\|_{\mathcal{L}(V',V)} \leqslant 1/a_{\min}$.

4.1.2 Reduced problem

We reformulate the optimal control problem (4.1), (4.2), (4.3), and (4.4) into the reduced formulation, i.e., a problem that only depends on the control.
In view of (4.6), (4.7) and (4.11) we can interpret the solution operator $S_{\boldsymbol{y}}$ of (4.8) as a linear continuous operator in $L^2(D)$

$$S_{\boldsymbol{y}} = E_2 A(\boldsymbol{y})^{-1} E_1 \tag{4.12}$$

The operator $S_{\boldsymbol{y}} : L^2(D) \to L^2(D)$ is the unique mapping, which for every $\boldsymbol{y} \in U$ assigns to each $f \in L^2(D)$ the unique solution $g \in L^2(D)$ of the weak problem: find $g \in V$ such that

$$b(\boldsymbol{y}, g, v) = \langle f, v\rangle \quad \forall v \in V.$$

Clearly the solution operator depends on the parameter $\boldsymbol{y} \in U$ as indicated by the subscript. Moreover, it is a self-adjoint operator, i.e.,

$$S_{\boldsymbol{y}} = S_{\boldsymbol{y}}^*, \tag{4.13}$$

where $S_{\boldsymbol{y}}^*$ is the adjoint operator of $S_{\boldsymbol{y}}$ defined by $\langle S_{\boldsymbol{y}}^* g, f\rangle = \langle g, S_{\boldsymbol{y}} f\rangle\ \forall f, g \in L^2(D)$. This is easily verified as for all $f, g \in L^2(D)$ we have $\langle S_{\boldsymbol{y}}^* g, f\rangle = \langle g, S_{\boldsymbol{y}} f\rangle = b(\boldsymbol{y}; S_{\boldsymbol{y}} g, S_{\boldsymbol{y}} f) = \langle S_{\boldsymbol{y}} g, f\rangle$. Thus, in the following we will omit the superscript $*$ in the notation for the adjoint operator $S_{\boldsymbol{y}}^*$.
By (4.12) and (4.8) it clearly holds that $u(\cdot, \boldsymbol{y}) = S_{\boldsymbol{y}} z$ for every $\boldsymbol{y} \in U$. Therefore, for each $\boldsymbol{y} \in U$ we can write the state u as a function of the control $z \in \mathcal{Z}$:

$$u(\cdot, \boldsymbol{y}, z) := S_{\boldsymbol{y}} z. \tag{4.14}$$

We call $u(\cdot, \boldsymbol{y}, z)$ the state corresponding to the control $z \in L^2(D)$. The optimal control problem (4.1), (4.2), (4.3), and (4.4) then becomes a quadratic problem in the Hilbert space $L^2(D)$: find

$$\min_{z \in \mathcal{Z}} J(z)\,, \quad J(z) := \frac{1}{2} \int_U \|S_{\boldsymbol{y}} z - \widehat{u}\|_{L^2(D)}^2 \,\mathrm{d}\boldsymbol{y} + \frac{\alpha}{2} \|z\|_{L^2(D)}^2\,. \tag{4.15}$$

4.1.3 Derivatives and adjoint problem

We observe that $\|S_{\boldsymbol{y}} z - \widehat{u}\|^2_{L^2(D)} \leqslant \frac{c_1 c_2}{a_{\min}} \|z\|_{L^2(D)} + \|\widehat{u}\|_{L^2(D)}$ is integrable. Furthermore, for each $\boldsymbol{y} \in U$, using (4.13), the Riesz representation of the Fréchet derivative of $F(z) := \|S_{\boldsymbol{y}} z - \widehat{u}\|^2_{L^2(D)}$ is given by $\nabla F(z) = S_{\boldsymbol{y}}(S_{\boldsymbol{y}} z - \widehat{u})$, which is bounded from above by $\frac{c_1 c_2}{a_{\min}} (\frac{c_1 c_2}{a_{\min}} \|z\|_{L^2(D)} + \|\widehat{u}\|_{L^2(D)})$. The same upper bound holds for the Fréchet derivative since the Riesz operator is an isometry. From Theorem 3.3.11 we conclude that the Riesz representation of the Fréchet derivative of $J(z)$ is given by

$$\nabla J(z) = \int_U S_{\boldsymbol{y}}(S_{\boldsymbol{y}} z - \widehat{u}) \,\mathrm{d}\boldsymbol{y} + \alpha z. \tag{4.16}$$

We use the symbol ∇J to emphasize that (4.16) is the Riesz representation (see Section 2.1) of the Fréchet derivative J' of J. Note that in general the Fréchet derivative J' at $z \in \mathcal{Z}$ of a functional $J : \mathcal{Z} \to \mathbb{R}$ is an element of the dual space $\mathcal{Z}'$, i.e., $J'(z) \in \mathcal{L}(\mathcal{Z}, \mathbb{R})$. We also call $\nabla J(z)$ the gradient of J at z.
Recalling that $S_{\boldsymbol{y}}$ is the solution operator of a PDE, and defining $f_{\text{adjoint}} := S_{\boldsymbol{y}} z - \widehat{u}$, it is easy to see that (4.16) can be computed by solving a second PDE, namely the adjoint PDE $S_{\boldsymbol{y}} f_{\text{adjoint}}$. The solution $q(\boldsymbol{y})$ of the adjoint PDE is called the adjoint state. Clearly, for each $\boldsymbol{y} \in U$ we can write the adjoint state as a function of the control $z \in \mathcal{Z}$:

$$q(\cdot, \boldsymbol{y}, z) := S_{\boldsymbol{y}}(S_{\boldsymbol{y}} z - \widehat{u}), \tag{4.17}$$

which we call the adjoint state corresponding to the control $z \in L^2(D)$. Using the adjoint state, the Fréchet derivative (4.16) can be written as

$$\nabla J(z) = \int_U q(\cdot, \boldsymbol{y}, z) \,\mathrm{d}\boldsymbol{y} + \alpha z.$$

4.1.4 Optimality conditions

Existence of an optimal control $z^* \in \mathcal{Z}_{\text{ad}}$ follows from Theorem 3.7.4. If in addition the regularization parameter $\alpha > 0$, then the optimal control z^* is the unique minimizer.
Since Fréchet differentiability implies Gâteaux differentiability, we obtain the following optimality conditions from Theorem 3.6.2: a control $z^* \in \mathcal{Z}_{\text{ad}}$ is the unique solution of (4.1), (4.2), (4.3), and (4.4) if and only if

$$\left.\begin{aligned} u(\cdot, \boldsymbol{y}, z^*) &= S_{\boldsymbol{y}} z^* \\ q(\boldsymbol{y}, \cdot, z^*) &= S_{\boldsymbol{y}}(u(\cdot, \boldsymbol{y}, z^*) - \widehat{u}) \end{aligned}\right\} \forall \boldsymbol{y} \in U,$$

$$\left\langle \int_U q(\cdot, \boldsymbol{y}, z^*) \,\mathrm{d}\boldsymbol{y} + \alpha z^*, z - z^* \right\rangle_{L^2(D)} \geqslant 0 \quad \forall z \in \mathcal{Z}_{\text{ad}},$$

where the first two equations are the state PDE and adjoint PDE, respectively, and the last condition is called variational inequality, which is equivalent to the following conditions:

- for arbitrary $\gamma > 0$, and $P_{\mathcal{Z}_{\text{ad}}} = \min(\max(z_{\min}, z^*), z_{\max})$ it holds for $z^* \in \mathcal{Z}_{\text{ad}}$ that
$$z^* - P_{\mathcal{Z}_{\text{ad}}}(z^* - \gamma \nabla J(z^*)) = 0,$$
- for $z^* \in \mathcal{Z}_{\text{ad}}$ it holds that
$$\nabla J(z^*)(\boldsymbol{x}) \begin{cases} = 0, & \text{if } z_{\min}(\boldsymbol{x}) < z^*(\boldsymbol{x}) < z_{\max}(\boldsymbol{x}), \\ \geqslant 0, & \text{if } z_{\min}(\boldsymbol{x}) = z^*(\boldsymbol{x}) < z_{\max}(\boldsymbol{x}), \\ \leqslant 0, & \text{if } z_{\min}(\boldsymbol{x}) < z^*(\boldsymbol{x}) = z_{\max}(\boldsymbol{x}), \end{cases} \quad \text{for a.e. } \boldsymbol{x} \in D.$$

- there exist $\mu_{\min}, \mu_{\max} \in L^2(D)$ with

$$\begin{aligned} \nabla J(z^*) + \mu_{\max} - \mu_{\min} &= 0, \\ z_{\min} \leqslant z^* &\leqslant z_{\max}, \\ \mu_{\max}, \mu_{\min} &\geqslant 0, \\ \mu_{\max}(z_{\max} - z^*) = \mu_{\min}(z^* - z_{\min}) &= 0. \end{aligned}$$

4.2 Parabolic PDE constraint

Let $D \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, denote a bounded physical domain with Lipschitz boundary, let $I := [0, T]$ denote the time interval with finite time horizon $0 < T < \infty$, and let $U := [-\frac{1}{2}, \frac{1}{2}]^{\mathbb{N}}$ denote a space of parameters. Given a target state $\widehat{u} \in \mathcal{X}$ and, that the regularization constants α_1, α_2 are nonnegative with $\alpha_1 + \alpha_2 > 0$ and $\alpha_3 > 0$, we study the problem of minimizing the following objective function:

$$\widetilde{J}(u, z) := \mathcal{R}\Big(\frac{\alpha_1}{2}\|u^{\boldsymbol{y}} - \widehat{u}\|^2_{L^2(V;I)} + \frac{\alpha_2}{2}\|u^{\boldsymbol{y}}(\cdot, T) - \widehat{u}(\cdot, T)\|^2_{L^2(D)}\Big) + \frac{\alpha_3}{2}\|z\|^2_{L^2(V';I)}\,, \quad (4.18)$$

subject to the control constraint

$$z \in \mathcal{Z}_{\mathrm{ad}} \quad (4.19)$$

and the heat equation over the time interval $I = [0, T]$

$$\begin{aligned} &\tfrac{\partial}{\partial t}u^{\boldsymbol{y}}(\boldsymbol{x}, t) - \nabla \cdot \big(a^{\boldsymbol{y}}(\boldsymbol{x}, t)\nabla u^{\boldsymbol{y}}(\boldsymbol{x}, t)\big) = z(\boldsymbol{x}, t), && \boldsymbol{x} \in D, && t \in I, \\ &u^{\boldsymbol{y}}(\boldsymbol{x}, t) = 0, && \boldsymbol{x} \in \partial D, && t \in I, \\ &u^{\boldsymbol{y}}(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}), && \boldsymbol{x} \in D, \end{aligned} \quad (4.20)$$

for all $\boldsymbol{y} \in U$. Here $z(\boldsymbol{x}, t)$ is the control and $u_0 \in L^2(D)$ denotes the initial heat distribution. We denote the input functions collectively by $f := (z, u_0)$. We have imposed homogeneous Dirichlet boundary conditions.
To ensure wellposedness of (4.1)-(4.3) we make the following assumtions:

(AP1) the feasible set of controls $\mathcal{Z}_{\mathrm{ad}} \subseteq \mathcal{Z} = L^2(V';)$ is nonempty, bounded, closed and convex.

(AP2) The sequence of parameters $\boldsymbol{y} = (y_j)_{j \geqslant 1}$ is i.i.d. uniformly in $[-\frac{1}{2}, \frac{1}{2}]$ for each $j \in \mathbb{N}$, i.e., $\boldsymbol{y}$ is distributed on U with probability measure μ, where $\mu(\mathrm{d}\boldsymbol{y}) = \bigotimes_{j \geqslant 1} \mathrm{d}y_j = \mathrm{d}\boldsymbol{y}$.

(AP3) The input uncertainty is described by the diffusion coefficient $a(\boldsymbol{y}, \boldsymbol{x})$ in (4.2), which is assumed to depend linearly on the parameters y_j, i.e., let

$$a^{\boldsymbol{y}}(\boldsymbol{x}, t) := a_0(\boldsymbol{x}, t) + \sum_{j \geqslant 1} y_j\, \psi_j(\boldsymbol{x}, t), \qquad \boldsymbol{x} \in D, \quad \boldsymbol{y} \in U, \quad t \in I, \quad (4.21)$$

be an uncertain (thermal) diffusion coefficient.

(AP4) For a.e. $t \in I$ we have $a_0(\cdot, t) \in L^\infty(D)$, $\psi_j(\cdot, t) \in L^\infty(D)$ for all $j \geqslant 1$, and that we have $(\sup_{t \in I} \|\psi_j(\cdot, t)\|_{L^\infty(D)})_{j \geqslant 1} \in \ell^1$;

(AP5) The mapping $t \mapsto a^{\boldsymbol{y}}(\boldsymbol{x}, t)$ is measurable on I;

(AP6) The uniform ellipticity assumption holds, i.e.,

$$0 < a_{\min} \leqslant a^{\boldsymbol{y}}(\boldsymbol{x}, t) \leqslant a_{\max} < \infty, \quad \boldsymbol{x} \in D, \quad \boldsymbol{y} \in U, \quad \text{a.e. } t \in I,$$

for some positive constants $a_{\min}$ and $a_{\max}$.

Time-varying diffusion coefficients occur e.g., in finance, cancer tomography. However, the presented setting clearly also includes time-constant diffusion coefficients, i.e., $a^{\boldsymbol{y}}(\boldsymbol{x}, t) = a^{\boldsymbol{y}}(\boldsymbol{x}) \, \forall t \in I$. By $\mathcal{R}$ in eq. (4.18) we denote a risk measure, which is a functional that maps a set of random variables into the extended real numbers. Specifically, $\mathcal{R}$ will in this section be either the expected value or the entropic risk measure.
We will first introduce a function space setting to describe the problem properly, including the definition of the $L^2(V; I)$ and $L^2(V'; I)$ norms. To this end, we define $V := H_0^1(D)$ and its (topological) dual space $V' := H^{-1}(D)$, and identify $L^2(D)$ with its own dual. Let $\langle \cdot, \cdot \rangle_{V',V}$ denotes the duality pairing between V' and V. The norm and inner product in V are defined as usual by

$$\|v\|_V := \|\nabla v\|_{L^2(D)}, \quad \langle v_1, v_2 \rangle_V := \langle \nabla v_1, \nabla v_2 \rangle_{L^2(D)}.$$

We shall make use of the Riesz operator $R_V \colon V \to V'$ defined by

$$\langle R_V v_1, v_2 \rangle_{V',V} = \langle v_1, v_2 \rangle_V \quad \forall \, v_1, v_2 \in V, \tag{4.22}$$

as well as its inverse $R_V^{-1} \colon V' \to V$ satisfying $R_V^{-1} w = v \Leftrightarrow w = R_V v \; \forall v \in V$, $w \in V'$. It follows from (4.22) that

$$\langle w, v \rangle_{V',V} = \langle R_V^{-1} w, v \rangle_V \quad \forall \, v \in V, w \in V'. \tag{4.23}$$

In turn we define the inner product in V' by

$$\langle w_1, w_2 \rangle_{V'} := \langle R_V^{-1} w_1, R_V^{-1} w_2 \rangle_V.$$

The norm induced by this inner product is equal to the usual dual norm: indeed we have

$$\|w\|_{V'} := \sup_{0 \neq v \in V} \frac{|\langle w, v \rangle_{V',V}|}{\|v\|_V} = \sup_{0 \neq v \in V} \frac{|\langle R_V^{-1} w, v \rangle_V|}{\|v\|_V} \leqslant \|R_V^{-1} w\|_V,$$

where we used (4.23) and Cauchy–Schwarz inequality; similar arguments yield

$$\|R_V^{-1} w\|_V^2 = \langle R_V^{-1} w, R_V^{-1} w \rangle_V = \langle w, R_V^{-1} w \rangle_{V',V} \leqslant \|w\|_{V'} \, \|R_V^{-1} w\|_V,$$

leading to $\|w\|_{V'} = \|R_V^{-1} w\|_V = \sqrt{\langle w, w \rangle_{V'}}$ as claimed.
We use analogous notations for inner products and duality pairings between function spaces on the space-time cylinder $D \times I$. The space $L^2(V; I)$ consists of all measurable functions $v : I \to V$ with finite norm

$$\|v\|_{L^2(V;I)} := \Big(\int_I \|v(\cdot, t)\|_V^2 \, \mathrm{d}t \Big)^{1/2}.$$

Note that $(L^2(V; I))' = L^2(V'; I)$, with the duality pairing given by

$$\langle w, v \rangle_{L^2(V';I),L^2(V;I)} = \int_I \langle w(\cdot, t), v(\cdot, t) \rangle_{V',V} \, \mathrm{d}t.$$

We extend the Riesz operator R_V to $R_V : L^2(V;I) \to L^2(V';I)$ so that

$$\begin{aligned}\langle v_1, v_2\rangle_{L^2(V;I)} &= \int_I \langle v_1(\cdot,t), v_2(\cdot,t)\rangle_V \,\mathrm{d}t = \int_I \big\langle R_V v_1(\cdot,t), v_2(\cdot,t)\big\rangle_{V',V} \,\mathrm{d}t \\ &= \big\langle R_V v_1, v_2\big\rangle_{L^2(V';I),L^2(V;I)} \quad \forall\, v_1, v_2 \in L^2(V;I),\end{aligned}$$

and we extend the inverse $R_V^{-1} : L^2(V';I) \to L^2(V;I)$ analogously.
We define the space of solutions $u^{\boldsymbol{y}}$ for $\boldsymbol{y} \in U$ by

$$\mathcal{X} := \Big\{ v \in L^2(V;I) : \tfrac{\partial}{\partial t} v \in L^2(V';I) \Big\},$$

which is the space of all functions v in $L^2(V;I)$ with (distributional) derivative $\frac{\partial}{\partial t} v$ in $L^2(V';I)$, and which is equipped with the (graph) norm

$$\|v\|_{\mathcal{X}} := \Big(\int_I \Big(\|v(\cdot,t)\|_V^2 + \|\tfrac{\partial}{\partial t} v(\cdot,t)\|_{V'}^2 \Big) \,\mathrm{d}t \Big)^{1/2} = \Big(\|v\|_{L^2(V;I)}^2 + \|\tfrac{\partial}{\partial t} v\|_{L^2(V';I)}^2 \Big)^{1/2}.$$

Finally, because there are two inputs in equation (4.20), namely $z \in L^2(V';I)$ and $u_0 \in L^2(D)$, it is convenient to define the product space $\mathcal{Y} := L^2(V;I) \times L^2(D)$, and define its dual space by $\mathcal{Y}' := L^2(V';I) \times L^2(D)$, with the norms

$$\begin{aligned}\|v\|_{\mathcal{Y}} &:= \Big(\int_I \|v_1(\cdot,t)\|_V^2 \,\mathrm{d}t + \|v_2\|_{L^2(D)}^2 \Big)^{1/2}, \\ \|w\|_{\mathcal{Y}'} &:= \Big(\int_I \|w_1(\cdot,t)\|_{V'}^2 \,\mathrm{d}t + \|w_2\|_{L^2(D)}^2 \Big)^{1/2}.\end{aligned}$$

In particular, we extend $\mathcal{X}$ to $\mathcal{Y}$ as follows. For all $v \in \mathcal{X}$ we interpret v as an element of $\mathcal{Y}$ as $v = (v(\boldsymbol{x},t), v(\boldsymbol{x},0))$. This gives $\mathcal{X} \subseteq \mathcal{Y}$. We further know from [47, Theorem 5.9.3] that $\mathcal{X} \hookrightarrow \mathcal{C}(L^2(D);I)$ and $\max_{t\in I} \|v(\cdot,t)\|_{L^2(D)} \leqslant C_1(\|v\|_{L^2(V;I)} + \|\frac{\partial}{\partial t} v\|_{L^2(V';I)}) \leqslant \sqrt{2}\, C_1 \|v\|_{\mathcal{X}}$ for $v \in \mathcal{X}$, where C_1 depends on T only. Hence we obtain for all $v \in \mathcal{X}$ that

$$\begin{aligned}&\|v\|_{\mathcal{Y}}^2 = \|v\|_{L^2(V;I)\times L^2(D)}^2 = \|v\|_{L^2(V;I)}^2 + \|v(\cdot,0)\|_{L^2(D)}^2 \\ &\leqslant \|v\|_{L^2(V;I)}^2 + \Big(\max_{t\in I} \|v(\cdot,t)\|_{L^2(D)} \Big)^2 \leqslant \|v\|_{\mathcal{X}}^2 + 2\, C_1^2 \|v\|_{\mathcal{X}}^2 = (1 + 2\, C_1^2) \|v\|_{\mathcal{X}}^2,\end{aligned}$$

and thus we get that $\mathcal{X}$ is continuously embedded into $\mathcal{Y}$, i.e., $\mathcal{X} \hookrightarrow \mathcal{Y}$.

4.2.1 Weak formulation

Based on these spaces, using integration by parts with respect to $\boldsymbol{x}$ we can write (4.20) as a variational problem as follows. Given the input functions $f = (z, u_0) \in \mathcal{Y}'$ and $\boldsymbol{y} \in U$, find a function $u^{\boldsymbol{y}} \in \mathcal{X}$ such that

$$b(\boldsymbol{y}; u^{\boldsymbol{y}}, v) \;=\; \langle f, v\rangle_{\mathcal{Y}',\mathcal{Y}} \quad \forall\, v = (v_1, v_2) \in \mathcal{Y}\,, \tag{4.24}$$

where for all $w \in \mathcal{X}$, $v = (v_1, v_2) \in \mathcal{Y}$ and $\boldsymbol{y} \in U$,

$$\begin{aligned}b(\boldsymbol{y}; w, v) &:= \langle B^{\boldsymbol{y}} w, v\rangle_{\mathcal{Y}',\mathcal{Y}} \\ &:= \underbrace{\int_I \big\langle \tfrac{\partial}{\partial t} w, v_1\big\rangle_{V',V} \,\mathrm{d}t + \int_I \int_D \big(a^{\boldsymbol{y}} \nabla w \cdot \nabla v_1\big) \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t}_{=:\langle B_1^{\boldsymbol{y}} w, v_1\rangle_{L^2(V';I),L^2(V;I)}} + \underbrace{\int_D w(\cdot,0)\, v_2 \,\mathrm{d}\boldsymbol{x}}_{=:\langle B_2^{\boldsymbol{y}} w, v_2\rangle_{L^2(D)}}\,,\end{aligned} \tag{4.25}$$

$$\langle f, v\rangle_{\mathcal{Y}',\mathcal{Y}} := \int_I \langle z, v_1\rangle_{V',V}\,\mathrm{d}t + \int_D u_0\, v_2\,\mathrm{d}\boldsymbol{x}\,,$$

with operators $B^{\boldsymbol{y}} : \mathcal{X} \to \mathcal{Y}'$, $B_1^{\boldsymbol{y}} : \mathcal{X} \to L^2(V';I)$, $B_2^{\boldsymbol{y}} : \mathcal{X} \to L^2(D)$, and $B^{\boldsymbol{y}}w = (B_1^{\boldsymbol{y}}w, B_2^{\boldsymbol{y}}w)$. For better readability we have omitted the parameter dependence $v = (v_1(\boldsymbol{x},t), v_2(\boldsymbol{x}))$, $f = (z(\boldsymbol{x},t), u_0(\boldsymbol{x}))$, $w = w(\boldsymbol{x},t)$ and $a^{\boldsymbol{y}} = a^{\boldsymbol{y}}(\boldsymbol{x},t)$. Note that a solution of (4.24) automatically satisfies $u^{\boldsymbol{y}}(\cdot,0) = u_0$, as can be seen by setting $v_1 = 0$ and allowing arbitrary v_2.

The parametric family of parabolic evolution operators $\{B^{\boldsymbol{y}},\, \boldsymbol{y} \in U\}$ associated with this bilinear form is a family of isomorphisms from $\mathcal{X}$ to $\mathcal{Y}'$, see, e.g., [36]. In [147] a shorter proof based on the characterization of the bounded invertibility of linear operators between Hilbert spaces is presented, together with precise bounds on the norms of the operator and its inverse: there exist constants $0 < \beta_1 \leqslant \beta_2 < \infty$ such that

$$\sup_{\boldsymbol{y}\in U} \|(B^{\boldsymbol{y}})^{-1}\|_{\mathcal{Y}'\to\mathcal{X}} \leqslant \frac{1}{\beta_1} \quad \text{and} \quad \sup_{\boldsymbol{y}\in U} \|B^{\boldsymbol{y}}\|_{\mathcal{X}\to\mathcal{Y}'} \leqslant \beta_2\,, \tag{4.26}$$

where $\beta_1 \geqslant \frac{\min\{a_{\min}a_{\max}^{-2}, a_{\min}\}}{\sqrt{2\max\{a_{\min}^{-2},1\}+\varrho^2}}$ and $\beta_2 \leqslant \sqrt{2\max\{1,a_{\max}^2\}+\varrho^2}$ with $\varrho := \sup_{w\in\mathcal{X}} \frac{\|w(0,\cdot)\|_{L^2(D)}}{\|w\|_{\mathcal{X}}}$, and hence for all $\boldsymbol{y} \in U$ we have the *a priori* estimate

$$\|u^{\boldsymbol{y}}\|_{\mathcal{X}} \leqslant \frac{\|f\|_{\mathcal{Y}'}}{\beta_1} = \frac{1}{\beta_1}\|(z,u_0)\|_{\mathcal{Y}'} = \frac{1}{\beta_1}\big(\|z\|^2_{L^2(V';I)} + \|u_0\|^2_{L^2(D)}\big)^{1/2}. \tag{4.27}$$

For our later derivation of the optimality conditions for the optimal control problem, it is helpful to write the weak form of the PDE (4.24) as an operator equation using (4.25):

$$B^{\boldsymbol{y}}u^{\boldsymbol{y}} \;=\; (B_1^{\boldsymbol{y}}u^{\boldsymbol{y}}, B_2^{\boldsymbol{y}}u^{\boldsymbol{y}}) \;=\; (z,u_0) \qquad \text{in } \mathcal{Y}'\,, \tag{4.28}$$

with $B_1^{\boldsymbol{y}} : \mathcal{X} \to L^2(V';I)$ and $B_2^{\boldsymbol{y}} : \mathcal{X} \to L^2(D)$ given by

$$B_1^{\boldsymbol{y}} = \Lambda_1 B^{\boldsymbol{y}} \quad \text{and} \quad B_2^{\boldsymbol{y}} = \Lambda_2 B^{\boldsymbol{y}}\,,$$

where $\Lambda_1 : \mathcal{Y}' \to L^2(V';I)$ and $\Lambda_2 : \mathcal{Y}' \to L^2(D)$ are the restriction operators defined, for any $v = (v_1, v_2) \in \mathcal{Y}'$, by

$$\Lambda_1(v_1,v_2) := v_1 \quad \text{and} \quad \Lambda_2(v_1,v_2) := v_2\,.$$

For the definition of a meaningful inverse of the operators $B_1^{\boldsymbol{y}}$ and $B_2^{\boldsymbol{y}}$, we first define the trivial extension operators $\Xi_1 : L^2(V';I) \to \mathcal{Y}'$ and $\Xi_2 : L^2(D) \to \mathcal{Y}'$, for any $v_1 \in L^2(V';I)$ and $v_2 \in L^2(D)$, by

$$\Xi_1 v_1 := (v_1, 0) \quad \text{and} \quad \Xi_2 v_2 := (0, v_2)\,.$$

We observe that $P_1 := \Xi_1\Lambda_1$ is an orthogonal projection on the $L^2(V';I)$-component in $\mathcal{Y}'$ and analogously $P_2 := \Xi_2\Lambda_2$ is an orthogonal projection on the $L^2(D)$-component in $\mathcal{Y}'$. This is verified as follows. For all $v, u \in \mathcal{Y}'$ it is true that

$$\langle(\mathcal{I}_{\mathcal{Y}'} - P_1)v, P_1u\rangle_{\mathcal{Y}'} = 0 \quad \text{and} \quad \langle(\mathcal{I}_{\mathcal{Y}'} - P_2)v, P_2u\rangle_{\mathcal{Y}'} = 0\,,$$

where $\mathcal{I}_{\mathcal{Y}'}$ denotes the identity operator on $\mathcal{Y}'$. We clearly have $\mathcal{I}_{\mathcal{Y}'} = P_1 + P_2$. Therefore we can write any element v in $\mathcal{Y}'$ as $v = P_1v + P_2v$ in $\mathcal{Y}'$, and by linearity of $(B^{\boldsymbol{y}})^{-1}$ we get

$$(B^{\boldsymbol{y}})^{-1}v = (B^{\boldsymbol{y}})^{-1}(P_1v + P_2v) = (B^{\boldsymbol{y}})^{-1}P_1v + (B^{\boldsymbol{y}})^{-1}P_2v\,.$$

A meaningful inverse of the operators $B_1^{\boldsymbol{y}} : \mathcal{X} \to L^2(V';I)$ and $B_2^{\boldsymbol{y}} : \mathcal{X} \to L^2(D)$ are then given by $(B_1^{\boldsymbol{y}})^\dagger : L^2(V';I) \to \mathcal{X}$ and $(B_2^{\boldsymbol{y}})^\dagger : L^2(D) \to \mathcal{X}$, defined as

$$(B_1^{\boldsymbol{y}})^\dagger := (B^{\boldsymbol{y}})^{-1}\Xi_1 \quad \text{and} \quad (B_2^{\boldsymbol{y}})^\dagger := (B^{\boldsymbol{y}})^{-1}\Xi_2\,. \tag{4.29}$$

We call the operator $(B_1^{\boldsymbol{y}})^\dagger$ the pseudoinverse of $B_1^{\boldsymbol{y}}$ and the operator $(B_2^{\boldsymbol{y}})^\dagger$ the pseudoinverse of $B_2^{\boldsymbol{y}}$. Clearly, the pseudoinverse operators are linear and bounded operators.

Lemma 4.2.1. *The pseudoinverse operators $(B_1^{\boldsymbol{y}})^\dagger$ and $(B_2^{\boldsymbol{y}})^\dagger$ defined by (4.29) satisfy*

$$\begin{aligned} \mathcal{I}_{L^2(V';I)} &= B_1^{\boldsymbol{y}}(B_1^{\boldsymbol{y}})^\dagger\,, \quad \mathcal{I}_{L^2(D)} = B_2^{\boldsymbol{y}}(B_2^{\boldsymbol{y}})^\dagger\,, \quad \text{and} \\ \mathcal{I}_{\mathcal{X}} &= (B_1^{\boldsymbol{y}})^\dagger B_1^{\boldsymbol{y}} + (B_2^{\boldsymbol{y}})^\dagger B_2^{\boldsymbol{y}}, \end{aligned} \tag{4.30}$$

which are the identity operators on $L^2(V';I)$, $L^2(D)$, and $\mathcal{X}$, respectively.

Proof. From the definition of various operators, we have

$$\begin{aligned} B_1^{\boldsymbol{y}}(B_1^{\boldsymbol{y}})^\dagger &= \Lambda_1 B^{\boldsymbol{y}}(B^{\boldsymbol{y}})^{-1}\Xi_1 = \Lambda_1\mathcal{I}_{\mathcal{Y}'}\Xi_1 = \Lambda_1\Xi_1 = \mathcal{I}_{L^2(V';I)}\,, \\ B_2^{\boldsymbol{y}}(B_2^{\boldsymbol{y}})^\dagger &= \Lambda_2 B^{\boldsymbol{y}}(B^{\boldsymbol{y}})^{-1}\Xi_2 = \Lambda_2\mathcal{I}_{\mathcal{Y}'}\Xi_2 = \Lambda_2\Xi_2 = \mathcal{I}_{L^2(D)}\,, \\ (B_1^{\boldsymbol{y}})^\dagger B_1^{\boldsymbol{y}} + (B_2^{\boldsymbol{y}})^\dagger B_2^{\boldsymbol{y}} &= (B^{\boldsymbol{y}})^{-1}\Xi_1\Lambda_1 B^{\boldsymbol{y}} + (B^{\boldsymbol{y}})^{-1}\Xi_2\Lambda_2 B^{\boldsymbol{y}} \\ &= (B^{\boldsymbol{y}})^{-1}(P_1+P_2)B^{\boldsymbol{y}} = (B^{\boldsymbol{y}})^{-1}\mathcal{I}_{\mathcal{Y}'}B^{\boldsymbol{y}} = \mathcal{I}_{\mathcal{X}}\,, \end{aligned}$$

as required. □

Lemma 4.2.2. *For $\boldsymbol{y} \in U$ and given $(z, u_0) \in \mathcal{Y}'$, the solution $u^{\boldsymbol{y}}$ of the operator equation (4.28) can be written as*

$$u^{\boldsymbol{y}} \;=\; (B^{\boldsymbol{y}})^{-1}(z,u_0) \;=\; (B_1^{\boldsymbol{y}})^\dagger z + (B_2^{\boldsymbol{y}})^\dagger u_0 \quad \textit{in } \mathcal{X}\,. \tag{4.31}$$

Proof. From (4.30) we have $u^{\boldsymbol{y}} = (B_1^{\boldsymbol{y}})^\dagger B_1^{\boldsymbol{y}} u^{\boldsymbol{y}} + (B_2^{\boldsymbol{y}})^\dagger B_2^{\boldsymbol{y}} u^{\boldsymbol{y}} = (B_1^{\boldsymbol{y}})^\dagger z + (B_2^{\boldsymbol{y}})^\dagger u_0$, as required. □

4.2.2 Dual problem

In the following we will need the dual operators $(B^{\boldsymbol{y}})'$, $(B_1^{\boldsymbol{y}})'$ and $(B_2^{\boldsymbol{y}})'$ of $B^{\boldsymbol{y}}$, $B_1^{\boldsymbol{y}}$ and $B_2^{\boldsymbol{y}}$, respectively, which are formally defined by

$$\begin{aligned} \langle w, (B^{\boldsymbol{y}})'v\rangle_{\mathcal{X},\mathcal{X}'} &:= \langle B^{\boldsymbol{y}}w, v\rangle_{\mathcal{Y}',\mathcal{Y}} \\ \langle w, (B_1^{\boldsymbol{y}})'v_1\rangle_{\mathcal{X},\mathcal{X}'} &:= \langle B_1^{\boldsymbol{y}}w, v_1\rangle_{L^2(V';I),L^2(V;I)} \\ \langle w, (B_2^{\boldsymbol{y}})'v_2\rangle_{\mathcal{X},\mathcal{X}'} &:= \langle B_2^{\boldsymbol{y}}w, v_2\rangle_{L^2(D)} \end{aligned}$$

for all $w \in \mathcal{X}$, $v = (v_1, v_2) \in \mathcal{Y}$ and $\boldsymbol{y} \in U$, with $(B^{\boldsymbol{y}})'v = (B_1^{\boldsymbol{y}})'v_1 + (B_2^{\boldsymbol{y}})'v_2$.
The dual problem to (4.24) (or equivalently (4.28)) is as follows. Given the input function $f_{\text{dual}} \in \mathcal{X}'$ and $\boldsymbol{y} \in U$, find a function $q^{\boldsymbol{y}} = (q_1^{\boldsymbol{y}}, q_2^{\boldsymbol{y}}) \in \mathcal{Y}$ such that

$$\langle w, (B^{\boldsymbol{y}})'q^{\boldsymbol{y}}\rangle_{\mathcal{X},\mathcal{X}'} = \langle w, f_{\text{dual}}\rangle_{\mathcal{X},\mathcal{X}'} \quad \forall w \in \mathcal{X}, \tag{4.32}$$

or in operator form $(B^{\boldsymbol{y}})'q^{\boldsymbol{y}} = f_{\text{dual}}$, which has the unique solution $q^{\boldsymbol{y}} = \big((B^{\boldsymbol{y}})'\big)^{-1} f_{\text{dual}}$. Existence and uniqueness of the solution of the dual problem follow directly from the bounded invertibility of $B^{\boldsymbol{y}}$. We know that its inverse, $(B^{\boldsymbol{y}})^{-1}$, is a bounded linear operator and thus the dual of $(B^{\boldsymbol{y}})^{-1}$ is (uniquely) defined (see, e.g., [164, Theorem 1 and Definition 1, Chapter VII]). The operator $(B^{\boldsymbol{y}})^{-1}$ and its dual operator $((B^{\boldsymbol{y}})^{-1})' = ((B^{\boldsymbol{y}})')^{-1}$

are equal in their operator norms (see, e.g., [164, Theorem 2, Chapter VII]), i.e., the operator norms of the dual operator $(B^{\boldsymbol{y}})'$ and its inverse are bounded by the constants β_2 and $\frac{1}{\beta_1}$ in (4.26).
Applying integration by parts with respect to the time variable in (4.25), the left-hand side of the dual problem (4.32) can be written as

$$\begin{aligned}
&\langle w, (B^{\boldsymbol{y}})' q^{\boldsymbol{y}} \rangle_{\mathcal{X},\mathcal{X}'} = \langle B^{\boldsymbol{y}} w, q^{\boldsymbol{y}} \rangle_{\mathcal{Y}',\mathcal{Y}} \\
&= \bigg(\int_I \langle w, -\tfrac{\partial}{\partial t} q_1^{\boldsymbol{y}} \rangle_{V,V'} \,\mathrm{d}t + \int_I \int_D (a^{\boldsymbol{y}} \nabla w \cdot \nabla q_1^{\boldsymbol{y}}) \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \\
&\quad + \int_D w(\cdot,T)\, q_1^{\boldsymbol{y}}(\cdot,T) \,\mathrm{d}\boldsymbol{x} - \int_D w(\cdot,0)\, q_1^{\boldsymbol{y}}(\cdot,0) \,\mathrm{d}\boldsymbol{x} \bigg) + \int_D w(\cdot,0)\, q_2^{\boldsymbol{y}} \,\mathrm{d}\boldsymbol{x} \qquad (4.33) \\
&= \big\langle w, (B_1^{\boldsymbol{y}})' q_1^{\boldsymbol{y}} \big\rangle_{\mathcal{X},\mathcal{X}'} + \big\langle w, (B_2^{\boldsymbol{y}})' q_2^{\boldsymbol{y}} \big\rangle_{\mathcal{X},\mathcal{X}'} .
\end{aligned}$$

We may express the solution $q^{\boldsymbol{y}} = (q_1^{\boldsymbol{y}}, q_2^{\boldsymbol{y}}) \in \mathcal{Y}$ of the dual problem (4.32) in terms of the dual operators of the pseudoinverse operators $(B_1^{\boldsymbol{y}})^\dagger$ and $(B_2^{\boldsymbol{y}})^\dagger$. This is true because we get an analogous result to Lemma 4.2.1 in the dual spaces.

Lemma 4.2.3. *The dual operators $((B_1^{\boldsymbol{y}})^\dagger)'$ and $((B_2^{\boldsymbol{y}})^\dagger)'$ of the pseudoinverse operators defined in* (4.29) *satisfy*

$$\begin{aligned}
\mathcal{I}_{L^2(V;I)} &= ((B_1^{\boldsymbol{y}})^\dagger)'(B_1^{\boldsymbol{y}})', \quad \mathcal{I}_{L^2(D)} = ((B_2^{\boldsymbol{y}})^\dagger)'(B_2^{\boldsymbol{y}})', \quad \text{and} \\
\mathcal{I}_{\mathcal{X}'} &= (B_1^{\boldsymbol{y}})'((B_1^{\boldsymbol{y}})^\dagger)' + (B_2^{\boldsymbol{y}})'((B_2^{\boldsymbol{y}})^\dagger)', \qquad (4.34)
\end{aligned}$$

which are the identity operators on $L^2(V;I)$, $L^2(D)$ and $\mathcal{X}'$, respectively.

Proof. For all $v_1 \in L^2(V';I)$, $w_1 \in L^2(V;I)$, $v_2, w_2 \in L^2(D)$, it follows from (4.30) that

$$\begin{aligned}
\langle v_1, w_1 \rangle_{L^2(V';I),L^2(V;I)} &= \big\langle B_1^{\boldsymbol{y}} (B_1^{\boldsymbol{y}})^\dagger v_1, w_1 \big\rangle_{L^2(V';I),L^2(V;I)} \\
&= \big\langle v_1, ((B_1^{\boldsymbol{y}})^\dagger)'(B_1^{\boldsymbol{y}})' w_1 \big\rangle_{L^2(V';I),L^2(V;I)} , \text{ and} \\
\langle v_2, w_2 \rangle_{L^2(D)} &= \big\langle B_2^{\boldsymbol{y}} (B_2^{\boldsymbol{y}})^\dagger v_2, w_2 \big\rangle_{L^2(D)} = \big\langle v_2, ((B_2^{\boldsymbol{y}})^\dagger)'(B_2^{\boldsymbol{y}})' w_2 \big\rangle_{L^2(D)} .
\end{aligned}$$

Similarly, for all $v \in \mathcal{X}$ and $w \in \mathcal{X}'$ we have

$$\begin{aligned}
\langle v, w \rangle_{\mathcal{X},\mathcal{X}'} &= \big\langle \big((B_1^{\boldsymbol{y}})^\dagger B_1^{\boldsymbol{y}} + (B_2^{\boldsymbol{y}})^\dagger B_2^{\boldsymbol{y}}\big) v, w \big\rangle_{\mathcal{X},\mathcal{X}'} \\
&= \big\langle (B_1^{\boldsymbol{y}})^\dagger B_1^{\boldsymbol{y}} v, w \big\rangle_{\mathcal{X},\mathcal{X}'} + \big\langle (B_2^{\boldsymbol{y}})^\dagger B_2^{\boldsymbol{y}} v, w \big\rangle_{\mathcal{X},\mathcal{X}'} \\
&= \big\langle v, (B_1^{\boldsymbol{y}})'((B_1^{\boldsymbol{y}})^\dagger)' w \big\rangle_{\mathcal{X},\mathcal{X}'} + \big\langle v, (B_2^{\boldsymbol{y}})'((B_2^{\boldsymbol{y}})^\dagger)' w \big\rangle_{\mathcal{X},\mathcal{X}'} \\
&= \langle v, \big((B_1^{\boldsymbol{y}})'((B_1^{\boldsymbol{y}})^\dagger)' + (B_2^{\boldsymbol{y}})'((B_2^{\boldsymbol{y}})^\dagger)'\big) w \rangle_{\mathcal{X},\mathcal{X}'} .
\end{aligned}$$

This completes the proof. □

Lemma 4.2.4. *Given the input function $f_{\mathrm{dual}} \in \mathcal{X}'$ and $\boldsymbol{y} \in U$, the (unique) solution of the dual problem* (4.32) *is given by*

$$q^{\boldsymbol{y}} = (q_1^{\boldsymbol{y}}, q_2^{\boldsymbol{y}}) = \big(((B_1^{\boldsymbol{y}})^\dagger)' f_{\mathrm{dual}}, ((B_2^{\boldsymbol{y}})^\dagger)' f_{\mathrm{dual}}\big) \quad \text{in } \mathcal{Y} . \qquad (4.35)$$

Proof. Existence and uniqueness follow from the bounded invertibility of $(B^{\boldsymbol{y}})'$, see Section 4.2.1. Thus, we only need to verify that (4.35) solves the dual problem (4.32). It follows from (4.34) that

$$f_{\mathrm{dual}} = \big((B_1^{\boldsymbol{y}})'((B_1^{\boldsymbol{y}})^\dagger)' + (B_2^{\boldsymbol{y}})'((B_2^{\boldsymbol{y}})^\dagger)'\big) f_{\mathrm{dual}}$$

$$
\begin{aligned}
&= (B_1^{\boldsymbol{y}})'((B_1^{\boldsymbol{y}})^\dagger)' f_{\mathrm{dual}} + (B_2^{\boldsymbol{y}})'((B_2^{\boldsymbol{y}})^\dagger)' f_{\mathrm{dual}} \\
&= (B_1^{\boldsymbol{y}})' q_1^{\boldsymbol{y}} + (B_2^{\boldsymbol{y}})' q_2^{\boldsymbol{y}} = (B^{\boldsymbol{y}})' q^{\boldsymbol{y}} \,,
\end{aligned}
$$

as required. □

We will see in the next section that, with the correct choice of the right-hand side f_{dual}, the Fréchet derivative of the objective function (4.18) can be computed using the solution $q^{\boldsymbol{y}}$ of the dual problem.

4.2.3 Reduced problem

We want to analyze the problem in its reduced form, i.e., expressing the state $u^{\boldsymbol{y}} = (B^{\boldsymbol{y}})^{-1}(z, u_0)$ in (4.18) in terms of the control z. This reformulation is possible because of the bounded invertibility of the operator $B^{\boldsymbol{y}}$ for every $\boldsymbol{y} \in U$, see Section 4.2.1 and the references therein. We therefore introduce an alternative notation $u(z) = (u^{\boldsymbol{y}}(z))(\boldsymbol{x}, t) = u^{\boldsymbol{y}}(\boldsymbol{x}, t)$. Clearly, $u^{\boldsymbol{y}}$ depends also on u_0, which is assumed to be given. The reduced problem is then to minimize

$$
\begin{aligned}
J(z) := \widetilde{J}\big(u(z), z\big) = \mathcal{R}\Big(\frac{\alpha_1}{2} \big\| u^{\boldsymbol{y}}(z) - \widehat{u} \big\|^2_{L^2(V;I)} + \frac{\alpha_2}{2} \big\| E_T\big(u^{\boldsymbol{y}}(z) - \widehat{u}\big) \big\|^2_{L^2(D)} \Big) \\
+ \frac{\alpha_3}{2} \|z\|^2_{L^2(V';I)},
\end{aligned} \tag{4.36}
$$

where $E_T \colon \mathcal{X} \to L^2(D)$ is the bounded linear operator (see, e.e., [47, Chapter 5, Theorem 3]) defined by $v \mapsto v(\cdot, T)$ for some fixed terminal time $T > 0$.
Defining

$$
\Phi^{\boldsymbol{y}}(z) := \frac{\alpha_1}{2} \big\| (B^{\boldsymbol{y}})^{-1}(z, u_0) - \widehat{u} \big\|^2_{L^2(V;I)} + \frac{\alpha_2}{2} \big\| E_T\big((B^{\boldsymbol{y}})^{-1}(z, u_0) - \widehat{u}\big) \big\|^2_{L^2(D)}, \tag{4.37}
$$

we can equivalently write the reduced problem as

$$
\min_{z \in \mathcal{Z}_{\mathrm{ad}}} \Big(\mathcal{R}(\Phi^{\boldsymbol{y}}(z)) + \frac{\alpha_3}{2} \|z\|^2_{L^2(V';I)} \Big) \,. \tag{4.38}
$$

With the uniformly boundedly invertible forward operator $B^{\boldsymbol{y}}$, our setting fits into the framework of theorem 3.5.4. In particular, the forward operator $B^{\boldsymbol{y}}$, the regularization term $\frac{\alpha_3}{2}\|z\|^2_{L^2(V';I)}$ and the random variable tracking-type objective function $\Phi^{\boldsymbol{y}}$ satisfy assumption 3.5.1, assumption 3.5.3 and assumption 3.3.1. We obtain the following result.

Corollary 4.2.5. *Let $\alpha_1, \alpha_2 \geqslant 0$ and $\alpha_3 > 0$ with $\alpha_1 + \alpha_2 > 0$ and let $\mathcal{R}$ be proper, closed, convex and monotonic, then there exists a unique solution of* (4.38).

Proof. The existence of the solution follows from theorem 3.5.4. We thus only prove the strong convexity of the objective function, which implies strict convexity and hence uniqueness of the solution. Clearly $\frac{\alpha_3}{2}\|z\|^2_{L^2(V';I)}$ is strongly convex. Since the sum of a convex and a strongly convex function is strongly convex it remains to show the convexity of $\mathcal{R}(\Phi^{\boldsymbol{y}}(z))$. By the linearity and the bounded invertibility of the linear forward operator $B^{\boldsymbol{y}}$, the tracking-type objective functional $\Phi^{\boldsymbol{y}}(z)$ is quadratic in z and hence convex, i.e., we have for $z, \tilde{z} \in L^2(V'; I)$ and $\lambda \in [0, 1]$ that $\Phi^{\boldsymbol{y}}(\lambda z + (1-\lambda)\tilde{z}) \leqslant \lambda \Phi^{\boldsymbol{y}}(z) + (1-\lambda)\Phi^{\boldsymbol{y}}(\tilde{z})$. Then, by lemma 3.2.1 we obtain that $\mathcal{R}(\Phi^{\boldsymbol{y}}(z))$ is convex. □

Having ensured the existence of a unique optimal control $z^* \in \mathcal{Z}_{\mathrm{ad}}$, we derive necessary and sufficient optimality conditions. To this end, we compute derivatives of the reduced objective function eq. (4.36).

4.2.4 Derivatives for linear risk measures, including the expected value

First we derive a formula for the Fréchet derivative of (4.36) when $\mathcal{R}$ is linear, which includes the special case $\mathcal{R}(\cdot) = \int_U (\cdot)\,\mathrm{d}\boldsymbol{y}$.

Lemma 4.2.6. *Let $\mathcal{R}$ be linear. Then the Fréchet derivative of* (4.36) *as an element of $(L^2(V';I))' = L^2(V;I)$ is given by*

$$J'(z) \;=\; \mathcal{R}\Big(\big((B_1^{\boldsymbol{y}})^\dagger\big)'\big(\alpha_1 R_V + \alpha_2 E_T' E_T\big)\big(u^{\boldsymbol{y}}(z) - \widehat{u}\big)\Big) + \alpha_3 R_V^{-1} z \tag{4.39}$$

for $z \in L^2(V';I)$.

Proof. For $z, \delta \in L^2(V';I)$, we can write

$$\begin{aligned}
J(z+\delta) &= \mathcal{R}\Big(\frac{\alpha_1}{2}\big\|u^{\boldsymbol{y}}(z+\delta) - u^{\boldsymbol{y}}(z) + u^{\boldsymbol{y}}(z) - \widehat{u}\big\|^2_{L^2(V;I)} \\
&\qquad + \frac{\alpha_2}{2}\big\|E_T\big(u^{\boldsymbol{y}}(z+\delta) - u^{\boldsymbol{y}}(z) + u^{\boldsymbol{y}}(z) - \widehat{u}\big)\big\|^2_{L^2(D)}\Big) + \frac{\alpha_3}{2}\|z+\delta\|^2_{L^2(V';I)} \\
&= \mathcal{R}\Big(\frac{\alpha_1}{2}\big\|(B_1^{\boldsymbol{y}})^\dagger\delta + \big(u^{\boldsymbol{y}}(z) - \widehat{u}\big)\big\|^2_{L^2(V;I)} \\
&\qquad + \frac{\alpha_2}{2}\big\|E_T(B_1^{\boldsymbol{y}})^\dagger\delta + E_T\big(u^{\boldsymbol{y}}(z) - \widehat{u}\big)\big\|^2_{L^2(D)}\Big) + \frac{\alpha_3}{2}\|z+\delta\|^2_{L^2(V';I)},
\end{aligned}$$

where we used (4.31) to write $u^{\boldsymbol{y}}(z+\delta) - u^{\boldsymbol{y}}(z) = [(B_1^{\boldsymbol{y}})^\dagger(z+\delta) + (B_2^{\boldsymbol{y}})^\dagger u_0] - [(B_1^{\boldsymbol{y}})^\dagger(z) + (B_2^{\boldsymbol{y}})^\dagger u_0] = (B_1^{\boldsymbol{y}})^\dagger\delta$. Expanding the squared norms using $\|v+w\|^2 = \langle v+w, v+w\rangle = \|v\|^2 + 2\langle v, w\rangle + \|w\|^2$, we obtain

$$J(z+\delta) \;=\; J(z) + \big(\partial_z J(z)\big)\,\delta + \mathrm{o}(\delta),$$

with the Fréchet derivative $\partial_z J(z) : L^2(V';I) \to \mathbb{R}$ defined by

$$\begin{aligned}
\big(\partial_z J(z)\big)\,\delta := \mathcal{R}\Big(\alpha_1 \overbrace{\big\langle (B_1^{\boldsymbol{y}})^\dagger\delta, u^{\boldsymbol{y}}(z) - \widehat{u}\big\rangle_{L^2(V;I)}}^{=:\,\mathrm{Term}_1} \\
+ \alpha_2 \underbrace{\big\langle E_T(B_1^{\boldsymbol{y}})^\dagger\delta, E_T\big(u^{\boldsymbol{y}}(z) - \widehat{u}\big)\big\rangle_{L^2(D)}}_{=:\,\mathrm{Term}_2}\Big) + \alpha_3 \underbrace{\langle z, \delta\rangle_{L^2(V';I)}}_{=:\,\mathrm{Term}_3}.
\end{aligned}$$

It remains to simplify the three terms. Using the extended Riesz operator $R_V : L^2(V;I) \to L^2(V';I)$, we have

$$\begin{aligned}
\mathrm{Term}_1 &= \big\langle u^{\boldsymbol{y}}(z) - \widehat{u}, (B_1^{\boldsymbol{y}})^\dagger\delta\big\rangle_{L^2(V;I)} = \big\langle R_V\big(u^{\boldsymbol{y}}(z) - \widehat{u}\big), (B_1^{\boldsymbol{y}})^\dagger\delta\big\rangle_{L^2(V';I),L^2(V;I)} \\
&= \big\langle R_V\big(u^{\boldsymbol{y}}(z) - \widehat{u}\big), (B_1^{\boldsymbol{y}})^\dagger\delta\big\rangle_{\mathcal{X}',\mathcal{X}} = \big\langle \big((B_1^{\boldsymbol{y}})^\dagger\big)' R_V\big(u^{\boldsymbol{y}}(z) - \widehat{u}\big), \delta\big\rangle_{L^2(V;I),L^2(V';I)},
\end{aligned}$$

where the third equality follows since $(B_1^{\boldsymbol{y}})^\dagger\delta \in \mathcal{X} \hookrightarrow L^2(V;I)$, and the fourth equality follows from the definition of the dual operator $((B_1^{\boldsymbol{y}})^\dagger)' : \mathcal{X}' \to L^2(V;I)$, noting that $(L^2(V';I))' = L^2(V;I)$.
Next, using the definition of the dual operator $(E_T)' : L^2(D) \to \mathcal{X}'$, we can write

$$\begin{aligned}
\mathrm{Term}_2 &= \big\langle E_T\big(u^{\boldsymbol{y}}(z) - \widehat{u}\big), E_T(B_1^{\boldsymbol{y}})^\dagger\delta\big\rangle_{L^2(D)} = \big\langle E_T' E_T\big(u^{\boldsymbol{y}}(z) - \widehat{u}\big), (B_1^{\boldsymbol{y}})^\dagger\delta\big\rangle_{\mathcal{X}',\mathcal{X}} \\
&= \big\langle \big((B_1^{\boldsymbol{y}})^\dagger\big)' E_T' E_T\big(u^{\boldsymbol{y}}(z) - \widehat{u}\big), \delta\big\rangle_{L^2(V;I),L^2(V';I)}.
\end{aligned}$$

Finally, using the definition of the $L^2(V', I)$ inner product and the extended inverse Riesz operator $R_V^{-1}\colon L^2(V'; I) \to L^2(V; I)$, we obtain

$$\text{Term}_3 \;=\; \langle z, \delta\rangle_{L^2(V';I)} \;=\; \langle R_V^{-1} z, R_V^{-1}\delta\rangle_{L^2(V;I)} \;=\; \big\langle R_V^{-1} z, \delta\big\rangle_{L^2(V;I),L^2(V';I)}.$$

Writing $(\partial_z J(z))\,\delta = \langle J'(z), \delta\rangle_{L^2(V;I),L^2(V';I)}$ and collecting the terms above leads to the expression for $J'(z)$ in (4.39). □

Next, we show that the Fréchet derivative $J'(z)$ of $J(z)$ can be computed using the solution of the dual problem (4.32) with

$$f_{\text{dual}} := (\alpha_1 R_V + \alpha_2 E_T' E_T)(u^{\boldsymbol{y}} - \widehat{u}) \in \mathcal{X}'\,. \tag{4.40}$$

We show this first for the special case when $\mathcal{R}$ is linear.

Lemma 4.2.7. *Let $\alpha_1, \alpha_2 \geqslant 0$ and $\alpha_3 > 0$, with $\alpha_1 + \alpha_2 > 0$. Let $f = (z, u_0) \in \mathcal{Y}'$ and $\widehat{u} \in \mathcal{X}$. For every $\boldsymbol{y} \in U$, let $u^{\boldsymbol{y}} \in \mathcal{X}$ be the solution of (4.20) and then let $q^{\boldsymbol{y}} \in \mathcal{Y}$ be the solution of (4.32) with f_{dual} given by (4.40). Then for $\mathcal{R}$ linear, the Fréchet derivative of (4.36) is given as an element of $L^2(V; I)$ by*

$$J'(z) = \mathcal{R}(q_1) + \alpha_3 R_V^{-1} z \tag{4.41}$$

for $z \in L^2(V'; I)$.

Proof. This follows immediately from (4.40), Lemma 4.2.6 and Lemma 4.2.4. □

Proposition 4.2.8. *Under the conditions of Lemma 4.2.7, with f_{dual} given by (4.40), the dual solution $q^{\boldsymbol{y}} = (q_1^{\boldsymbol{y}}, q_2^{\boldsymbol{y}}) \in \mathcal{Y}$ satisfies*

$$q_2^{\boldsymbol{y}} = q_1^{\boldsymbol{y}}(\cdot, 0).$$

Consequently, the left-hand side of (4.32) reduces to

$$\int_I \big\langle w, -\tfrac{\partial}{\partial t} q_1^{\boldsymbol{y}}\big\rangle_{V,V'} \,\mathrm{d}t + \int_I \int_D \big(a^{\boldsymbol{y}} \nabla w \cdot \nabla q_1^{\boldsymbol{y}}\big)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t + \int_D w(\cdot, T)\, q_1^{\boldsymbol{y}}(\cdot, T)\,\mathrm{d}\boldsymbol{x}\,, \tag{4.42}$$

and hence $q_1^{\boldsymbol{y}}$ is the solution to

$$\begin{cases} -\frac{\partial}{\partial t} q_1^{\boldsymbol{y}}(\boldsymbol{x}, t) - \nabla \cdot \big(a^{\boldsymbol{y}}(\boldsymbol{x}, t) \nabla q_1^{\boldsymbol{y}}(\boldsymbol{x}, t)\big) = \alpha_1 R_V\big(u^{\boldsymbol{y}}(\boldsymbol{x}, t) - \widehat{u}(\boldsymbol{x}, t)\big) \\ q_1^{\boldsymbol{y}}(\boldsymbol{x}, t) = 0 \\ q_1^{\boldsymbol{y}}(\boldsymbol{x}, T) = \alpha_2\big(u^{\boldsymbol{y}}(\boldsymbol{x}, T) - \widehat{u}(\boldsymbol{x}, T)\big), \end{cases} \tag{4.43}$$

where the first equation holds for $\boldsymbol{x} \in D$, $t \in I$, and the second equation holds for $\boldsymbol{x} \in \partial D$, $t \in I$, and the last equation holds for $x \in D$.

Proof. Since (4.32) holds for arbitrary $w \in \mathcal{X}$, it holds in particular for the special case

$$w = w_n(\boldsymbol{x}, t) := \begin{cases} \big(1 - \frac{nt}{T}\big)\, v(\boldsymbol{x}) & \text{for } t \in \big[0, \frac{T}{n}\big]\,, \\ 0 & \text{for } t \in \big(\frac{T}{n}, T\big]\,, \end{cases}$$

with arbitrary $v \in V$. For f_{dual} given by (4.40), the right-hand side of (4.32) becomes

$$\langle w_n, f_{\text{dual}}\rangle_{\mathcal{X},\mathcal{X}'}$$

$$= \big\langle w_n, \alpha_1 R_V(u^{\boldsymbol{y}} - \widehat{u})\big\rangle_{\mathcal{X},\mathcal{X}'} + \big\langle w_n(\cdot,T), \alpha_2\big(u^{\boldsymbol{y}}(\cdot,T) - \widehat{u}(\cdot,T)\big)\big\rangle_{L^2(D)}$$
$$= \int_0^{\frac{T}{n}} \int_D \big(1 - \tfrac{nt}{T}\big)\, v\, \alpha_1 R_V(u^{\boldsymbol{y}} - \widehat{u})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \;\; \to 0 \;\text{ as }\; n \to \infty\,.$$

From (4.33) the left-hand side of (4.32) is now

$$\begin{aligned}
&\langle w_n, (B^{\boldsymbol{y}})' q^{\boldsymbol{y}}\rangle_{\mathcal{X},\mathcal{X}'} \\
&= \int_0^{\frac{T}{n}} \big(1 - \tfrac{nt}{T}\big)\big\langle v, -\tfrac{\partial}{\partial t} q_1^{\boldsymbol{y}}\big\rangle_{V,V'}\,\mathrm{d}t + \int_0^{\frac{T}{n}} \int_D \big(1 - \tfrac{nt}{T}\big)\big(a^{\boldsymbol{y}} \nabla v \cdot \nabla q_1^{\boldsymbol{y}}\big)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \\
&\qquad - \int_D v\, q_1^{\boldsymbol{y}}(\cdot,0)\,\mathrm{d}\boldsymbol{x} + \int_D v\, q_2^{\boldsymbol{y}}\,\mathrm{d}\boldsymbol{x} \\
&\to \int_D v\,\big(q_2^{\boldsymbol{y}} - q_1^{\boldsymbol{y}}(\cdot,0)\big)\,\mathrm{d}\boldsymbol{x} \quad \text{as} \quad n \to \infty\,.
\end{aligned}$$

Equating the two sides, letting $n \to \infty$, and noting that $v \in V$ is arbitrary, we conclude that necessarily $q_2^{\boldsymbol{y}} = q_1^{\boldsymbol{y}}(\cdot,0)$.
Hence, the left-hand side of (4.32) reduces to (4.42). By analogy with the weak form of (4.20), using the transformation $t \mapsto T - t$, we conclude that $q_1^{\boldsymbol{y}}$ is the solution to (4.43). □

4.2.5 Derivatives of the entropic risk measure

The expected value is risk neutral. Next, we consider the risk averse entropic risk measure (see section 4.2.5):

$$\mathcal{R}_{\mathrm{e}}(Y(\boldsymbol{y})) := \frac{1}{\theta} \ln\Big(\int_U \exp\big(\theta\, Y(\boldsymbol{y})\big)\,\mathrm{d}\boldsymbol{y}\Big)\,,$$

for an essentially bounded random variable $Y(\boldsymbol{y})$ and some $\theta \in (0,\infty)$. Using $\mathcal{R} = \mathcal{R}_{\mathrm{e}}$ in (4.36), the optimal control problem becomes $\min_{z\in\mathcal{Z}} J(z)$, with

$$J(z) = \frac{1}{\theta} \ln\Big(\int_U \exp\big(\theta\, \Phi^{\boldsymbol{y}}(z)\big)\,\mathrm{d}\boldsymbol{y}\Big) + \frac{\alpha_3}{2}\|z\|^2_{L^2(V';I)}\,, \tag{4.44}$$

for some $\theta \in (0,\infty)$ and $\Phi^{\boldsymbol{y}}$ defined in (4.37).
In the following we want to compute the Fréchet derivative of $J(z)$ with respect to $z \in L^2(V';I)$. To this end, we verify that $\Phi^{\boldsymbol{y}}(z) \leqslant C < \infty$ is uniformly bounded in $\boldsymbol{y} \in U$ for any $z \in L^2(V';I)$, i.e. the constant $C > 0$ is independent of $\boldsymbol{y} \in U$.

Lemma 4.2.9. *Let $f = (z, u_0) \in \mathcal{Y}'$ and $\widehat{u} \in \mathcal{X}$, and let $\alpha_1, \alpha_2 \geqslant 0$ with $\alpha_1 + \alpha_2 > 0$. Then for all $\boldsymbol{y} \in U$, the function $\Phi^{\boldsymbol{y}}$ defined by (4.37) satisfies*

$$0 \leqslant \Phi^{\boldsymbol{y}} \leqslant \frac{\alpha_1 + \alpha_2\,\|E_T\|^2_{\mathcal{X}\to L^2(D)}}{2}\Big(\frac{\|f\|_{\mathcal{Y}'}}{\beta_1} + \|\widehat{u}\|_{\mathcal{X}}\Big)^2 < \infty. \tag{4.45}$$

Thus for all $\theta > 0$ we have

$$1 \leqslant \exp\big(\theta\,\Phi^{\boldsymbol{y}}\big) \leqslant e^{\sigma} < \infty, \quad \textit{with} \tag{4.46}$$
$$\sigma := \frac{\alpha_1 + \alpha_2\,\|E_T\|^2_{\mathcal{X}\to L^2(D)}}{2}\Big(\frac{\|f\|_{\mathcal{Y}'}}{\beta_1} + \|\widehat{u}\|_{\mathcal{X}}\Big)^2\theta. \tag{4.47}$$

Proof. We have from (4.37) that

$$\begin{aligned}\Phi^{\boldsymbol{y}}(z) &\leqslant \frac{\alpha_1}{2}\big\|(B^{\boldsymbol{y}})^{-1}f-\widehat{u}\big\|_{\mathcal{X}}^2+\frac{\alpha_2}{2}\|E_T\|_{\mathcal{X}\to L^2(D)}^2\big\|(B^{\boldsymbol{y}})^{-1}f-\widehat{u}\big\|_{\mathcal{X}}^2\\ &\leqslant \frac{\alpha_1+\alpha_2\,\|E_T\|_{\mathcal{X}\to L^2(D)}^2}{2}\left(\big\|(B^{\boldsymbol{y}})^{-1}f\big\|_{\mathcal{X}}+\|\widehat{u}\|_{\mathcal{X}}\right)^2,\end{aligned}$$

which yields (4.45) after applying (4.27). □

Using the preceding lemma, we compute the Fréchet derivative of (4.44).

Lemma 4.2.10. *Let $\alpha_1,\alpha_2\geqslant 0$ and $\alpha_3>0$, with $\alpha_1+\alpha_2>0$, and let $0<\theta<\infty$. Let $f=(z,u_0)\in\mathcal{Y}'$ and $\widehat{u}\in\mathcal{X}$. For every $\boldsymbol{y}\in U$, let $u^{\boldsymbol{y}}\in\mathcal{X}$ be the solution of (4.20) and then let $q^{\boldsymbol{y}}=(q_1^{\boldsymbol{y}},q_2^{\boldsymbol{y}})\in\mathcal{Y}$ be the solution of (4.32) with f_{dual} given by (4.40). Then the Fréchet derivative of (4.44) is given as an element of $L^2(V;I)$ for $z\in L^2(V';I)$ by*

$$J'(z)=\frac{1}{\int_U\exp\big(\theta\,\Phi^{\boldsymbol{y}}(z)\big)\,\mathrm{d}\boldsymbol{y}}\int_U\exp\big(\theta\,\Phi^{\boldsymbol{y}}(z)\big)\,q_1^{\boldsymbol{y}}\,\mathrm{d}\boldsymbol{y}+\alpha_3R_V^{-1}z \tag{4.48}$$

where $\Phi^{\boldsymbol{y}}(z)$ is defined in (4.37).

Proof. The application of the chain rule gives

$$\partial_z\mathcal{R}_{\mathrm{e}}(\Phi^{\boldsymbol{y}}(z))=\frac{1}{\theta\int_U\exp\big(\theta\,\Phi^{\boldsymbol{y}}(z)\big)\,\mathrm{d}\boldsymbol{y}}\partial_z\Big(\int_U\exp\big(\theta\,\Phi^{\boldsymbol{y}}(z)\big)\,\mathrm{d}\boldsymbol{y}\Big).$$

Lemma 4.2.9 implies that $1\leqslant\int_U\exp\big(\theta\,\Phi^{\boldsymbol{y}}(z)\big)\,\mathrm{d}\boldsymbol{y}<\infty$. Then the integral is a bounded and linear operator and hence its Fréchet derivative is the operator itself. Exploiting this fact, we obtain that $\partial_z\big(\int_U\exp\big(\theta\,\Phi^{\boldsymbol{y}}(z)\big)\,\mathrm{d}\boldsymbol{y}\big)=\int_U\big(\partial_z\exp\big(\theta\,\Phi^{\boldsymbol{y}}(z)\big)\big)\,\mathrm{d}\boldsymbol{y}$. By the chain rule it follows for each $\boldsymbol{y}\in U$ that $\partial_z\exp\big(\theta\,\Phi^{\boldsymbol{y}}(z)\big)=\theta\,\exp\big(\theta\,\Phi^{\boldsymbol{y}}(z)\big)\,\partial_z\Phi^{\boldsymbol{y}}(z)$. Recalling from the previous subsection that $\partial_z(\frac{\alpha_3}{2}\|z\|_{L^2(V';I)}^2)=\alpha_3R_V^{-1}z$ and $\partial_z\Phi^{\boldsymbol{y}}(z)=\big((B_1^{\boldsymbol{y}})^\dagger\big)'(\alpha_1R_V+\alpha_2E_T'E_T)(u^{\boldsymbol{y}}(z)-\widehat{u})=q_1^{\boldsymbol{y}}$, and collecting terms gives (4.48). □

4.2.6 Optimality conditions

In the case when the feasible set of controls $\mathcal{Z}_{\mathrm{ad}}$ is a nonempty and convex set, we know (see, e.g., theorem 3.6.2) variational inequality

$$\big\langle J'(z^*),z-z^*\big\rangle_{L^2(V;I),L^2(V';I)}\geqslant 0\quad\forall z\in\mathcal{Z}_{\mathrm{ad}}\,. \tag{4.49}$$

For convex objective functionals $J(z)$, like the ones considered in this work, the variational inequality is a necessary and sufficient condition for optimality. The complete optimality conditions are then given by the following result.

Theorem 4.2.11. *Let $\mathcal{R}$ be the expected value or the entropic risk measure. A control $z^*\in\mathcal{Z}_{\mathrm{ad}}$ is the unique minimizer of (4.18) subject to (4.19) and (4.20) if and only if it satisfies the optimality system:*

$$\left.\begin{aligned}\big\langle B^{\boldsymbol{y}}u^{\boldsymbol{y}},(v_1,v_2)\big\rangle_{\mathcal{Y}',\mathcal{Y}}&=\langle z^*,v_1\rangle_{L^2(V';I),L^2(V;I)}+\langle u_0,v_2\rangle_{L^2(D)}\quad&\forall\,v\in\mathcal{Y},\\ \big\langle w,(B^{\boldsymbol{y}})'q^{\boldsymbol{y}}\big\rangle_{\mathcal{X},\mathcal{X}'}&=\big\langle w,\alpha_1R_V(u^{\boldsymbol{y}}-\widehat{u})\big\rangle_{\mathcal{X},\mathcal{X}'}&\\ &\quad+\big\langle w(T),\alpha_2(u^{\boldsymbol{y}}(T)-\widehat{u}(T))\big\rangle_{L^2(D)}&\forall\,w\in\mathcal{X},\end{aligned}\right\}\forall\boldsymbol{y}\in U,$$

$$\big\langle J'(z^*),z-z^*\big\rangle_{L^2(V;I),L^2(V';I)}\geqslant 0\quad\forall z\in\mathcal{Z}_{\mathrm{ad}}\,,$$

where $J'(z)$ is given by (4.41) for the expected value, or (4.48) for the entropic risk measure.

Observe that the optimality system in Theorem 4.2.11 contains the variational formulations of the state PDE (4.24) and the dual PDE (4.32) in the first and second equation, respectively.
It is convenient to reformulate the variational inequality (4.49) in terms of an orthogonal projection onto $\mathcal{Z}_{\mathrm{ad}}$. The orthogonal projection onto a nonemtpy, closed and convex subset $\mathcal{Z}_{\mathrm{ad}} \subset H$ of a Hilbert space H, denoted by $P_{\mathcal{Z}_{\mathrm{ad}}} : \mathcal{Z}_{\mathrm{ad}} \to H$, is defined as

$$P_{\mathcal{Z}}(h) \in \mathcal{Z}\,, \quad \|P_{\mathcal{Z}}(h) - h\|_H = \min_{v \in \mathcal{Z}} \|v - h\|_H\,, \quad \forall h \in H\,.$$

Then, see, e.g., [89, Lemma 1.11], for all $h \in H$ and $\gamma > 0$ the condition $h \in \mathcal{Z}_{\mathrm{ad}}$, $\langle h, v - z\rangle_H \geqslant 0\, \forall v \in \mathcal{Z}$ is equivalent to $z - P_{\mathcal{Z}}(z - \gamma h) = 0$. Using the definition of the Riesz operator and $H = L^2(V'; I)$, we conclude that (4.49) is equivalent to

$$z^* - P_{\mathcal{Z}_{\mathrm{ad}}}(z^* - \gamma R_V J'(z^*)) = 0\,.$$

This equivalence can then be used to develop projected descent methods to solve the optimal control problem, see, e.g., [89, Chapter 2.2.2].

Remark 4.2.12. *If $\mathcal{Z}_{\mathrm{ad}}$ is the closed ball with radius $r > 0$ in a Hilbert space H, then the orthogonal projection $P_{\mathcal{Z}_{\mathrm{ad}}}$ onto $\mathcal{Z}_{\mathrm{ad}}$ is given by*

$$P_{\mathcal{Z}_{\mathrm{ad}}}(h) = \min\Big(1, \frac{r}{\|h\|_H}\Big)\, h \qquad \text{for all } h \in H.$$

4.3 Analytic parametric linear operator constraints

Let $U = [-\frac{1}{2}, \frac{1}{2}]^{\mathbb{N}}$ be the space of parameters and assume that the sequence of parameters $\boldsymbol{y} = (y_j)_{j \geqslant 1}$ is i.i.d. uniformly in $[-\frac{1}{2}, \frac{1}{2}]$ for each $j \in \mathbb{N}$, i.e., $\boldsymbol{y}$ is distributed on U with probability measure μ, where $\mu(\mathrm{d}\boldsymbol{y}) = \bigotimes_{j \geqslant 1} \mathrm{d}y_j = \mathrm{d}\boldsymbol{y}$. Let $\alpha > 0$ and $\mathcal{Z}_{\mathrm{ad}} \subseteq \mathcal{Z}$ be closed and convex. Given a target state $\widehat{u}$, our goal of computation is the following optimal control problem

$$\min_{z \in \mathcal{Z}_{\mathrm{ad}}, u \in \mathcal{X}_{\mathrm{ad}}} \hat{J}(u, z)\,, \quad \hat{J}(u, z) := \frac{1}{2}\mathcal{R}\left(\|\mathcal{Q}u - \widehat{u}\|_{\mathfrak{J}}^2\right) + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2\,, \tag{4.50}$$

subject to the parametric linear operator equation in $\mathcal{Y}'$

$$A(\boldsymbol{y})u = \mathcal{B}z\,, \tag{4.51}$$

for $1 \leqslant q < \infty$. Let $\mathcal{Z}$ be a Hilbert space, and $\mathcal{X}_{\mathrm{ad}} \subset \mathcal{X}$, where $\mathcal{X}$ is a separable, reflexive Banach space. Moreover, let $\mathfrak{J}$ be a Hilbert space, $\widehat{u} \in \mathfrak{J}$, and $\mathcal{Q} \in \mathcal{L}(\mathcal{X}, \mathfrak{J})$, $\mathcal{B} \in \mathcal{L}(\mathcal{Z}, \mathcal{Y}')$. In particular, the operators $\mathcal{B}$ and $\mathcal{Q}$ are not dependent on $\boldsymbol{y}$ and thus can be uniformly bounded for all $\boldsymbol{y}$, i.e., $\|\mathcal{B}\|_{\mathcal{L}(\mathcal{Z}, \mathcal{Y}')} \leqslant C_{\mathcal{B}}$ and $\|\mathcal{Q}\|_{\mathcal{L}(\mathcal{X}, \mathfrak{J})} \leqslant C_{\mathcal{Q}}$ for some $C_1, C_2 > 0$ and all $\boldsymbol{y} \in U$. The risk measure $\mathcal{R}$ will again be either the expected value or the entropic risk measure.
Assume that the parametric family of operators $A(\boldsymbol{y}) \in \{\mathcal{L}(\mathcal{X}, \mathcal{Y}') : \boldsymbol{y} \in U\}$ satisfies Assumption 2.3.1, i.e., is a regular p-analytic operator family for some $0 < p \leqslant 1$. This implies in particular that the constraint (4.51) is well-posed, that is for each $z \in \mathcal{Z}$ there exists a unique $u \in \mathcal{X}$ such that (4.51) is true for all $\boldsymbol{y} \in U$.
Hence, we can substitute $u = A^{-1}(\boldsymbol{y})\mathcal{B}z$ into J, which gives $J(z) := J(A^{-1}(\boldsymbol{y})\mathcal{B}z, z)$ and leads to the reduced formulation of problem (4.50) – (4.51)

$$\min_{z \in \mathcal{Z}_{\mathrm{ad}}} J(z)\,, \quad J(z) := \frac{1}{2}\mathcal{R}\left(\|\mathcal{Q}A^{-1}(\boldsymbol{y})\mathcal{B}z - \widehat{u}\|_{\mathfrak{J}}^2\right) + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2\,, \tag{4.52}$$

where $\mathcal{Z}_{\mathrm{ad}} := \{z \in \mathcal{Z} : z \in \mathcal{Z}_{\mathrm{ad}}, A^{-1}(\boldsymbol{y})\mathcal{B}z \in \mathcal{X}_{\mathrm{ad}}\}$.

4.3.1 Derivatives and dual problem

For linear risk measures $\mathcal{R}$, the Fréchet derivative of J at z is given by

$$J'(z) = \mathcal{R}\left(\mathcal{B}'(A^{-1}(\boldsymbol{y}))'\mathcal{Q}'R_{\mathfrak{J}}(u^{\boldsymbol{y}}(z) - \widehat{u})\right) + \alpha R_{\mathcal{Z}} z. \tag{4.53}$$

$$J(z+\delta) = \frac{1}{2}\mathcal{R}\left(\|\mathcal{Q}A^{-1}(\boldsymbol{y})\mathcal{B}\delta + \mathcal{Q}u(z) - \widehat{u}\|_{\mathfrak{J}}^2\right) + \frac{\alpha}{2}\|z+\delta\|_{\mathcal{Z}}^2$$

Expanding the squared norms using $\|v+w\|^2 = \langle v+w, v+w\rangle = \|v\|^2 + 2\langle v, w\rangle + \|w\|^2$, we obtain

$$J(z+\delta) \;=\; J(z) + (\partial_z J(z))\,\delta + \mathrm{o}(\delta),$$

with the Fréchet derivative $\partial_z J(z) : \mathcal{Z} \to \mathbb{R}$ at $z \in \mathcal{Z}$ defined by

$$\begin{aligned}(\partial_z J(z))\,\delta &:= \mathcal{R}\left(\left\langle \mathcal{Q}A^{-1}(\boldsymbol{y})\mathcal{B}\delta, u^{\boldsymbol{y}}(z) - \widehat{u}\right\rangle_{\mathfrak{J}}\right) + \alpha\langle\delta, z\rangle_{\mathcal{Z}} \\ &= \mathcal{R}\left(\left\langle \delta, \mathcal{B}'(A^{-1}(\boldsymbol{y}))'\mathcal{Q}'R_{\mathfrak{J}}(u^{\boldsymbol{y}}(z) - \widehat{u})\right\rangle_{\mathcal{Z},\mathcal{Z}'}\right) + \alpha\langle\delta, R_{\mathcal{Z}} z\rangle_{\mathcal{Z},\mathcal{Z}'}\end{aligned}$$

where $R_{\mathcal{Z}} : \mathcal{Z} \to \mathcal{Z}'$ denotes the Riesz operator $\langle v, w\rangle_{\mathcal{Z}} = \langle v, R_{\mathcal{Z}} w\rangle_{\mathcal{Z},\mathcal{Z}'}$ for arbitrary $v, w \in \mathcal{Z}$, and $R_{\mathfrak{J}} : \mathfrak{J} \to \mathfrak{J}'$ denotes the Riesz operator $\langle v, w\rangle_{\mathfrak{J}} = \langle v, R_{\mathfrak{J}} w\rangle_{\mathfrak{J},\mathfrak{J}'}$ for arbitrary $v, w \in \mathfrak{J}$.
Defining

$$q(\boldsymbol{y}) := (A^{-1}(\boldsymbol{y}))'\mathcal{Q}'R_{\mathfrak{J}}(u^{\boldsymbol{y}}(z) - \widehat{u}) \in \mathcal{Y}, \tag{4.54}$$

we observe that $q(\boldsymbol{y})$ for each $\boldsymbol{y} \in U$ solves the dual PDE problem in $\mathcal{X}'$

$$A(\boldsymbol{y})'q(\boldsymbol{y}) = \mathcal{Q}'R_{\mathfrak{J}}(u^{\boldsymbol{y}}(z) - \widehat{u}). \tag{4.55}$$

Moreover, the Fréchet derivative for linear risk measures (4.53) can directly be computed using (4.55)

$$J'(z) = \mathcal{R}\left(\mathcal{B}'q(\boldsymbol{y})\right) + \alpha R_{\mathcal{Z}} z. \tag{4.56}$$

In order to derive the Fréchet derivative of J with the entropic risk measure $\mathcal{R} = \mathcal{R}_{\mathrm{e}}$, we observe that the random variable objective funciton can be uniformly bounded for all $\boldsymbol{y} \in U$ by

$$\|\mathcal{Q}u^{\boldsymbol{y}}(z) - \widehat{u}\|_{\mathfrak{J}}^2 \leqslant 2\|\mathcal{Q}u^{\boldsymbol{y}}(z)\|_{\mathfrak{J}}^2 + 2\|\widehat{u}\|_{\mathfrak{J}}^2 \leqslant 2C_{\mathcal{Q}}CC_{\mathcal{B}}\|z\|_{\mathcal{Z}} + 2\|\widehat{u}\|_{\mathfrak{J}}^2.$$

Together with the chain rule, this leads

$$J'(z) = \frac{1}{\int_U \exp\left(\theta\|\mathcal{Q}u^{\boldsymbol{y}}(z) - \widehat{u}\|_{\mathfrak{J}}^2\right)\mathrm{d}\boldsymbol{y}} \int_U \exp\left(\theta\|\mathcal{Q}u^{\boldsymbol{y}}(z) - \widehat{u}\|_{\mathfrak{J}}^2\right)\mathcal{B}'q^{\boldsymbol{y}}\mathrm{d}\boldsymbol{y} + \alpha R_{\mathcal{Z}} z. \tag{4.57}$$

4.3.2 Optimality conditions

The uniformly boundedly invertible forward operator $A(\boldsymbol{y})$, fits into the framework of Theorem 3.5.4. In particular, the forward operator $A(\boldsymbol{y})$, the regularization term $\frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2$ and the random variable tracking-type objective function $\|\mathcal{Q}u^{\boldsymbol{y}}(z) - \widehat{u}\|_{\mathfrak{J}}^2$ satisfy Assumption 3.5.1, Assumption 3.5.3 and Assumption 3.3.1. We obtain the following result.

Corollary 4.3.1. *Let $\alpha > 0$ and let $\mathcal{R}$ be proper, closed, convex and monotonic, then there exists a unique solution of* (4.52).

Proof. The existence of the solution follows from Theorem 3.5.4. We thus only prove the strong convexity of the objective function, which implies strict convexity and hence uniqueness of the solution. Clearly the regularization $\frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2$ is strongly convex. Since the sum of a convex and a strongly convex function is strongly convex it remains to show the convexity of $\mathcal{R}(\|\mathcal{Q}u^{\boldsymbol{y}}(z)-\widehat{u}\|_{\mathfrak{J}}^2)$. By the linearity and the bounded invertibility of the linear forward operator $B^{\boldsymbol{y}}$, the tracking-type objective functional $\|\mathcal{Q}u^{\boldsymbol{y}}(z)-\widehat{u}\|_{\mathfrak{J}}^2$ is quadratic in z and hence convex. Then, by Lemma 3.2.1 we obtain that $\mathcal{R}(\|\mathcal{Q}u^{\boldsymbol{y}}(z)-\widehat{u}\|_{\mathfrak{J}}^2)$ is convex. □

Having ensured the existence of a unique optimal control $z^* \in \mathcal{Z}_{\mathrm{ad}}$, we derive necessary and sufficient optimality conditions. Therefore, let the feasible set of controls $\mathcal{Z}_{\mathrm{ad}}$ be nonempty and convex set, then we know (see, e.g., Theorem 3.6.2) that the variational inequality holds for an optimal control $z^* \in \mathcal{Z}_{\mathrm{ad}}$:

$$\langle J'(z^*), z - z^* \rangle_{\mathcal{Z}',\mathcal{Z}} \geqslant 0 \quad \forall z \in \mathcal{Z}_{\mathrm{ad}}\,. \tag{4.58}$$

For convex objective functionals $J(z)$, like the ones considered in this work, the variational inequality is a necessary and sufficient condition for optimality. The complete optimality conditions are then given by the following result.

Theorem 4.3.2. *Let $\mathcal{R}$ be the expected value or the entropic risk measure. A control $z^* \in \mathcal{Z}_{\mathrm{ad}}$ is the unique minimizer of* (4.52) *if and only if it satisfies the optimality system:*

$$\left.\begin{aligned} A(\boldsymbol{y})u(\boldsymbol{y}) &= \mathcal{B}z^* \\ A(\boldsymbol{y})'q(\boldsymbol{y}) &= (A^{-1}(\boldsymbol{y}))'\mathcal{Q}'R_{\mathfrak{J}}(u^{\boldsymbol{y}}(z) - \widehat{u}) \end{aligned}\right\} \forall \boldsymbol{y} \in U,$$
$$\langle J'(z^*), z - z^* \rangle_{\mathcal{Z}',\mathcal{Z}} \geqslant 0 \quad \forall z \in \mathcal{Z}_{\mathrm{ad}}\,,$$

where $J'(z)$ is given by (4.56) *for the expected value, or* (4.57) *for the entropic risk measure.*

The optimality system in Theorem 4.3.2 contains the forward problem and the dual problem in the first and second equation, respectively.
It is convenient to reformulate the variational inequality (4.58) in terms of an orthogonal projection onto $\mathcal{Z}_{\mathrm{ad}}$. Using the definition of the Riesz operator, we conclude that (4.58) is equivalent to

$$z^* - P_{\mathcal{Z}_{\mathrm{ad}}}(z^* - \gamma R_{\mathcal{Z}} J'(z^*)) = 0\,.$$

This equation can then be used to develop algorithms, such as the projected descent method, to solve the optimal control problem, see, e.g., Section 4.4 or [89, Chapter 2.2.2].

4.4 Projected gradient descent

Consider an abstract optimal control problem

$$\min_{z \in \mathcal{Z}_{\mathrm{ad}}} f(z) \tag{4.59}$$

with control constraints $z \in \mathcal{Z}_{\mathrm{ad}}$, where $\mathcal{Z}_{\mathrm{ad}}$ is a closed and convex subset of a Hilbert space $\mathcal{Z}$, and where $f : \mathcal{Z} \to \mathbb{R}$ is continuously Fréchet differentiable.
The application of a standard gradient descent step to feasible z_i might lead to infeasibility of $z_i - \eta \nabla J(z_i)$ even for small stepsizes $\eta > 0$. On the other hand, considering only those $\eta > 0$ for which $z_i - \eta \nabla J(z_i)$ stays feasible is not viable since this might result in very small step sizes η. The following algorithms are based on the orthogonal projection onto $\mathcal{Z}_{\mathrm{ad}}$ and can be used to compute a minimizer of (4.59).

Algorithm 1 Projected gradient descent

Input: feasible starting value $z \in \mathcal{Z}$

1: **while** $\|z - P_{\mathcal{Z}_{\mathrm{ad}}}(z - \nabla J(z))\|_{\mathcal{Z}} >$TOL **do**
2: find step size η using Algorithm 2
3: set $z := P_{\mathcal{Z}_{\mathrm{ad}}}(z - \eta \nabla J(z))$
4: **end while**

Algorithm 2 Projected Armijo rule

Input: current z, parameters $\beta, \gamma \in (0, 1)$
Output: step size $\eta > 0$

1: set $\eta := 1$
2: **while** $J(P_{\mathcal{Z}_{\mathrm{ad}}}(z - \eta \nabla J(z))) - J(z) > -\frac{\gamma}{\eta} \|z - P_{\mathcal{Z}_{\mathrm{ad}}}(z - \eta \nabla J(z))\|_{\mathcal{Z}}^2$ **do**
3: set $\eta := \beta \eta$
4: **end while**

Theorem 4.4.1. *Let $\mathcal{Z}_{\mathrm{ad}}$ be nonempty and let f be bounded from below. If ∇f is α-Hölder continuous on $\{w + s : f(w) \leqslant f(w^{(0)}), \|s\|_{\mathcal{Z}} \leqslant \rho\}$ for some $\alpha > 0$ and some $\rho > 0$, then the sequence $\{z_i\}$ generated by Algorithm 1 satisfies*

$$\lim_{i \to \infty} \|z_i - P_{\mathcal{Z}_{\mathrm{ad}}}(z_i - \nabla J(z_i))\|_{\mathcal{Z}} = 0 \,,$$

where $P_{\mathcal{Z}_{\mathrm{ad}}}$ is defined by (4.60). Moreover, the sequence $\{z_i\}$ converges to the unique solution z^* of (4.59).

Proof. See. e.g., [89, Theorem 2.4]. □

Both, the elliptic example and the parabolic example fit into this framework: consider now problem (4.15) with

$$-\infty < z_{\min} < z_{\max} < \infty \quad \text{a.e. in } D \,,$$

i.e., $\mathcal{Z}_{\mathrm{ad}} \subsetneq \mathcal{Z} = L^2(D)$.
To incorporate these constraints we use the projection $P_{\mathcal{Z}_{\mathrm{ad}}}$ onto $\mathcal{Z}_{\mathrm{ad}}$ given by

$$P_{\mathcal{Z}_{\mathrm{ad}}}(z)(\boldsymbol{x}) = P_{[z_{\min}(\boldsymbol{x}), z_{\max}(\boldsymbol{x})]}(z(\boldsymbol{x})) = \max(z_{\min}(\boldsymbol{x}), \min(z(\boldsymbol{x}), z_{\max}(\boldsymbol{x}))) \,, \tag{4.60}$$

and perform a line search along the projected path $\{P_{\mathcal{Z}_{\mathrm{ad}}}(z_i - \eta \nabla J(z_i)) : \eta > 0\}$. Using the optimality condition $z^* - P_{\mathcal{Z}_{\mathrm{ad}}}(z^* - \nabla J(z^*)) = 0$ leads to the projected gradient descent algorithm Algorithm 1, which is justified by Theorem 4.4.1.

Proof. For the proof of we refer to [89, Theorem 2.4]. □

Consider now problem (3.25) with $\mathcal{Z}_{\mathrm{ad}} = \overline{B_r(0)} \subset L^2(V'; I)$ being the closed ball in V' with radius $r > 0$ centered at the origin. In view of Remark 4.2.12 we use the orthogonal projection

$$P_{\mathcal{Z}_{\mathrm{ad}}}(z)(\boldsymbol{x}) = \min\left(1, \frac{r}{\|h\|_{\mathcal{Z}}}\right) z, \quad \forall z \in \mathcal{Z}, \tag{4.61}$$

to incorporate the control constraints. Using the Riesz operator R_V (see (4.22)), we can define the Riesz representation ∇J of the Fréchet derivative J' of J by

$$\langle \nabla J(z), h \rangle_{L^2(V';I),L^2(V';I)} = \langle R_V J'(z), h \rangle_{L^2(V';I),L^2(V';I)}, \quad \text{for all } z, h \in L^2(V'; I).$$

Hence, we can use Algorithm 1 together with Algorithm 2 to find the optimal control $z^* \in \mathcal{Z}_{\mathrm{ad}}$.

Remark 4.4.2. *In the cases without control constraints, i.e., $z_{\min} = -\infty$, $z_{\max} = \infty$ in* (4.60) *and $r = \infty$ in* (4.61)*, the projection becomes the identity.*

For more sophisticated methods, such as Newton-type methods, we refer the reader to [89, Chapter 2].

4.4.1 Numerical experiments

In the following we apply the gradient descent method and its projected version to solve the optimal control problems described in Section 4.1 and Section 4.2.

Elliptic example

We consider the optimal control problem described in Section 4.1, i.e., the problem of finding the optimal control $z \in \mathcal{Z}_{\mathrm{ad}}$ that minimizes (4.1) subject to the elliptic state PDE (4.2) – (4.3) and the control constraints (4.4). Suppose the PDE is defined in the two-dimensional physical domain $D = (0,1)^2$, and equipped with the diffusion coefficient (4.5). We set $a_0(\boldsymbol{x}) \equiv 1$ as the mean field and use the parameterized family of fluctuations

$$\psi_j(\boldsymbol{x}) = \frac{1}{(k_j^2 + \ell_j^2)^{\vartheta}} \sin(\pi k_j x_1) \sin(\pi \ell_j x_2) \quad \text{for } \vartheta > 1 \text{ and } j \in \mathbb{N}, \tag{4.62}$$

where the sequence $(k_j, \ell_j)_{j \geqslant 1}$ is an ordering of the elements of $\mathbb{N} \times \mathbb{N}$, so that the sequence $(\|\psi_j\|_{L^\infty(D)})_{j \geqslant 1}$ is non-increasing. This implies that $\|\psi_j\|_{L^\infty(D)} \sim j^{-\vartheta}$ as $j \to \infty$ by Weyl's asymptotic law for the spectrum of the Dirichlet Laplacian (cf. [152] as well as the examples in [39, 53]). The target state is chosen to be $\widehat{u}(\boldsymbol{x}) = x_1^2 - x_2^2$ for $\boldsymbol{x} = (x_1, x_2) \in D$.
We use piecewise linear finite elements with mesh width $h = 2^{-6}$ to discretize the spatial domain $D = (0,1)^2$, see Section 7.1 for more details on the FE method. The integrals over the parametric domain U are discretized using a lattice rule with a single fixed random shift with $n = 2^{15}$ points and the truncation dimension $s = 2^{12}$, see Chapter 5 and Chapter 6 for the details. More precisely, the lattice QMC rule was generated by using

the fast component-by-component (CBC) implementation of the QMC4PDE software [113, 114], with the weights chosen to appropriately accommodate the fluctuations (4.62) in accordance with theorem 6.2.7 (see Chapter 6 for details on quasi-Monte Carlo methods). In particular, we note that while all the lattice rules in the subsequent numerical examples were designed with the adjoint solution q in mind, the same lattice rules have been used in the sequel to analyze the behavior of the state PDE u as well.
We set $\vartheta = 1.5$, and fix the space of admissible controls $\mathcal{Z}_{\text{ad}} = \{z \in L^2(D) : z_{\min} \leqslant z \leqslant z_{\max} \text{ a.e. in } D\}$ with

$$z_{\min}(\boldsymbol{x}) = \begin{cases} 0 & \boldsymbol{x} \in \left[\frac{1}{8}, \frac{3}{8}\right] \times \left[\frac{5}{8}, \frac{7}{8}\right], \\ 0 & \boldsymbol{x} \in \left[\frac{5}{8}, \frac{7}{8}\right] \times \left[\frac{5}{8}, \frac{7}{8}\right], \\ -1 & \text{otherwise} \end{cases} \quad \text{and} \quad z_{\max}(\boldsymbol{x}) = \begin{cases} 0 & \boldsymbol{x} \in \left[\frac{1}{8}, \frac{3}{8}\right] \times \left[\frac{1}{8}, \frac{3}{8}\right], \\ 0 & \boldsymbol{x} \in \left[\frac{5}{8}, \frac{7}{8}\right] \times \left[\frac{1}{8}, \frac{3}{8}\right], \\ 1 & \text{otherwise.} \end{cases}$$

We consider the regularization parameters $\alpha \in \{0.1, 0.01\}$ for the minimization problem. To minimize the discretized target functional, we use the projected gradient descent algorithm (Algorithm 1) in conjunction with the projected Armijo (Algorithm 2) rule with control parameter $\gamma = 10^{-4}$ and backtracking from $\eta = 1$ with the update $\eta \leftarrow 0.5\eta$. For both experiments, we used $z_0(\boldsymbol{x}) = z_0(x_1, x_2) = P_{\mathcal{Z}}(0, x_2)$, with $P_{\mathcal{Z}_{\text{ad}}}(z)(x_1, x_2) := \max(z_{\min}(x_1, x_2), \min(z(x_1, x_2), z_{\max}(x_1, x_2)))$, as the initial guess and track the averaged least square difference of the state u and the target state $\widehat{u}$. The results are displayed in Figure 4.1. We observe that for a larger value of α the algorithm converges faster and the averaged difference between the state u and the target state $\widehat{u}$ increases.
The same behaviour is observed in the unconstrained case with $\mathcal{Z}_{\text{ad}} = L^2(D)$. We fix the same parameters as before and use the gradient descent algorithm together with the

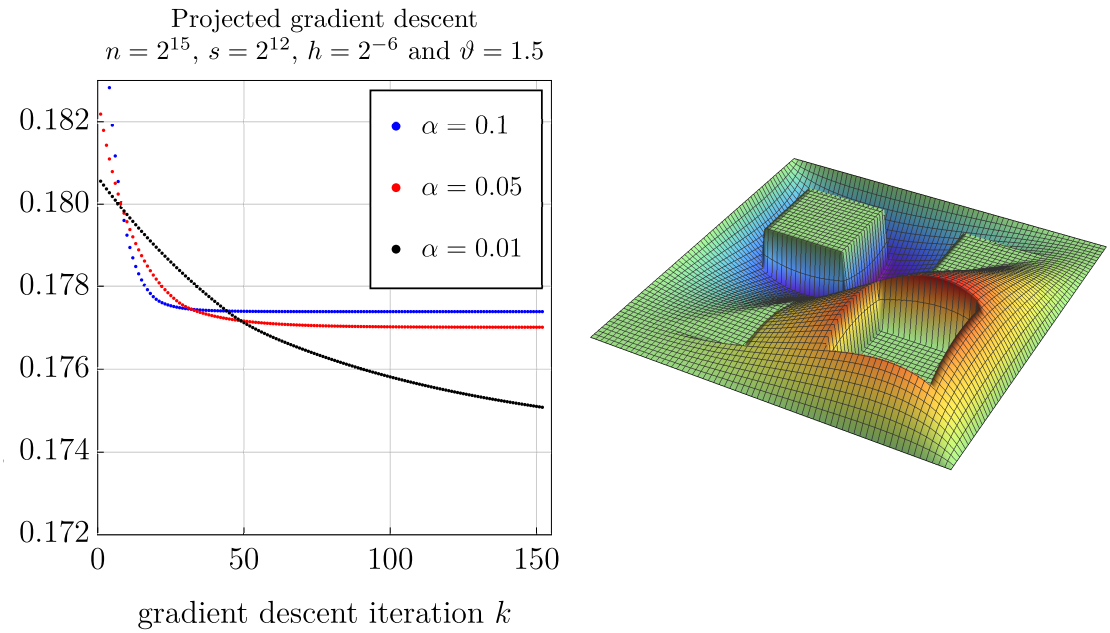


Figure 4.1: Left: Averaged least square difference of the state u and the target state $\widehat{u}$ at each step of the projected gradient descent algorithm $\int_{U_s} \|u(\cdot, \boldsymbol{y}, z_k) - \widehat{u}\|^2_{L^2(D)} \, \mathrm{d}\boldsymbol{y}$ for different values of the regularization parameter α. Right: The control corresponding to $\alpha = 0.1$ after 152 projected gradient descent iterations.

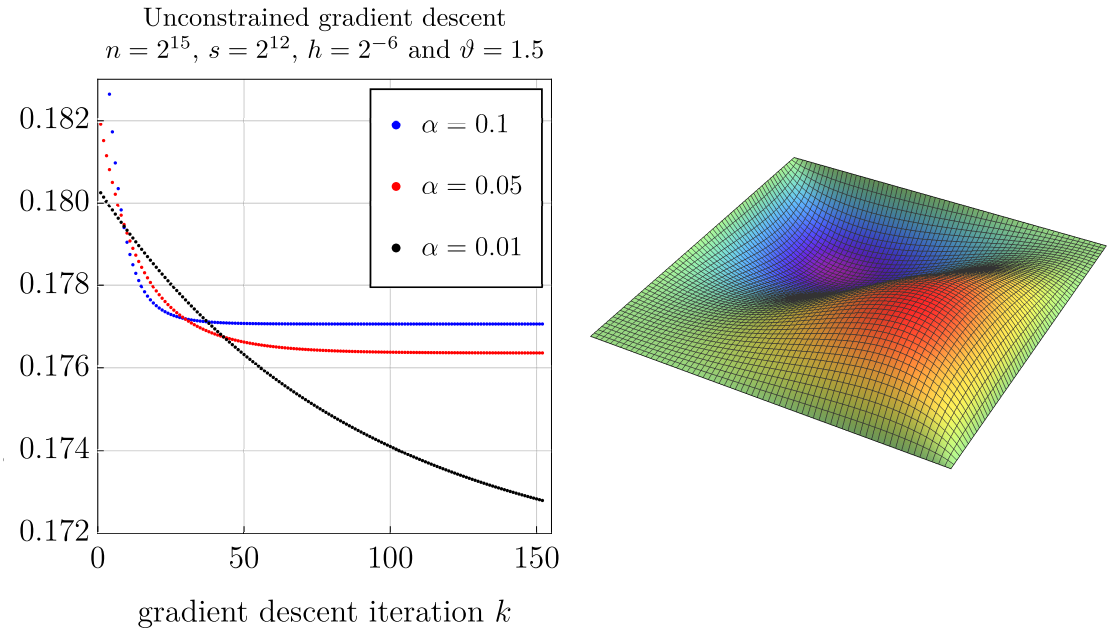


Figure 4.2: Left: Averaged least square difference of the state u and the target state $\widehat{u}$ at each step of the gradient descent algorithm $\int_{U_s} \|u(\cdot, \boldsymbol{y}, z_k) - \widehat{u}\|^2_{L^2(D)} \, \mathrm{d}\boldsymbol{y}$ for different values of the regularization parameter α. Right: The control corresponding to $\alpha = 0.1$ after 152 gradient descent iterations.

Armijo rule with control parameter $\gamma = 10^{-4}$ and backtracking from $\eta = 1$ with the update $\eta \leftarrow 0.5\eta$. Again we refer to [89] for details on the algorithm and convergence results. We choose $z_0(\boldsymbol{x}) = x_2$ as the initial guess and track the averaged least square difference of the state u and the target state $\widehat{u}$. The results are displayed in Figure 4.2.

Parabolic example

We consider the optimal control problem in Section 4.2, i.e., we aim to minimize (4.38). We fix the physical domain $D = (0,1)^2$ and the terminal time $T = 1$. The uncertain diffusion coefficient, defined as in (4.21), is independent of t, and parameterized with mean field $a_0(\boldsymbol{x}) \equiv 1$ and the fluctuations

$$\psi_j(\boldsymbol{x}) = \frac{1}{2} j^{-\vartheta} \sin(\pi j x_1) \sin(\pi j x_2) \quad \text{for } \vartheta > 1 \text{ and } j \in \mathbb{N}.$$

We use the implicit Euler finite difference scheme with step size $\Delta t = \frac{T}{500} = 2 \cdot 10^{-3}$ to discretize the PDE system with respect to the temporal variable[22]. The spatial part of the PDE system is discretized using a first order finite element method with mesh size $h = 2^{-5}$ (see Section 7.1 for details) and the Riesz operator in the loading term of the adjoint PDE can be evaluated using (4.22). The integrals in the experiments are approximated using lattice rules (see Chapter 6 for details) that are generated using the fast CBC algorithm with weights chosen as in Theorem 6.2.7, where we used the parameter value $\beta_1 = 1$ in (4.86). As the target state we choose

$$\widehat{u}(\boldsymbol{x}, t) := \chi_{\|\boldsymbol{x} - (c_1(t), c_2(t))\|_\infty \leq \frac{1}{10}}(\boldsymbol{x}) \, \widehat{u}_1(\boldsymbol{x}, t) + \chi_{\|\boldsymbol{x} + (c_1(t), c_2(t)) - (1,1)\|_\infty \leq \frac{1}{10}}(\boldsymbol{x}) \, \widehat{u}_2(\boldsymbol{x}, t),$$

[22] We refer to [154, Chapter 12] for details on discontinuous Galerkin FEM and the connection to the implicit Euler scheme.

where

$$\begin{aligned}
\widehat{u}_1(\boldsymbol{x},t) &:= 10240\Big(x_1 - c_1(t) - \frac{1}{10}\Big)\Big(x_2 - c_2(t) - \frac{1}{10}\Big) \\
&\quad \times \Big(x_1 - c_1(t) + \frac{1}{10}\Big)\Big(x_2 - c_2(t) + \frac{1}{10}\Big), \\
\widehat{u}_2(\boldsymbol{x},t) &:= 10240\Big(x_1 + c_1(t) - \frac{11}{10}\Big)\Big(x_2 + c_2(t) - \frac{11}{10}\Big) \\
&\quad \times \Big(x_1 + c_1(t) - \frac{9}{10}\Big)\Big(x_2 + c_2(t) - \frac{9}{10}\Big), \\
c_1(t) &:= \frac{1}{2} + \frac{1}{4}(1-t^{10})\cos(4\pi t^2) \quad \text{and} \quad c_2(t) := \frac{1}{2} + \frac{1}{4}(1-t^{10})\sin(4\pi t^2).
\end{aligned}$$

Moreover, we set the parameters appearing in the objective functional (4.18) and adjoint equation (4.43) to $\alpha_1 = 10^{-3}$, $\alpha_2 = 10^{-2}$, and $\alpha_3 = 10^{-7}$. Furthermore, the initial state is

$$u_0(\boldsymbol{x}) = \sin(2\pi x_1)\sin(2\pi x_2)$$

.

We consider the problem of finding the optimal control $z \in \mathcal{Z}_{\mathrm{ad}}$ that minimizes (4.18) subject to the PDE constraint (4.20). We consider constrained optimization over $\mathcal{Z}_{\mathrm{ad}} = \{z \in L^2(V';I) : \|z\|_{L^2(V';I)} \leqslant 2\}$ and compare our results with the reconstruction obtained by carrying out unconstrained optimization over $\mathcal{Z} = L^2(V';I)$. To this end, we define the projection operator

$$P_{\mathcal{Z}_{\mathrm{ad}}}(w) := \min\left\{2, \frac{2}{\|w\|_{L^2(V;I)}}\right\} w \quad \text{for } w \in L^2(V;I)$$

which is used in the constrained setting, while in the unconstrained setting we use $P_{\mathcal{Z}_{\mathrm{ad}}} := \mathcal{I}_{L^2(V;I)}$.

Algorithm 3 Projected gradient descent

Input: feasible starting value $w \in L^2(V;I)$ such that $z = R_V w \in \mathcal{Z}$

1: **while** $\|w - P_{\mathcal{Z}_{\mathrm{ad}}}(w - J'(R_V w))\|_{L^2(V;I)} >$TOL **do**
2: find step size η using Algorithm 4
3: set $w := P_{\mathcal{Z}_{\mathrm{ad}}}(w - \eta J'(R_V w))$
4: **end while**

Algorithm 4 Projected Armijo rule

Input: current $w \in L^2(V;I)$, parameters $\beta, \gamma \in (0,1)$ and $\eta_0 > 0$
Output: step size $\eta > 0$

1: set $\eta := \eta_0$
2: **while**
3: $J(P_{\mathcal{Z}_{\mathrm{ad}}}(w - \eta J'(R_V w))) - J(R_V w) > -\frac{\gamma}{\eta}\|w - P_{\mathcal{Z}_{\mathrm{ad}}}(w - \eta J'(R_V w))\|^2_{L^2(V;I)}$ **do**
4: set $\eta := \beta\eta$
5: **end while**

To be able to handle elements of $\mathcal{Z}_{\mathrm{ad}} \subseteq L^2(V';I)$ numerically, we apply the projected gradient method as described in Algorithm 3 together with the projected Armijo rule

stated in Algorithm 4. Note that, Algorithm 3 and Algorithm 4 coincide with Algorithm 1 and Algorithm 2. However, Algorithm 3 and Algorithm 4 are presented to illustrate the precise application of the Riesz operator R_V. Moreover, evaluating $J(R_V w)$ and $J'(R_V w)$ in Algorithm 3 and Algorithm 4 requires solving the state PDE with the source term $R_V w$. In particular, the Riesz operator appears in the loading term after finite element discretization and can thus be evaluated using (4.22). We use the initial guess $w_0 = 0$. The parameters of the gradient descent method were chosen to be $\eta_0 = 100$, $\gamma = 10^{-4}$, and $\beta = 0.1$.

We consider the entropic risk measure with $\theta = 10$ and set $\vartheta = 1.3$. The reconstructed optimal control obtained using the bounded set of feasible controls $\mathcal{Z}_{\text{ad}}$ is displayed in Figure 4.3. The reconstructed optimal control at the terminal time $T = 1$ and its pointwise difference to the control obtained without imposing control constraints are displayed in Figure 4.4. Finally, the evolution of the objective functional as the number of gradient descent iterations increases is plotted in Figure 4.5 for the constrained and unconstrained optimization problems.

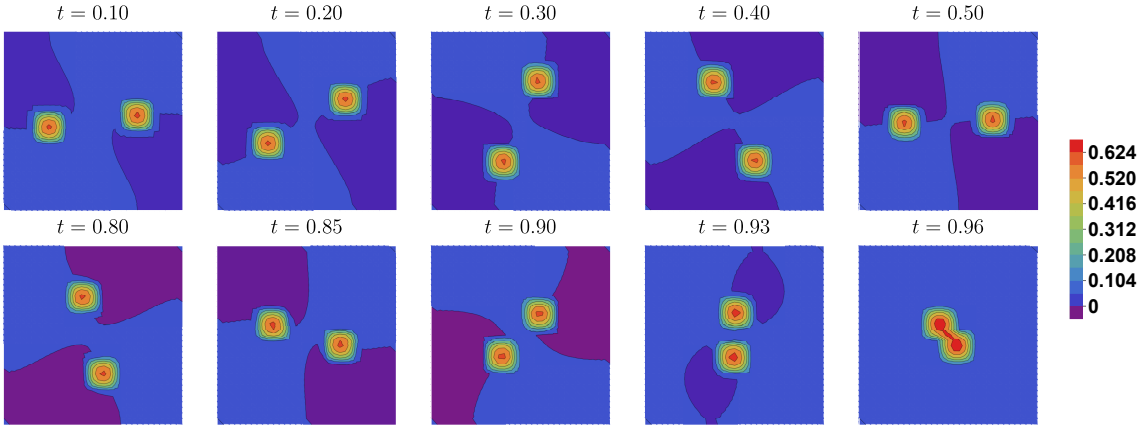


Figure 4.3: The inverse Riesz transform $R_V^{-1} z^*$ of the reconstructed optimal control z^* using the entropic risk measure for several values of t in the constrained setting.

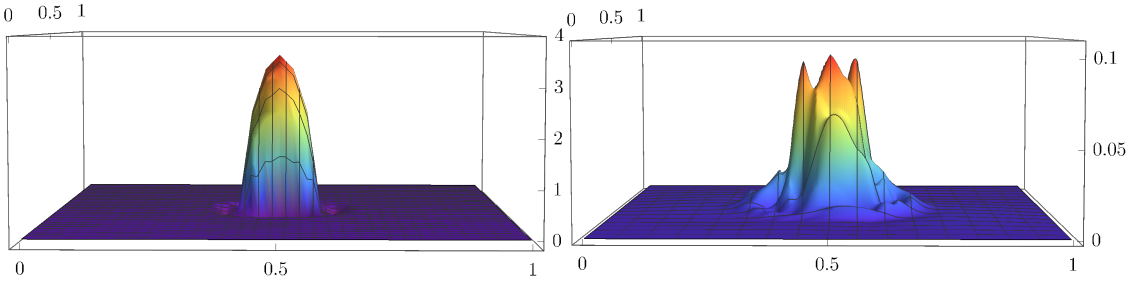


Figure 4.4: Left: the inverse Riesz transform of the control at time $t = 1$ in the constrained setting after 25 iterations of the projected gradient descent algorithm using the entropic risk measure. Right: The difference between the reconstruction obtained in the constrained setting and the corresponding solution in the unconstrained setting.

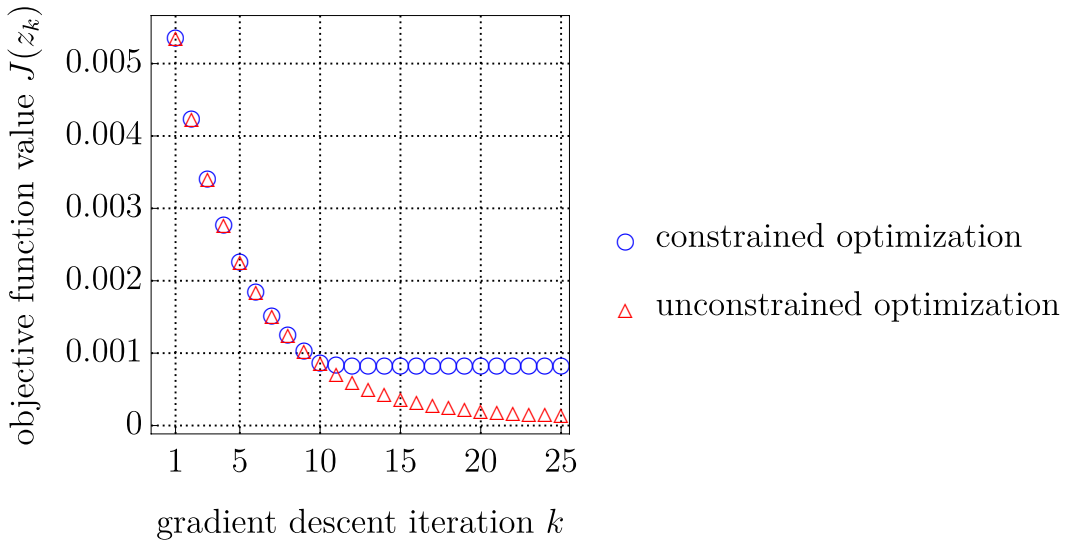


Figure 4.5: The value of the objective functional for each gradient descent iteration. The results corresponding to the constrained setting and the unconstrained setting are plotted in blue and red, respectively.

4.5 Error contributions and error expansion

In this section we start the error analysis of the optimal control problems by decomposing the overall error into its contributions. More precisely, we use the convexity of the objective functional of the optimal control problems to derive bounds on the error in the optimal control in terms of the corresponding adjoint states. The error in the adjoint states is then decomposed into its contributions and analyzed separately in Chapter 5, Chapter 6, and Section 7.1. In fact, the theoretical results for the different error contributions developed in this thesis are not limited to the application to optimal control problems, but cover a much broader class of problems, see Chapter 5, Chapter 6, and Section 7.1 for further details.
Let $\mathcal{X}$ be a separable Banach space and let the dimensionally truncated sequence $\boldsymbol{y} \in U$ be denoted by

$$\boldsymbol{y}_{\leqslant s} := (y_1, y_2, \ldots, y_s, 0, 0, \ldots).$$

In the case of an affine parameterization of the random input field (4.5) or (4.21) this corresponds to a truncation of the series at s terms, i.e.,

$$a(\boldsymbol{x}, \boldsymbol{y}) = a_0(\boldsymbol{x}) + \sum_{j \geqslant 1} y_j \psi_j(\boldsymbol{x}) \approx a_0(\boldsymbol{x}) + \sum_{j=1}^{s} y_j \psi_j(\boldsymbol{x}) = a(\boldsymbol{x}, \boldsymbol{y}_{\leqslant s}), \quad \boldsymbol{x} \in D.$$

The involved integrals then become integrals over a finite-dimensional domain U_s, i.e.,

$$I(f) := \int_U f(\boldsymbol{y})\, \mu(\mathrm{d}\boldsymbol{y}) \approx \int_{U_s} f(\boldsymbol{y}_{\leqslant s})\, \mu(\mathrm{d}\boldsymbol{y}_{\leqslant s}) =: I_s(f), \quad f \in L^1_\mu(U, \mathcal{X}).$$

Dimensionally truncated quantities are denoted with the subscript s, e.g., $u_s(\boldsymbol{y}) := u(\boldsymbol{y}_{\leqslant s})$. In Section 5.3 we provide error bounds and convergence rates for the dimension truncation error in a very general setting.

Once the dimension of the parameterization has been truncated, we employ an n-point cubature rule to approximate the s-dimensional integrals, i.e.,

$$I_s(f) = \int_{U_s} f(\boldsymbol{y}_{\leqslant s})\,\mu(\mathrm{d}\boldsymbol{y}_{\leqslant s}) \approx \sum_{i=1}^{n} \alpha_i f(\boldsymbol{y}^{(i)}) =: Q_{s,n}(f), \quad f \in L^1_\mu(U_s, \mathcal{X})$$

for cubature weights $\alpha_i \in \mathbb{R}$ for each $i = 1, \dots, n$ and nodes $\boldsymbol{y}^{(i)} \in \mathbb{R}^s$ carefully chosen according to a cubature rule. In this work we focus on randomly shifted rank-1 lattice rules, which are quasi-Monte Carlo rules for integration, see Section 7.2.2 for a precise describtion of the methods with rigorous error bounds and convergence rates. QMC methods are particularly well-suited for optimization since they preserve convexity due to their equal weights, i.e., $\alpha_i = \frac{1}{n}$ for all $i = 1, \dots, n$.
Quantities that depend on the number of cubature points n will be denoted with the subscript n.
Finally, one can use, e.g., finite element methods to discretize the PDEs in the spatial variables. Spatially discretized objects will be denoted with the subscript h, e.g., z_h denotes the discretized control on a finite element subspace.

4.5.1 Elliptic PDE constraint

In the following we consider a discretization of problem (4.1), (4.2), (4.3), and (4.4). Given $s \in \mathbb{N}$ and $\boldsymbol{y} \in U$, we truncate the sum in (4.5) after s terms, i.e., we set $y_j = 0$ for $j \geqslant s+1$. For every $\boldsymbol{y} \in U$ and every control $z \in L^2(D)$ we denote by

$$u_s(\boldsymbol{y}, \cdot, z) := u(\boldsymbol{y}_{\leqslant s}, \cdot, z)$$

the dimensionally truncated state, i.e., the unique solution of the parametric weak problem (4.8) corresponding to the dimensionally truncated diffusion coefficient $a_s(\boldsymbol{y}) := a(\boldsymbol{y}_{\leqslant s})$. Similarly we write $q_s(\boldsymbol{y}, \cdot, z) := q(\boldsymbol{y}_{\leqslant s}, \cdot, z)$ for any $\boldsymbol{y} \in U$ and any $z, \widehat{u} \in L^2(D)$ for the unique solution of the adjoint parametric weak problem (4.17) corresponding to the dimensionally truncated diffusion coefficient and truncated right-hand side $u_s(\boldsymbol{y}, \cdot, z) - \widehat{u}$. We further assume that we have access only to a finite element discretization $u_{s,h}(\boldsymbol{y}, \cdot, z)$ of the truncated state, to be defined precisely in Section 7.1. We write $q_{s,h}(\boldsymbol{y}, \cdot, z)$ for the finite element discretization of the truncated adjoint state corresponding to $u_{s,h}(\boldsymbol{y}, \cdot, z)$. We also write $u_s(\boldsymbol{y}, \cdot, z) = S_{\boldsymbol{y}_{\leqslant}} z$ and $q_s(\boldsymbol{y}, \cdot, z) = S_{\boldsymbol{y}_{\leqslant}}(u_s(\boldsymbol{y}, \cdot, z) - \widehat{u})$ in conjunction with $u_{s,h}(\boldsymbol{y}, \cdot, z) = S_{\boldsymbol{y}_{\leqslant s},h} z$ and $q_{s,h}(\boldsymbol{y}, \cdot, z) = S_{\boldsymbol{y}_{\leqslant s},h}(u_{s,h}(\boldsymbol{y}, \cdot, z) - \widehat{u})$.
Finally we use an n-point quasi-Monte Carlo approximation for the integral over U_s leading to the following discretization of (4.15)

$$\min_{z \in \mathcal{Z}_{\mathrm{ad}}} J_{s,h,n}(z)\,, \quad J_{s,h,n}(z) := \frac{1}{2n} \sum_{i=1}^{n} \|S_{\boldsymbol{y}^{(i)},h} z - \widehat{u}\|^2_{L^2(D)} + \frac{\alpha}{2} \|z\|^2_{L^2(D)}\,, \tag{4.63}$$

for quadrature points $\boldsymbol{y}^{(i)} \in U_s$, $i \in \{1, \dots, n\}$, to be defined precisely in Section 7.2.2. Since the error splitting will be based on the convexity of the problem, it is important to use cubature methods with positive weights in order to retain the convexity after discretization. In analogy to (4.16) it follows that the gradient of $J_{s,h,n}$, i.e., the representer of the Fréchet derivative of $J_{s,h,n}$ is given by

$$\nabla J_{s,h,n}(z) = \frac{1}{n} \sum_{i=1}^{n} q_{s,h}(\boldsymbol{y}^{(i)}, \cdot, z) + \alpha z\,.$$

Due to the positive weights of the quadrature rule, (4.63) is still a convex minimization problem. Existence and uniqueness of the solution $z^*_{s,h,n}$ of (4.63) follow by the previous arguments.
We note that the optimal control $z^*_{s,h,n}$ is implicitly discretized in terms of the FE discretization of the solution operator, see [88].

Lemma 4.5.1. *Let z^* be the unique minimizer of* (4.15) *and let $z^*_{s,h,n}$ be the unique minimizer of* (4.63)*. Then, we have*

$$\|z^* - z^*_{s,h,n}\|_{L^2(D)} \leqslant \frac{1}{\alpha} \left\| \int_U q(\boldsymbol{y}, \cdot, z^*) \,\mathrm{d}\boldsymbol{y} - \frac{1}{n} \sum_{i=1}^{n} q_{s,h}(\boldsymbol{y}^{(i)}, \cdot, z^*) \right\|_{L^2(D)}, \tag{4.64}$$

for quadrature points $\boldsymbol{y}^{(i)} \in \left[-\frac{1}{2}, \frac{1}{2}\right]^s$, $i \in \{1, \dots, n\}$.

Proof. By the optimality of $z^*_{s,h,n}$ it holds that $\langle \nabla J_{s,h,n}(z^*_{s,h,n}), z - z^*_{s,h,n} \rangle_{L^2(D)} \geqslant 0$ for all $z \in \mathcal{Z}_{\mathrm{ad}}$, and thus in particular $\langle \nabla J_{s,h,n}(z^*_{s,h,n}), z^* - z^*_{s,h,n} \rangle_{L^2(D)} \geqslant 0$. Similarly it holds for all $z\mathcal{Z}_{\mathrm{ad}}$ that $\langle \nabla J(z^*), z - z^* \rangle_{L^2(D)} \geqslant 0$ and thus in particular $\langle -\nabla J(z^*), z^* - z^*_{s,h,n} \rangle_{L^2(D)} \geqslant 0$. Adding these inequalities leads to

$$\big\langle \nabla J_{s,h,n}(z^*_{s,h,n}) - \nabla J(z^*), z^* - z^*_{s,h,n} \big\rangle_{L^2(D)} \geqslant 0\,.$$

Thus

$$\begin{aligned}
\alpha \|z^* - z^*_{s,h,n}\|^2_{L^2(D)} &\leqslant \alpha \|z^* - z^*_{s,h,n}\|^2_{L^2(D)} + \big\langle \nabla J_{s,h,n}(z^*_{s,h,n}) - \nabla J(z^*), z^* - z^*_{s,h,n} \big\rangle_{L^2(D)} \\
&= \big\langle \nabla J_{s,h,n}(z^*_{s,h,n}) - \alpha z^*_{s,h,n} - \nabla J(z^*) + \alpha z^*, z^* - z^*_{s,h,n} \big\rangle_{L^2(D)} \\
&= \big\langle \nabla J_{s,h,n}(z^*_{s,h,n}) - \alpha z^*_{s,h,n} - \nabla J_{s,h,n}(z^*) + \alpha z^*, z^* - z^*_{s,h,n} \big\rangle_{L^2(D)} \\
&\quad + \big\langle \nabla J_{s,h,n}(z^*) - \alpha z^* - \nabla J(z^*) + \alpha z^*, z^* - z^*_{s,h,n} \big\rangle_{L^2(D)} \\
&\leqslant \left\langle \frac{1}{n} \sum_{i=1}^{n} q_{s,h}(\boldsymbol{y}^{(i)}, \cdot, z^*) - \int_U q(\boldsymbol{y}, \cdot, z^*) \,\mathrm{d}\boldsymbol{y}, z^* - z^*_{s,h,n} \right\rangle_{L^2(D)} \\
&\leqslant \left\| \frac{1}{n} \sum_{i=1}^{n} q_{s,h}(\boldsymbol{y}^{(i)}, \cdot, z^*) - \int_U q(\boldsymbol{y}, \cdot, z^*) \,\mathrm{d}\boldsymbol{y} \right\|_{L^2(D)} \|z^* - z^*_{s,h,n}\|_{L^2(D)}\,,
\end{aligned}$$

where in the fourth step we used the strong convexity of the objective function, i.e., we used $\langle \nabla J_{s,h,n}(z^*_{s,h,n}) - \nabla J_{s,h,n}(z^*) + \alpha(z^* - z^*_{s,h,n}), z^* - z^*_{s,h,n} \rangle \leqslant 0$. The result then follows from $\alpha > 0$. □

We can split up the error on the right-hand side in (4.64) into dimension truncation error, FE discretization error and QMC cubature error as follows

$$\begin{aligned}
\int_U q(\boldsymbol{y}, \boldsymbol{x}, z) \,\mathrm{d}\boldsymbol{y} - \frac{1}{n} \sum_{i=1}^{n} q_{s,h}(\boldsymbol{y}^{(i)}, \boldsymbol{x}, z) &= \underbrace{\int_U (q(\boldsymbol{y}, \boldsymbol{x}, z) - q_s(\boldsymbol{y}, \boldsymbol{x}, z)) \,\mathrm{d}\boldsymbol{y}}_{\text{truncation error}} \\
&\quad + \underbrace{\int_{U_s} (q_s(\boldsymbol{y}, \boldsymbol{x}, z) - q_{s,h}(\boldsymbol{y}, \boldsymbol{x}, z)) \,\mathrm{d}\boldsymbol{y}_{\leqslant s}}_{\text{FE discretization error}} \\
&\quad + \underbrace{\int_{U_s} q_{s,h}(\boldsymbol{y}, \boldsymbol{x}, z) \,\mathrm{d}\boldsymbol{y}_{\leqslant s} - \frac{1}{n} \sum_{i=1}^{n} q_{s,h}(\boldsymbol{y}^{(i)}, \boldsymbol{x}, z)}_{\text{QMC quadrature error}}\,.
\end{aligned} \tag{4.65}$$

These errors can be controlled as shown in Section 5.3, Section 7.1 and Section 7.2.2 below.

4.5.2 Parabolic PDE constraint

Let z^* denote the solution of (4.38) and let $z^*_{s,n}$ be the minimizer of

$$J_{s,n}(z) := \mathcal{R}_{s,n}(\Phi_s^{\boldsymbol{y}}(z)) + \frac{\alpha_3}{2}\|z\|^2_{L^2(V';I)},$$

where $\Phi_s^{\boldsymbol{y}}(z) = \Phi^{\boldsymbol{y}_{\leqslant s}}(z)$ is the truncated version of $\Phi^{\boldsymbol{y}}(z)$ defined in (4.37), and $\mathcal{R}_{s,n}$ is an approximation of the risk measure $\mathcal{R}$, for which the integrals over the parameter domain $U = [-\frac{1}{2}, \frac{1}{2}]^{\mathbb{N}}$ are replaced by s-dimensional integrals over $U_s = [-\frac{1}{2}, \frac{1}{2}]^s$ and then approximated by an n-point randomly-shifted QMC rule:

$$\mathcal{R}_{s,n}(\Phi_s^{\boldsymbol{y}}(z)) = \begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \Phi_s^{\boldsymbol{y}^{(i)}}(z) & \text{for the expected value,} \\[2ex] \dfrac{1}{\theta}\ln\Big(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \exp\big(\theta\, \Phi_s^{\boldsymbol{y}^{(i)}}(z)\big)\Big) & \text{for the entropic risk measure,} \end{cases}$$

for some risk aversion parameter $\theta \in (0, \infty)$, and for carefully chosen QMC points $\boldsymbol{y}^{(i)}$, $i = 1, \ldots, n$, involving a uniformly sampled random shift $\boldsymbol{\Delta} \in [0, 1]^s$, see Section 7.2.2.
We have seen in the proof of Corollary 4.2.5 that the risk measures considered in the example for the parabolic PDE constraint are convex and the objective function J, see (4.36), is thus strongly convex. It is important to note that the n-point QMC rule preserves the convexity of the risk measure, so $J_{s,n}$ is a strongly convex function, because it is a sum of a convex and a strongly convex function. Similar to the elliptic example, we therefore have the optimality conditions $\langle J'_{s,n}(z^*_{s,n}), z - z^*_{s,n}\rangle_{L^2(V;I),L^2(V';I)} \geqslant 0$ for all $z \in \mathcal{Z}_{\text{ad}}$ and thus in particular $\langle J'_{s,n}(z^*_{s,n}), z^* - z^*_{s,n}\rangle_{L^2(V;I),L^2(V';I)} \geqslant 0$. Similarly, we have $\langle J'(z^*), z - z^*\rangle_{L^2(V;I),L^2(V';I)} \geqslant 0$, and in particular $\langle -J'(z^*), z^* - z^*_{s,n}\rangle_{L^2(V;I),L^2(V';I)} \geqslant 0$. Adding these inequalities gives

$$\langle J'_{s,n}(z^*_{s,n}) - J'(z^*), z^* - z^*_{s,n}\rangle_{L^2(V;I),L^2(V';I)} \geqslant 0\,.$$

Hence

$$\begin{aligned}
&\alpha_3\|z^* - z^*_{s,n}\|^2_{L^2(V';I)} \\
&\leqslant \alpha_3\|z^* - z^*_{s,n}\|^2_{L^2(V';I)} + \langle J'_{s,n}(z^*_{s,n}) - J'(z^*), z^* - z^*_{s,n}\rangle_{L^2(V;I),L^2(V';I)} \\
&= \langle J'_{s,n}(z^*_{s,n}) - \alpha_3 R_V^{-1} z^*_{s,n} - J'(z^*) + \alpha_3 R_V^{-1} z^*, z^* - z^*_{s,n}\rangle_{L^2(V;I),L^2(V';I)} \\
&= \langle J'_{s,n}(z^*_{s,n}) - \alpha_3 R_V^{-1} z^*_{s,n} - J'_{s,n}(z^*) + \alpha_3 R_V^{-1} z^*, z^* - z^*_{s,n}\rangle_{L^2(V;I),L^2(V';I)} \\
&\quad + \langle J'_{s,n}(z^*) - \alpha_3 R_V^{-1} z^* - J'(z^*) + \alpha_3 R_V^{-1} z^*, z^* - z^*_{s,n}\rangle_{L^2(V;I),L^2(V';I)} \\
&\leqslant \langle J'_{s,n}(z^*) - \alpha_3 R_V^{-1} z^* - J'(z^*) + \alpha_3 R_V^{-1} z^*, z^* - z^*_{s,n}\rangle_{L^2(V;I),L^2(V';I)} \\
&\leqslant \|J'_{s,n}(z^*) - \alpha_3 R_V^{-1} z^* - J'(z^*) + \alpha_3 R_V^{-1} z^*\|_{L^2(V;I)}\|z^* - z^*_{s,n}\|_{L^2(V';I)}\,,
\end{aligned}$$

where we used the α_3-strong convexity of $J'_{s,n}$ in the fourth step, i.e.,

$$\langle J'_{s,n}(z^*_{s,n}) - J'_{s,n}(z^*) - \alpha_3 R_V^{-1}(z^* - z^*_{s,n}), z^* - z^*_{s,n}\rangle_{L^2(V;I),L^2(V';I)} \leqslant 0\,.$$

It follows

$$\|z^* - z^*_{s,n}\|_{L^2(V';I)} \leqslant \frac{1}{\alpha_3}\|J'(z^*) - J'_{s,n}(z^*)\|_{L^2(V;I)}.$$

We will next expand this upper bound in order to split it into the different error contributions: dimension truncation error and QMC error. The different error contributions are then analyzed separately in the Chapter 5 and Chapter 6 for both risk measures.
In the case of the expected value, it follows from (4.41) that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Delta}}\|z^* - z^*_{s,n}\|^2_{L^2(V';I)} &\leqslant \frac{1}{\alpha_3^2}\mathbb{E}_{\boldsymbol{\Delta}}\Big\| \int_U q_1^{\boldsymbol{y}}\,\mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^n q_{1,s}^{\boldsymbol{y}^{(i)}}\Big\|^2_{L^2(V;I)}\\ &\leqslant \frac{2}{\alpha_3^2}\Big\| \int_U (q_1^{\boldsymbol{y}} - q_{1,s}^{\boldsymbol{y}})\,\mathrm{d}\boldsymbol{y}\Big\|^2_{L^2(V;I)} + \frac{2}{\alpha_3^2}\mathbb{E}_{\boldsymbol{\Delta}}\Big\| \int_{U_s} q_{1,s}^{\boldsymbol{y}}\,\mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^n q_{1,s}^{\boldsymbol{y}^{(i)}}\Big\|^2_{L^2(V;I)}, \end{aligned} \tag{4.66}$$

where $q_{1,s}^{\boldsymbol{y}} := q_1^{\boldsymbol{y}_{\leqslant s}}$ denotes the truncated version of $q_1^{\boldsymbol{y}}$, and $\mathbb{E}_{\boldsymbol{\Delta}}$ denotes the expected value with respect to the random shift $\boldsymbol{\Delta} \in [0,1]^s$.
In the case of the entropic risk measure, we recall that $J'(z)$ is given by (4.48). Let

$$\begin{aligned} T &:= \int_U \exp\big(\theta\,\Phi^{\boldsymbol{y}}(z^*)\big)\,\mathrm{d}\boldsymbol{y}\,, \qquad\qquad T_{s,n} := \frac{1}{n}\sum_{i=1}^n \exp\big(\theta\,\Phi_s^{\boldsymbol{y}^{(i)}}(z^*_{s,n})\big)\,,\\ S &:= \int_U \exp\big(\theta\,\Phi^{\boldsymbol{y}}(z^*)\big)\,q_1^{\boldsymbol{y}}(z^*)\,\mathrm{d}\boldsymbol{y}\,,\ S_{s,n} := \frac{1}{n}\sum_{i=1}^n \exp\big(\theta\,\Phi_s^{\boldsymbol{y}^{(i)}}(z^*_{s,n})\big)\,q_{1,s}^{\boldsymbol{y}^{(i)}}(z^*_{s,n}), \end{aligned}$$

then we have

$$\begin{aligned}\alpha_3\|z^* - z^*_{s,n}\|_{L^2(V';I)} &\leqslant \Big\|\frac{S}{T} - \frac{S_{s,n}}{T_{s,n}}\Big\|_{L^2(V;I)} = \frac{\|S\,T_{s,n} - S_{s,n}\,T\|_{L^2(V;I)}}{T\,T_{s,n}}\\ &= \frac{\|S\,T_{s,n} - S\,T + S\,T - S_{s,n}\,T\|_{L^2(V;I)}}{T\,T_{s,n}}\\ &\leqslant \frac{\|S\|_{L^2(V;I)}\,|T - T_{s,n}|}{T\,T_{s,n}} + \frac{\|S - S_{s,n}\|_{L^2(V;I)}}{T_{s,n}}\\ &\leqslant \mu\,|T - T_{s,n}| + \|S - S_{s,n}\|_{L^2(V;I)}, \end{aligned} \tag{4.67}$$

where we used $T \geqslant 1$ and $T_{s,n} \geqslant 1$ and, using the abbreviation $g^{\boldsymbol{y}}(\boldsymbol{x},t) := \exp(\theta\,\Phi^{\boldsymbol{y}}(z))\,q_1^{\boldsymbol{y}}(\boldsymbol{x},t)$ we get

$$\begin{aligned}\|S\|^2_{L^2(V;I)} &= \int_I \Big\| \int_U g^{\boldsymbol{y}}(\cdot,t)\,\mathrm{d}\boldsymbol{y}\Big\|_V^2\,\mathrm{d}t = \int_I\int_D \Big|\nabla\Big(\int_U g^{\boldsymbol{y}}(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{y}\Big)\Big|^2\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\\ &\leqslant \int_U\int_I\int_D \big|\nabla g^{\boldsymbol{y}}(\boldsymbol{x},t)\big|^2\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t\,\mathrm{d}\boldsymbol{y} = \int_U \|g^{\boldsymbol{y}}\|^2_{L^2(V;I)}\,\mathrm{d}\boldsymbol{y} \leqslant \mu^2\,, \end{aligned}$$

where we used Theorem 4.6.12 with $\boldsymbol{\nu} = \mathbf{0}$.
We can write

$$\mathbb{E}_{\boldsymbol{\Delta}}\Big\|\frac{S}{T} - \frac{S_{s,n}}{T_{s,n}}\Big\|^2_{L^2(V;I)} \leqslant 2\mu^2\,\mathbb{E}_{\boldsymbol{\Delta}}\big|T - T_{s,n}\big|^2 + 2\mathbb{E}_{\boldsymbol{\Delta}}\|S - S_{s,n}\|^2_{L^2(V;I)}. \tag{4.68}$$

For the first term on the right-hand side of (4.68) we obtain

$$\mathbb{E}_{\boldsymbol{\Delta}}\big|T - T_{s,n}\big|^2 \leqslant 2\big|T - T_s\big|^2 + 2\mathbb{E}_{\boldsymbol{\Delta}}\big|T_s - T_{s,n}\big|^2, \tag{4.69}$$

and for the second term we have

$$\mathbb{E}_{\boldsymbol{\Delta}}\|S - S_{s,n}\|^2_{L^2(V;I)} \leqslant 2\|S - S_s\|^2_{L^2(V;I)} + 2\mathbb{E}_{\boldsymbol{\Delta}}\|S_s - S_{s,n}\|^2_{L^2(V;I)}. \tag{4.70}$$

Remark 4.5.2. *Since we have* $\|v_1\|_{L^2(V;I)} \leqslant \|v\|_{\mathcal{Y}}$ *for all* $v = (v_1, v_2) \in \mathcal{Y}$ *by definition, and thus in particular* $\|\int_U (q_1^{\boldsymbol{y}} - q_{1,s}^{\boldsymbol{y}})\,\mathrm{d}\boldsymbol{y}\|_{L^2(V;I)} \leqslant \|\int_U (q^{\boldsymbol{y}} - q_s^{\boldsymbol{y}})\,\mathrm{d}\boldsymbol{y}\|_{\mathcal{Y}}$*, we can replace* $q_1^{\boldsymbol{y}}, q_{1,s}^{\boldsymbol{y}} \in L^2(V;I)$ *in* (4.66) *and* (4.70) *by* $q^{\boldsymbol{y}}, q_s^{\boldsymbol{y}} \in \mathcal{Y}$*. In order to obtain error bounds and convergence rates for* (4.66) *and* (4.70)*, it is then sufficient to derive the results in the* $\mathcal{Y}$*-norm, which is slightly stronger than the* $L^2(V;I)$*-norm.*

The errors can be controlled as shown in Section 5.3 and Section 7.2.2.
For the parabolic PDE constraint and the parametric operator equation constraints, we leave out the spatial discretization and instead analyze the remaining error contributions directly in the respective function space. This technique has the advantage that, in the presented setting, the derived error bounds and convergence results immediately carry over to finite element discretizations belonging to conforming finite element subspaces of the respective function space.

4.5.3 Parametric linear operator constraints

Let z^* denote the solution of (4.52) and let $z_{s,n}^*$ be the minimizer of

$$J_{s,n}(z) := \frac{1}{2}\mathcal{R}_{s,n}\Big(\underbrace{\|\mathcal{Q}A^{-1}(\boldsymbol{y}^{(i)})\mathcal{B}z - \widehat{u}\|_{\mathfrak{Z}}^2}_{=:\widetilde{\Phi}_s^{\boldsymbol{y}^{(i)}}(z)}\Big) + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2$$

where, analogously to the setting with the parabolic PDE constraint, $\mathcal{R}_{s,n}$ is an approximation of the risk measure $\mathcal{R}$, for which the integrals over the parameter domain $U = [-\frac{1}{2}, \frac{1}{2}]^{\mathbb{N}}$ are replaced by s-dimensional integrals over $U_s = [-\frac{1}{2}, \frac{1}{2}]^s$ and then approximated by an n-point randomly-shifted QMC rule:

$$\mathcal{R}_{s,n}(\Phi_s^{\boldsymbol{y}}(z)) = \begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \widetilde{\Phi}_s^{\boldsymbol{y}^{(i)}}(z) & \text{for the expected value,} \\ \dfrac{1}{\theta}\ln\Big(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \exp\big(\theta\,\widetilde{\Phi}_s^{\boldsymbol{y}^{(i)}}(z)\big)\Big) & \text{for the entropic risk measure,} \end{cases}$$

for some risk aversion parameter $\theta \in (0, \infty)$, and for carefully chosen QMC points $\boldsymbol{y}^{(i)}$, $i = 1, \ldots, n$, involving a uniformly sampled random shift $\boldsymbol{\Delta} \in [0,1]^s$, see Section 7.2.2.
Since the two considered risk measured are convex and thus the objective function J, see (4.52), is strongly convex. The n-point QMC rule preserves the convexity of the risk measure, so $J_{s,n}$ is a strongly convex function. Analogously to the elliptic and the parabolic examples, we therefore have the optimality conditions $\langle J_{s,n}'(z_{s,n}^*), z - z_{s,n}^*\rangle_{\mathcal{Z}',\mathcal{Z}} \geqslant 0$ for all $z \in \mathcal{Z}_{\mathrm{ad}}$ and thus in particular $\langle J_{s,n}'(z_{s,n}^*), z^* - z_{s,n}^*\rangle_{\mathcal{Z}',\mathcal{Z}} \geqslant 0$. Similarly, we have $\langle J'(z^*), z - z^*\rangle_{\mathcal{Z}',\mathcal{Z}} \geqslant 0$, and in particular $\langle -J'(z^*), z^* - z_{s,n}^*\rangle_{\mathcal{Z}',\mathcal{Z}} \geqslant 0$. Adding these inequalities gives

$$\langle J_{s,n}'(z_{s,n}^*) - J'(z^*), z^* - z_{s,n}^*\rangle_{\mathcal{Z}',\mathcal{Z}} \geqslant 0\,.$$

Hence, by the same arguments as in the elliptic and parabolic examples, we obtain

$$\|z^* - z_{s,n}^*\|_{\mathcal{Z}} \leqslant \frac{1}{\alpha}\|J'(z^*) - J_{s,n}'(z^*)\|_{\mathcal{Z}'}.$$

In order to analyzed the error contributions separately in Section 5.3 and Section 7.2.2 we will expand this upper bound for both risk measures.

In the case of the expected value, it follows from (4.53) that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Delta}}\|z^* - z^*_{s,n}\|^2_{\mathcal{Z}} &\leqslant \frac{1}{\alpha^2}\mathbb{E}_{\boldsymbol{\Delta}}\Big\| \int_U \mathcal{B}' q^{\boldsymbol{y}} \,\mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^n \mathcal{B}' q_s^{\boldsymbol{y}^{(i)}} \Big\|^2_{\mathcal{Z}} \leqslant \frac{C^2_{\mathcal{B}'}}{\alpha^2}\mathbb{E}_{\boldsymbol{\Delta}}\Big\| \int_U q^{\boldsymbol{y}} \,\mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^n q_s^{\boldsymbol{y}^{(i)}} \Big\|^2_{\mathcal{Z}} \\ &\leqslant \frac{2C^2_{\mathcal{B}}}{\alpha^2}\Big\| \int_U (q^{\boldsymbol{y}} - q_s^{\boldsymbol{y}}) \,\mathrm{d}\boldsymbol{y}\Big\|^2_{\mathcal{Y}} + \frac{2C^2_{\mathcal{B}}}{\alpha^2}\mathbb{E}_{\boldsymbol{\Delta}}\Big\| \int_{U_s} q_s^{\boldsymbol{y}} \,\mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^n q_s^{\boldsymbol{y}^{(i)}} \Big\|^2_{\mathcal{Y}}, \end{aligned} \tag{4.71}$$

where $C = \|\mathcal{B}'\|_{\mathcal{L}(\mathcal{Y},\mathcal{Z}')}$ and $q_s^{\boldsymbol{y}} := q^{\boldsymbol{y}_{\leqslant s}}$ denotes the truncated version of $q^{\boldsymbol{y}}$, defined in (4.54), and $\mathbb{E}_{\boldsymbol{\Delta}}$ denotes the expected value with respect to the random shift $\boldsymbol{\Delta} \in [0,1]^s$. In the case of the entropic risk measure, we recall that $J'(z)$ is given by (4.57). Let

$$\begin{aligned} T &:= \int_U \exp\big(\theta\, \widetilde{\Phi}^{\boldsymbol{y}}(z^*)\big) \,\mathrm{d}\boldsymbol{y}\,, \qquad\qquad T_{s,n} := \frac{1}{n}\sum_{i=1}^n \exp\big(\theta\, \widetilde{\Phi}_s^{\boldsymbol{y}^{(i)}}(z^*_{s,n})\big)\,, \\ S &:= \int_U \exp\big(\theta\, \widetilde{\Phi}^{\boldsymbol{y}}(z^*)\big)\, \mathcal{B}' q^{\boldsymbol{y}}(z^*) \,\mathrm{d}\boldsymbol{y}\,,\ S_{s,n} := \frac{1}{n}\sum_{i=1}^n \exp\big(\theta\, \widetilde{\Phi}_s^{\boldsymbol{y}^{(i)}}(z^*_{s,n})\big)\, \mathcal{B}' q_s^{\boldsymbol{y}^{(i)}}(z^*_{s,n}), \end{aligned}$$

then we have

$$\begin{aligned} \big\|z^* - z^*_{s,n}\big\|_{L^2(V';I)} &\leqslant \Big\|\frac{S}{T} - \frac{S_{s,n}}{T_{s,n}}\Big\|_{\mathcal{Z}'} = \frac{\big\|S\,T_{s,n} - S_{s,n}\,T\big\|_{\mathcal{Z}'}}{T\,T_{s,n}} \\ &= \frac{\big\|S\,T_{s,n} - S\,T + S\,T - S_{s,n}\,T\big\|_{\mathcal{Z}'}}{T\,T_{s,n}} \\ &\leqslant \frac{\|S\|_{\mathcal{Z}'}\big|T - T_{s,n}\big|}{T\,T_{s,n}} + \frac{\big\|S - S_{s,n}\big\|_{\mathcal{Z}'}}{T_{s,n}} \\ &\leqslant \mu\big|T - T_{s,n}\big| + \big\|S - S_{s,n}\big\|_{\mathcal{Z}'}, \end{aligned} \tag{4.72}$$

where we used $T \geqslant 1$ and $T_{s,n} \geqslant 1$ and, using the abbreviation $g^{\boldsymbol{y}}(\boldsymbol{x},t) := \exp(\theta\,\Phi^{\boldsymbol{y}}(z))\,\mathcal{B}' q^{\boldsymbol{y}}(\boldsymbol{x},t)$ we get

$$\begin{aligned} \|S\|^2_{\mathcal{Z}'} \;&=\; \Big\| \int_U \exp\big(\theta\, \widetilde{\Phi}^{\boldsymbol{y}}(z^*)\big)\, \mathcal{B}' q^{\boldsymbol{y}}(z^*) \,\mathrm{d}\boldsymbol{y}\Big\|_{\mathcal{Z}'} \;=\; \int_U \Big\| \exp\big(\theta\, \widetilde{\Phi}^{\boldsymbol{y}}(z^*)\big)\, \mathcal{B}' q^{\boldsymbol{y}}(z^*)\Big\|_{\mathcal{Z}'} \,\mathrm{d}\boldsymbol{y} \\ &= \int_U \big\|g^{\boldsymbol{y}}\big\|^2_{\mathcal{Z}'} \,\mathrm{d}\boldsymbol{y} \;\leqslant\; \mu^2\,, \end{aligned}$$

where we used Theorem 4.6.19 with $\boldsymbol{\nu} = \mathbf{0}$.
We can write

$$\mathbb{E}_{\boldsymbol{\Delta}}\Big\|\frac{S}{T} - \frac{S_{s,n}}{T_{s,n}}\Big\|^2_{\mathcal{Z}'} \;\leqslant\; 2\mu^2\,\mathbb{E}_{\boldsymbol{\Delta}}\big|T - T_{s,n}\big|^2 + 2\mathbb{E}_{\boldsymbol{\Delta}}\big\|S - S_{s,n}\big\|^2_{\mathcal{Z}'}. \tag{4.73}$$

For the first term on the right-hand side of (4.73) we obtain

$$\mathbb{E}_{\boldsymbol{\Delta}}\big|T - T_{s,n}\big|^2 \;\leqslant\; 2\big|T - T_s\big|^2 + 2\mathbb{E}_{\boldsymbol{\Delta}}\big|T_s - T_{s,n}\big|^2, \tag{4.74}$$

and for the second term we have

$$\mathbb{E}_{\boldsymbol{\Delta}}\|S - S_{s,n}\|^2_{\mathcal{Z}'} \;\leqslant\; 2\|S - S_s\|^2_{\mathcal{Z}'} + 2\mathbb{E}_{\boldsymbol{\Delta}}\|S_s - S_{s,n}\|^2_{\mathcal{Z}'}. \tag{4.75}$$

The different errors can be controlled as shown in Section 5.3 and Section 7.2.2. Recall that $\mathcal{B}'$ is a bounded linear operator and can be pulled out before the error estimation, i.e.,

$$\begin{aligned}\|S - S_{s,n}\|_{\mathcal{Z}'} &= \left\| \int_U \exp\left(\theta\, \widetilde{\Phi}^{\boldsymbol{y}}(z^*)\right) \mathcal{B}' q^{\boldsymbol{y}}(z^*)\, \mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^{n} \exp\left(\theta\, \widetilde{\Phi}_s^{\boldsymbol{y}^{(i)}}(z_{s,n}^*)\right) \mathcal{B}' q_s^{\boldsymbol{y}^{(i)}}(z_{s,n}^*) \right\|_{\mathcal{Z}'} \\ &\leqslant C_{\mathcal{B}} \left\| \int_U \exp\left(\theta\, \widetilde{\Phi}^{\boldsymbol{y}}(z^*)\right) q^{\boldsymbol{y}}(z^*)\, \mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^{n} \exp\left(\theta\, \widetilde{\Phi}_s^{\boldsymbol{y}^{(i)}}(z_{s,n}^*)\right) q_s^{\boldsymbol{y}^{(i)}}(z_{s,n}^*) \right\|_{\mathcal{Y}}.\end{aligned}$$

4.6 Regularity analysis

In the previous section we have seen that the error in the optimal control can be bounded by the error in the derivative of the objective functional. This derivative typically involves a function of the adjoint or dual PDE solution. By a simple application of the triangle inequality, the overall error can then be decomposed into its different contributions. The error bounds and convergence rates for the different error contributions rely fundamentally on the parametric regularity of the quantity of interest. In preparation for the application of the theoretical results derived in Chapter 5 and Chapter 6 to the optimal control problems described in Chapter 4, we investigate the parametric regularity of the integrands in the bounds of the errors in the optimal control appearing in (4.64), (4.66), (4.67), (4.71), and (4.72).

We start this section with recalling some well-known and frequently used results. An important result for the differentiation is Leibniz generalized product rule: Let $\boldsymbol{\nu}, \boldsymbol{m} \in \mathcal{F}$. Then it holds for sufficiently regular functions f, g, that

$$\partial^{\boldsymbol{\nu}}(fg) = \sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \partial^{\boldsymbol{m}} f \partial^{\boldsymbol{\nu}-\boldsymbol{m}} g\,.$$

Another frequently used result is the following recursive bound.

Lemma 4.6.1 ([119, Lemma 5])**.** *Given a sequence of nonnegative numbers $\boldsymbol{b} = (b_j)_{j\in\mathbb{N}_0}$, let $(\mathbb{A}_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ and $(\mathbb{B}_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ be nonnegative numbers satisfying for any $\boldsymbol{\nu} \in \mathcal{F}$ the inequality*

$$\mathbb{A}_{\boldsymbol{\nu}} \leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}, \boldsymbol{m}\neq\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{m}} \mathbb{A}_{\boldsymbol{m}} + \mathbb{B}_{\boldsymbol{\nu}}\,.$$

Then

$$\mathbb{A}_{\boldsymbol{\nu}} \leqslant \sum_{\boldsymbol{k}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{k}} \frac{|\boldsymbol{k}|!}{(\ln 2)^{|\boldsymbol{k}|}} \boldsymbol{b}^{\boldsymbol{k}} \mathbb{B}_{\boldsymbol{\nu}-\boldsymbol{k}}$$

for all $\boldsymbol{\nu} \in \mathcal{F}$.

In the regularity analysis we will frequently use a number of combinatorial identities, which are listed below. As stated in [113, equation (9.3)] the identity

$$\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}, |\boldsymbol{m}|=i} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} = \binom{|\boldsymbol{\nu}|}{i} = \frac{|\boldsymbol{\nu}|!}{i!(|\boldsymbol{\nu}|-i)!}. \tag{4.76}$$

follows from considering the number of ways to pick i objects from a set of bags containing in total $|\boldsymbol{\nu}|$ objects. It then follows that

$$\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} |\boldsymbol{m}|!|\boldsymbol{\nu}-\boldsymbol{m}|! = \sum_{i=0}^{|\boldsymbol{\nu}|} \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},|\boldsymbol{m}|=i} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} i!(|\boldsymbol{\nu}|-i)! = \sum_{i=0}^{|\boldsymbol{\nu}|} |\boldsymbol{\nu}|! = (|\boldsymbol{\nu}|+1)!, \quad (4.77)$$

as can be found in [113, equation (9.4)]. Moreover, it follows that

$$\begin{aligned}
\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \frac{|\boldsymbol{m}|!}{(\ln 2)^{|\boldsymbol{m}|}} &= \sum_{i=0}^{|\boldsymbol{\nu}|} \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},|m|=i} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \frac{i!}{(\ln 2)^i} = \sum_{i=0}^{|\boldsymbol{\nu}|} \frac{|\boldsymbol{\nu}|!}{(|\boldsymbol{\nu}|-i)!} \frac{1}{(\ln 2)^i} \\
&= |\boldsymbol{\nu}|! \left(\frac{1}{0!(\ln 2)^{|\boldsymbol{\nu}|}} + \frac{1}{1!(\ln 2)^{|\boldsymbol{\nu}|-1}} + \ldots + \frac{1}{|\boldsymbol{\nu}|!(\ln 2)^0} \right) \\
&\leqslant \frac{|\boldsymbol{\nu}|!}{(\ln 2)^{|\boldsymbol{\nu}|}} \left(\frac{1}{0!(\ln 2)^0} + \frac{1}{1!(\ln 2)^{-1}} + \ldots + \frac{1}{|\boldsymbol{\nu}|!(\ln 2)^{-|\boldsymbol{\nu}|}} \right) \\
&\leqslant \frac{|\boldsymbol{\nu}|!}{(\ln 2)^{|\boldsymbol{\nu}|}} e^{\ln 2} = 2\frac{|\boldsymbol{\nu}|!}{(\ln 2)^{|\boldsymbol{\nu}|}} \qquad (4.78)
\end{aligned}$$

and

$$\begin{aligned}
\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},\boldsymbol{m}\neq\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \frac{(|\boldsymbol{m}|+1)!}{(\ln 2)^{|\boldsymbol{m}|}} &= \sum_{i=0}^{|\boldsymbol{\nu}|-1} \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},|\boldsymbol{m}|=i} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \frac{(i+1)!}{(\ln 2)^i} \\
&= |\boldsymbol{\nu}|! \sum_{i=0}^{|\boldsymbol{\nu}|-1} \frac{i+1}{(|\boldsymbol{\nu}|-i)!(\ln 2)^i} \\
&= |\boldsymbol{\nu}|! \left(\frac{|\boldsymbol{\nu}|}{1!(\ln 2)^{|\boldsymbol{\nu}|-1}} + \frac{|\boldsymbol{\nu}|-1}{2!(\ln 2)^{|\boldsymbol{\nu}|-2}} + \ldots + \frac{1}{|\boldsymbol{\nu}|!(\ln 2)^0} \right) \\
&\leqslant \frac{|\boldsymbol{\nu}|!}{(\ln 2)^{|\boldsymbol{\nu}|}} \left(\frac{|\boldsymbol{\nu}|}{1!(\ln 2)^{-1}} + \frac{|\boldsymbol{\nu}|}{2!(\ln 2)^{-2}} + \ldots + \frac{|\boldsymbol{\nu}|}{|\boldsymbol{\nu}|!(\ln 2)^{-|\boldsymbol{\nu}|}} \right) \\
&\leqslant \frac{|\boldsymbol{\nu}|!|\boldsymbol{\nu}|(e^{\ln 2}-1)}{(\ln 2)^{|\boldsymbol{\nu}|}} \leqslant \frac{(|\boldsymbol{\nu}|+1)!}{(\ln 2)^{|\boldsymbol{\nu}|}}, \qquad (4.79)
\end{aligned}$$

By adding the $\boldsymbol{m} = \boldsymbol{\nu}$ term on both sides of (4.79) we get

$$\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \frac{(|\boldsymbol{m}|+1)!}{(\ln 2)^{|\boldsymbol{m}|}} \leqslant 2\frac{(|\boldsymbol{\nu}|+1)!}{(\ln 2)^{|\boldsymbol{\nu}|}}. \quad (4.80)$$

4.6.1 Elliptic PDE

In this subsection we will derive bounds on the mixed first partial derivatives of the parametric solution u as well as bounds on the mixed first partial derivatives of the adjoint parametric solution q. For the solution $u(\cdot, \boldsymbol{y}, z)$ of the state equation (4.8) the following result is well-known.

Lemma 4.6.2. *For every $z \in V'$, every $\boldsymbol{y} \in U$ and every $\boldsymbol{\nu} \in \mathcal{D}$ we have*

$$\|(\partial^{\boldsymbol{\nu}} u)(\cdot, \boldsymbol{y}, z)\|_V := \|\nabla(\partial^{\boldsymbol{\nu}} u)(\cdot, \boldsymbol{y}, z)\|_{L^2(D)} \leqslant |\boldsymbol{\nu}|!\, \boldsymbol{b}^{\boldsymbol{\nu}} \frac{\|z\|_{V'}}{a_{\min}}.$$

This lemma can be found, e.g., in [31].
In contrast to the parametric weak problem (4.8), the right-hand side of the adjoint parametric weak problem (4.17) depends on the parameters $\boldsymbol{y} \in U$. In particular, the problem is of the following form: for every $\boldsymbol{y} \in U$, find $q(\cdot, \boldsymbol{y}, z) \in V$ such that

$$\int_D a(\boldsymbol{x}, \boldsymbol{y}) \nabla q(\boldsymbol{x}, \boldsymbol{y}, z) \cdot \nabla v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int_D \tilde{f}(\boldsymbol{x}, \boldsymbol{y}, z) v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \,, \quad v \in V \,, \tag{4.81}$$

where the right-hand side $\tilde{f}(\boldsymbol{x}, \boldsymbol{y}, z) := u(\boldsymbol{x}, \boldsymbol{y}, z) - \widehat{u}(\boldsymbol{x})$ depends on $z \in L^2(D)$ and $\boldsymbol{y} \in U$. Lemma 4.6.3 below gives a bound for the mixed derivatives of the solution $q(\cdot, \boldsymbol{y}, z) \in V$ of (4.81). Similar regularity results to the following can be found in [111] (uniform case) and [24] (log-normal case) for problems with stochastic controls z, depending on $\boldsymbol{y}$. In particular, in the unconstrained case $\mathcal{Z}_{\mathrm{ad}} = L^2(D)$ the optimality conditions in Section 4.1.4 reduce to an affine parametric linear saddle point operator and the theory, e.g., from [111, 146] can be applied.

Lemma 4.6.3. *For every $z \in L^2(D)$, every $\boldsymbol{y} \in U$ and every $\boldsymbol{\nu} \in \mathcal{F}$, we have for the corresponding adjoint state $q(\cdot, \boldsymbol{y}, z)$ that*

$$\|(\partial^{\boldsymbol{\nu}} q)(\cdot, \boldsymbol{y}, z)\|_V \leqslant (|\boldsymbol{\nu}| + 1)! \, \boldsymbol{b}^{\boldsymbol{\nu}} \, C_q \left(\|z\|_{L^2(D)} + \|\widehat{u}\|_{L^2(D)} \right),$$

where $C_q := \max\{\frac{c_1}{a_{\min}}, \frac{c_1^2 c_2}{a_{\min}^2}\}$.

Proof. The case $\boldsymbol{\nu} = \boldsymbol{0}$ follows from the uniform bounded invertibility of $S_{\boldsymbol{y}}$

$$\|q(\cdot, \boldsymbol{y}, z)\|_V \leqslant C_q (\|z\|_{L^2(D)} + \|\widehat{u}\|_{L^2(D)}). \tag{4.82}$$

Now consider $\boldsymbol{\nu} \neq \boldsymbol{0}$. Applying the mixed derivative operator $\partial^{\boldsymbol{\nu}}$ to (4.81) and using the Leibniz product rule, we obtain the identity

$$\begin{aligned} \int_D \left(\sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} (\partial^{\boldsymbol{m}} a)(\boldsymbol{x}, \boldsymbol{y}) \nabla (\partial^{\boldsymbol{\nu} - \boldsymbol{m}} q)(\boldsymbol{x}, \boldsymbol{y}, z) \cdot \nabla v(\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{x} \\ = \int_D (\partial^{\boldsymbol{\nu}} \tilde{f})(\boldsymbol{x}, \boldsymbol{y}, z) \; v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \quad \forall v \in V \,. \end{aligned} \tag{4.83}$$

Due to the linear dependence of $a(\boldsymbol{x}, \boldsymbol{y})$ on the parameters $\boldsymbol{y}$, the partial derivative $\partial^{\boldsymbol{m}}$ of a with respect to $\boldsymbol{y}$ satisfies

$$(\partial^{\boldsymbol{m}} a)(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} a(\boldsymbol{x}, \boldsymbol{y}) & \text{if } \boldsymbol{m} = \boldsymbol{0} \,, \\ \psi_j(\boldsymbol{x}) & \text{if } \boldsymbol{m} = \boldsymbol{e}_j \,, \\ 0 & \text{else} \,. \end{cases}$$

Setting $v = (\partial^{\boldsymbol{\nu}} q)(\cdot, \boldsymbol{y}, z)$ and separating out the $\boldsymbol{m} = \boldsymbol{0}$ term, we obtain

$$\begin{aligned} &\int_D a(\boldsymbol{x}, \boldsymbol{y}) |\nabla (\partial^{\boldsymbol{\nu}} q)(\boldsymbol{x}, \boldsymbol{y}, z)|^2 \, \mathrm{d}\boldsymbol{x} \\ &\quad = - \sum_{j \in \mathrm{supp}(\boldsymbol{\nu})} \nu_j \int_D \psi_j(\boldsymbol{x}) \nabla (\partial^{\boldsymbol{\nu} - \boldsymbol{e}_j} q)(\boldsymbol{x}, \boldsymbol{y}, z) \cdot \nabla (\partial^{\boldsymbol{\nu}} q)(\boldsymbol{x}, \boldsymbol{y}, z) \, \mathrm{d}\boldsymbol{x} \\ &\quad + \int_D (\partial^{\boldsymbol{\nu}} \tilde{f})(\boldsymbol{x}, \boldsymbol{y}, z) (\partial^{\boldsymbol{\nu}} q)(\boldsymbol{x}, \boldsymbol{y}, z) \, \mathrm{d}\boldsymbol{x} \,, \end{aligned}$$

which yields

$$\begin{aligned} a_{\min}\|(\partial^{\boldsymbol{\nu}} q)(\cdot,\boldsymbol{y},z)\|_V^2 &\leqslant \sum_{j\geqslant 1}\nu_j\|\psi_j\|_{L^\infty(D)}\|(\partial^{\boldsymbol{\nu}-\boldsymbol{e}_j} q)(\cdot,\boldsymbol{y},z)\|_V\|(\partial^{\boldsymbol{\nu}} q)(\cdot,\boldsymbol{y},z)\|_V \\ &\quad + \|(\partial^{\boldsymbol{\nu}}\tilde{f})(\cdot,\boldsymbol{y},z)\|_{V'}\|(\partial^{\boldsymbol{\nu}} q)(\cdot,\boldsymbol{y},z)\|_V \\ &= \sum_{j\geqslant 1}\nu_j\|\psi_j\|_{L^\infty(D)}\|(\partial^{\boldsymbol{\nu}-\boldsymbol{e}_j} q)(\cdot,\boldsymbol{y},z)\|_V\|(\partial^{\boldsymbol{\nu}} q)(\cdot,\boldsymbol{y},z)\|_V \\ &\quad + \|(\partial^{\boldsymbol{\nu}}\tilde{f})(\cdot,\boldsymbol{y},z)\|_{V'}\|(\partial^{\boldsymbol{\nu}} q)(\cdot,\boldsymbol{y},z)\|_V\,, \end{aligned}$$

and hence

$$\|(\partial^{\boldsymbol{\nu}} q)(\cdot,\boldsymbol{y},z)\|_V \leqslant \sum_{j\geqslant 1}\nu_j b_j\|(\partial^{\boldsymbol{\nu}-\boldsymbol{e}_j} q)(\cdot,\boldsymbol{y},z)\|_V + \frac{\|(\partial^{\boldsymbol{\nu}}\tilde{f})(\cdot,\boldsymbol{y},z)\|_{V'}}{a_{\min}}\,,$$

where $b_j := \frac{\|\psi_j\|_{L^\infty(D)}}{a_{\min}}$ for $j\in\mathbb{N}$. With $\tilde{f}(\cdot,\boldsymbol{y},z) = u(\cdot,\boldsymbol{y},z) - \widehat{u}(\cdot)$ this reduces to

$$\|(\partial^{\boldsymbol{\nu}} q)(\cdot,\boldsymbol{y},z)\|_V \leqslant \sum_{j\geqslant 1}\nu_j b_j\|(\partial^{\boldsymbol{\nu}-\boldsymbol{e}_j} q)(\cdot,\boldsymbol{y},z)\|_V + \frac{\|(\partial^{\boldsymbol{\nu}} u)(\cdot,\boldsymbol{y},z)\|_{V'}}{a_{\min}}\,. \tag{4.84}$$

With Lemma 4.6.2 we get

$$\|(\partial^{\boldsymbol{\nu}} u)(\cdot,\boldsymbol{y},z)\|_{V'} \leqslant c_1c_2\|(\partial^{\boldsymbol{\nu}} u)(\cdot,\boldsymbol{y},z)\|_V \leqslant c_1c_2\,|\boldsymbol{\nu}|!\,\boldsymbol{b}^{\boldsymbol{\nu}}\frac{\|z\|_{V'}}{a_{\min}}\,,$$

where $c_1, c_2 > 0$ are embedding constants, see (4.6) and (4.7). Then (4.84) becomes, for $\boldsymbol{\nu}\neq\mathbf{0}$,

$$\|(\partial^{\boldsymbol{\nu}} q)(\cdot,\boldsymbol{y},z)\|_V \leqslant \sum_{j\geqslant 1}\nu_j b_j\|(\partial^{\boldsymbol{\nu}-\boldsymbol{e}_j} q)(\cdot,\boldsymbol{y},z)\|_V + c_1c_2\,|\boldsymbol{\nu}|!\,\boldsymbol{b}^{\boldsymbol{\nu}}\frac{\|z\|_{V'}}{a_{\min}^2}\,.$$

Now we apply Lemma 4.6.1 to obtain the final bound. For this to work we need the above recursion to hold also for the case $\boldsymbol{\nu}=\mathbf{0}$, which is not true when we compare it with the a-priori bound (4.82). We therefore enlarge the constants so that the recursion becomes

$$\|(\partial^{\boldsymbol{\nu}} q)(\cdot,\boldsymbol{y},z)\|_V \leqslant \sum_{j\geqslant 1}\nu_j b_j\|(\partial^{\boldsymbol{\nu}-\boldsymbol{e}_j} q)(\cdot,\boldsymbol{y},z)\|_V + |\boldsymbol{\nu}|!\,\boldsymbol{b}^{\boldsymbol{\nu}}\,C_q\,(\|z\|_{L^2(D)} + \|\widehat{u}\|_{L^2(D)})\,,$$

which by Lemma 4.6.1 gives

$$\begin{aligned} \|(\partial^{\boldsymbol{\nu}} q)(\cdot,\boldsymbol{y},z)\|_V &\leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}|\boldsymbol{m}|!\;\boldsymbol{b}^{\boldsymbol{m}}\,|\boldsymbol{\nu}-\boldsymbol{m}|!\;\boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{m}}\,C_q\,(\|z\|_{L^2(D)} + \|\widehat{u}\|_{L^2(D)}) \\ &= \boldsymbol{b}^{\boldsymbol{\nu}}\,C_q\,(\|z\|_{L^2(D)} + \|\widehat{u}\|_{L^2(D)})\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}|\boldsymbol{m}|!\;|\boldsymbol{\nu}-\boldsymbol{m}|! \\ &= \boldsymbol{b}^{\boldsymbol{\nu}}\,C_q\,(\|z\|_{L^2(D)} + \|\widehat{u}\|_{L^2(D)})\,(|\boldsymbol{\nu}|+1)!\,, \end{aligned}$$

where the last equality follows from (4.77). □

4.6.2 Parabolic PDE

The idea of the proofs in the regularity analysis of the parabolic PDE follow mainly the ideas of the proof of the elliptic PDE. A novelty is the regularity analysis of the solution of the adjoint state in conjunction with the entropic risk measure.
The following regularity result for the state $u^{\boldsymbol{y}}$ was proved in [111].

Lemma 4.6.4. *Let $f = (z, u_0) \in \mathcal{Y}'$. For all $\boldsymbol{\nu} \in \mathcal{F}$ and all $\boldsymbol{y} \in U$, we have*

$$\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} u^{\boldsymbol{y}}\|_{\mathcal{X}} \leqslant \frac{\|f\|_{\mathcal{Y}'}}{\beta_1} |\boldsymbol{\nu}|! \, \boldsymbol{b}^{\boldsymbol{\nu}}, \tag{4.85}$$

where β_1 is as described in (4.26) and the sequence $\boldsymbol{b} = (b_j)_{j \geqslant 1}$ is defined by

$$b_j := \frac{1}{\beta_1} \sup_{t \in I} \|\psi_j(\cdot, t)\|_{L^\infty(D)}. \tag{4.86}$$

Parametric regularity of the adjoint state

In this subsection we derive an a-priori bound for the adjoint state and the partial derivatives of the adjoint state with respect to the parametric variables. Existing results, e.g., [111, Theorem 4], do not directly apply to our case, since the right-hand side of the affine linear, parametric operator equation depends on the parametric variable, more specifically

$$(B^{\boldsymbol{y}})' q^{\boldsymbol{y}} = (\alpha_1 R_V + \alpha_2 E_T' E_T)(u^{\boldsymbol{y}} - \widehat{u}).$$

Lemma 4.6.5. *Let $\alpha_1, \alpha_2 \geqslant 0$ and $\alpha_3 > 0$, with $\alpha_1 + \alpha_2 > 0$. Let $f = (z, u_0) \in \mathcal{Y}'$ and $\widehat{u} \in \mathcal{X}$. For every $\boldsymbol{y} \in U$, let $u^{\boldsymbol{y}} \in \mathcal{X}$ be the solution of (4.20) and then let $q^{\boldsymbol{y}} \in \mathcal{Y}$ be the solution of (4.32) with f_{dual} given by (4.40). Then we have*

$$\|q^{\boldsymbol{y}}\|_{\mathcal{Y}} \leqslant \frac{\alpha_1 + \alpha_2 \|E_T\|^2_{\mathcal{X} \to L^2(D)}}{\beta_1} \left(\frac{\|f\|_{\mathcal{Y}'}}{\beta_1} + \|\widehat{u}\|_{\mathcal{X}} \right),$$

where β_1 is described in (4.26).

Proof. By the bounded invertibility of $B^{\boldsymbol{y}}$ and its dual operator, we have

$$\|q^{\boldsymbol{y}}\|_{\mathcal{Y}} \leqslant \|((B^{\boldsymbol{y}})')^{-1}\|_{\mathcal{X}' \to \mathcal{Y}} \, \|(\alpha_1 R_V + \alpha_2 E_T' E_T)(u^{\boldsymbol{y}} - \widehat{u})\|_{\mathcal{X}'},$$

with $\|((B^{\boldsymbol{y}})')^{-1}\|_{\mathcal{X}' \to \mathcal{Y}} \leqslant 1/\beta_1$,

$$\begin{aligned}
\|R_V(u^{\boldsymbol{y}} - \widehat{u})\|_{\mathcal{X}'} &\leqslant \|R_V(u^{\boldsymbol{y}} - \widehat{u})\|_{L^2(V';I)} = \|u^{\boldsymbol{y}} - \widehat{u}\|_{L^2(V;I)} \leqslant \|u^{\boldsymbol{y}} - \widehat{u}\|_{\mathcal{X}}, \\
\|E_T' E_T(u^{\boldsymbol{y}} - \widehat{u})\|_{\mathcal{X}'} &\leqslant \|E_T\|^2_{\mathcal{X} \to L^2(D)} \, \|u^{\boldsymbol{y}} - \widehat{u}\|_{\mathcal{X}}, \\
\|u^{\boldsymbol{y}} - \widehat{u}\|_{\mathcal{X}} &\leqslant \|u^{\boldsymbol{y}}\|_{\mathcal{X}} + \|\widehat{u}\|_{\mathcal{X}} \leqslant \frac{\|f\|_{\mathcal{Y}'}}{\beta_1} + \|\widehat{u}\|_{\mathcal{X}},
\end{aligned}$$

where we used (4.27). Combining the estimates gives the desired result. □

Theorem 4.6.6. *Let $\alpha_1, \alpha_2 \geqslant 0$ and $\alpha_3 > 0$, with $\alpha_1 + \alpha_2 > 0$. Let $f = (z, u_0) \in \mathcal{Y}'$ and $\widehat{u} \in \mathcal{X}$. For every $\boldsymbol{y} \in U$, let $u^{\boldsymbol{y}} \in \mathcal{X}$ be the solution of (4.20) and then let $q^{\boldsymbol{y}} \in \mathcal{Y}$ be the solution of (4.32) with f_{dual} given by (4.40). Then for every $\boldsymbol{\nu} \in \mathcal{F}$ we have*

$$\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} q^{\boldsymbol{y}}\|_{\mathcal{Y}} \leqslant \frac{\alpha_1 + \alpha_2 \|E_T\|^2_{\mathcal{X} \to L^2(D)}}{\beta_1} \left(\frac{\|f\|_{\mathcal{Y}'}}{\beta_1} + \|\widehat{u}\|_{\mathcal{X}} \right) (|\boldsymbol{\nu}| + 1)! \, \boldsymbol{b}^{\boldsymbol{\nu}},$$

where β_1 is described in (4.26) and the sequence $\boldsymbol{b} = (b_j)_{j \geqslant 1}$ is defined in (4.86).

Proof. For $\boldsymbol{\nu} = \mathbf{0}$ the assertion follows from the previous lemma. For $\boldsymbol{\nu} \neq \mathbf{0}$ we take derivatives $\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}((B^{\boldsymbol{y}})' q^{\boldsymbol{y}}) = \partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}((\alpha_1 R_V + \alpha_2 E_T' E_T)(u^{\boldsymbol{y}} - \widehat{u}))$ and use the Leibniz product rule to get

$$\sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \big(\partial_{\boldsymbol{y}}^{\boldsymbol{m}} (B^{\boldsymbol{y}})'\big)\big(\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}-\boldsymbol{m}} q^{\boldsymbol{y}}\big) = (\alpha_1 R_V + \alpha_2 E_T' E_T)\big(\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}(u^{\boldsymbol{y}} - \widehat{u})\big)\,.$$

Separating out the $\boldsymbol{m} = \mathbf{0}$ term, we obtain

$$\begin{aligned}
&(B^{\boldsymbol{y}})'(\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} q^{\boldsymbol{y}}) \\
&= -\sum_{\mathbf{0} \neq \boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \big(\partial_{\boldsymbol{y}}^{\boldsymbol{m}} (B^{\boldsymbol{y}})'\big)\big(\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}-\boldsymbol{m}} q^{\boldsymbol{y}}\big) + (\alpha_1 R_V + \alpha_2 E_T' E_T)\big(\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}(u^{\boldsymbol{y}} - \widehat{u})\big).
\end{aligned}$$

By the bounded invertibility of $(B^{\boldsymbol{y}})'$, we have $\|((B^{\boldsymbol{y}})')^{-1}\|_{\mathcal{X}' \to \mathcal{Y}} \leqslant \frac{1}{\beta_1}$ and

$$\begin{aligned}
\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} q^{\boldsymbol{y}}\|_{\mathcal{Y}} &\leqslant \sum_{\mathbf{0} \neq \boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \|((B^{\boldsymbol{y}})')^{-1} \partial_{\boldsymbol{y}}^{\boldsymbol{m}} (B^{\boldsymbol{y}})'\|_{\mathcal{Y} \to \mathcal{Y}}\, \|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}-\boldsymbol{m}} q^{\boldsymbol{y}}\|_{\mathcal{Y}} \\
&\quad + \|((B^{\boldsymbol{y}})')^{-1}\|_{\mathcal{X}' \to \mathcal{Y}}\, \|(\alpha_1 R_V + \alpha_2 E_T' E_T)(\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}(u^{\boldsymbol{y}} - \widehat{u}))\|_{\mathcal{X}'} \\
&\leqslant \sum_{\mathbf{0} \neq \boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \frac{1}{\beta_1} \|\partial_{\boldsymbol{y}}^{\boldsymbol{m}} (B^{\boldsymbol{y}})'\|_{\mathcal{Y} \to \mathcal{X}'}\, \|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}-\boldsymbol{m}} q^{\boldsymbol{y}}\|_{\mathcal{Y}} \\
&\quad + \frac{\alpha_1 + \alpha_2 \|E_T\|_{\mathcal{X} \to L^2(D)}^2}{\beta_1} \|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}(u^{\boldsymbol{y}} - \widehat{u})\|_{\mathcal{X}}.
\end{aligned}$$

Recall that

$$\begin{aligned}
&\langle v, (B^{\boldsymbol{y}})' w\rangle_{\mathcal{X},\mathcal{X}'} \\
&= \int_I \langle v, -\tfrac{\partial}{\partial t} w\rangle_{V,V'} \,\mathrm{d}t + \int_I \int_D a^{\boldsymbol{y}}\, \nabla v \cdot \nabla w \,\mathrm{d}x \,\mathrm{d}t + \int_D E_T w\, E_T v \,\mathrm{d}x.
\end{aligned}$$

For $\boldsymbol{m} \neq \mathbf{0}$, we conclude with (4.21) that $\langle v, \partial^{\boldsymbol{m}} (B^{\boldsymbol{y}})' w\rangle_{\mathcal{X},\mathcal{X}'} = \int_I \int_D \psi_j\, \nabla v \cdot \nabla w \,\mathrm{d}x \,\mathrm{d}t$ if $\boldsymbol{m} = \boldsymbol{e}_j$, and otherwise it is zero. Hence for $\boldsymbol{m} = \boldsymbol{e}_j$ we obtain for all $v \in \mathcal{Y}$ that

$$\begin{aligned}
\|\partial^{\boldsymbol{m}} (B^{\boldsymbol{y}})' v\|_{\mathcal{X}'} &= \sup_{w \in \mathcal{X}} \frac{|\langle v, \partial^{\boldsymbol{m}} (B^{\boldsymbol{y}})' w\rangle_{\mathcal{X},\mathcal{X}'}|}{\|w\|_{\mathcal{X}}} = \sup_{w \in \mathcal{X}} \frac{|\int_I \int_D \psi_j\, \nabla v \cdot \nabla w \,\mathrm{d}x \,\mathrm{d}t|}{\|w\|_{\mathcal{X}}} \\
&\leqslant b_j \sup_{w \in \mathcal{X}} \frac{\|v\|_{L^2(V;I)}\, \|w\|_{L^2(V;I)}}{\|w\|_{\mathcal{X}}} \leqslant b_j \|v\|_{\mathcal{Y}}\,.
\end{aligned}$$

Hence

$$\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} q^{\boldsymbol{y}}\|_{\mathcal{Y}} \leqslant \sum_{j \in \operatorname{supp}(\boldsymbol{\nu})} \nu_j\, b_j\, \|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}-\boldsymbol{e}_j} q^{\boldsymbol{y}}\|_{\mathcal{Y}} + \frac{\alpha_1 + \alpha_2 \|E_T\|_{\mathcal{X} \to L^2(D)}^2}{\beta_1} \|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}(u^{\boldsymbol{y}} - \widehat{u})\|_{\mathcal{X}}.$$

By Lemma 4.6.5 this recursion is true for $\boldsymbol{\nu} = \mathbf{0}$ and we may apply Lemma 4.6.1 to get

$$\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} q^{\boldsymbol{y}}\|_{\mathcal{Y}} \leqslant \sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} |\boldsymbol{m}|!\, \boldsymbol{b}^{\boldsymbol{m}} \Big(\frac{\alpha_1 + \alpha_2 \|E_T\|_{\mathcal{X} \to L^2(D)}^2}{\beta_1} \|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}-\boldsymbol{m}}(u^{\boldsymbol{y}} - \widehat{u})\|_{\mathcal{X}} \Big)\,.$$

From (4.27) and (4.85) we have

$$\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}(u^{\boldsymbol{y}}-\widehat{u})\|_{\mathcal{X}} \leqslant \begin{cases} \frac{1}{\beta_1}\|f\|_{\mathcal{Y}'} + \|\widehat{u}\|_{\mathcal{X}} & \text{if } \boldsymbol{\nu}=\mathbf{0}, \\ \frac{1}{\beta_1}\|f\|_{\mathcal{Y}'}\,|\boldsymbol{\nu}|!\,\boldsymbol{b}^{\boldsymbol{\nu}} & \text{if } \boldsymbol{\nu}\neq\mathbf{0}. \end{cases}$$

We finally arrive at

$$\begin{aligned}
\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} q^{\boldsymbol{y}}\|_{\mathcal{Y}} &\leqslant \sum_{\substack{\boldsymbol{m}\leqslant\boldsymbol{\nu}\\ \boldsymbol{m}\neq\boldsymbol{\nu}}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} |\boldsymbol{m}|!\,\boldsymbol{b}^{\boldsymbol{m}} \frac{\alpha_1+\alpha_2\,\|E_T\|^2_{\mathcal{X}\to L^2(D)}}{\beta_1}\,\frac{\|f\|_{\mathcal{Y}'}}{\beta_1}\,|\boldsymbol{\nu}-\boldsymbol{m}|!\,\boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{m}} \\
&\qquad + |\boldsymbol{\nu}|!\,\boldsymbol{b}^{\boldsymbol{\nu}}\,\frac{\alpha_1+\alpha_2\,\|E_T\|^2_{\mathcal{X}\to L^2(D)}}{\beta_1}\Big(\frac{\|f\|_{\mathcal{Y}'}}{\beta_1}+\|\widehat{u}\|_{\mathcal{X}}\Big) \\
&= (|\boldsymbol{\nu}|+1)!\,\boldsymbol{b}^{\boldsymbol{\nu}}\frac{\alpha_1+\alpha_2\,\|E_T\|^2_{\mathcal{X}\to L^2(D)}}{\beta_1}\,\frac{\|f\|_{\mathcal{Y}'}}{\beta_1} \\
&\qquad + |\boldsymbol{\nu}|!\,\boldsymbol{b}^{\boldsymbol{\nu}}\frac{\alpha_1+\alpha_2\,\|E_T\|^2_{\mathcal{X}\to L^2(D)}}{\beta_1}\|\widehat{u}\|_{\mathcal{X}} \\
&\leqslant (|\boldsymbol{\nu}|+1)!\,\boldsymbol{b}^{\boldsymbol{\nu}}\frac{\alpha_1+\alpha_2\,\|E_T\|^2_{\mathcal{X}\to L^2(D)}}{\beta_1}\Big(\frac{\|f\|_{\mathcal{Y}'}}{\beta_1}+\|\widehat{u}\|_{\mathcal{X}}\Big),
\end{aligned}$$

where the equality follows from (4.77). □

Regularity analysis for the entropic risk measure

Our goal is to use QMC to approximate the following high-dimensional integrals appearing in the denominator and numerator of the gradient (4.48). To this end, we develop regularity bounds for the integrands.

Lemma 4.6.7. *Let $\theta>0$, $\alpha_1,\alpha_2\geqslant 0$, with $\alpha_1+\alpha_2>0$. Let $f=(z,u_0)\in\mathcal{Y}'$ and $\widehat{u}\in\mathcal{X}$. For every $\boldsymbol{y}\in U$, let $u^{\boldsymbol{y}}\in\mathcal{X}$ be the solution of (4.20) and let $\Phi^{\boldsymbol{y}}$ be as in (4.37). Then for all $\boldsymbol{\nu}\in\mathcal{F}$ we have*

$$|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}\Phi^{\boldsymbol{y}}| \leqslant \frac{\alpha_1+\alpha_2\,\|E_T\|^2_{\mathcal{X}\to L^2(D)}}{2}\left(\frac{\|f\|_{\mathcal{Y}'}}{\beta_1}+\|\widehat{u}\|_{\mathcal{X}}\right)^2 (|\boldsymbol{\nu}|+1)!\,\boldsymbol{b}^{\boldsymbol{\nu}},$$

where the sequence $\boldsymbol{b}=(b_j)_{j\geqslant 1}$ is defined by (4.86).

Proof. The case $\boldsymbol{\nu}=\mathbf{0}$ is precisely (4.45). Consider now $\boldsymbol{\nu}\neq\mathbf{0}$. We estimate the partial derivatives of $\Phi^{\boldsymbol{y}}$ by differentiating under the integral sign and using the Leibniz product rule in conjunction with the Cauchy–Schwarz inequality to obtain

$$|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}\Phi^{\boldsymbol{y}}| \leqslant \frac{\alpha_1+\alpha_2\,\|E_T\|^2_{\mathcal{X}\to L^2(D)}}{2}\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}\|\partial^{\boldsymbol{m}}(u^{\boldsymbol{y}}-\widehat{u})\|_{\mathcal{X}}\,\|\partial^{\boldsymbol{\nu}-\boldsymbol{m}}(u^{\boldsymbol{y}}-\widehat{u})\|_{\mathcal{X}}.$$

Separating out the $\boldsymbol{m}=\mathbf{0}$ and $\boldsymbol{m}=\boldsymbol{\nu}$ terms and utilizing (4.85), we obtain

$$\begin{aligned}
&\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}\|\partial^{\boldsymbol{m}}(u^{\boldsymbol{y}}-\widehat{u})\|_{\mathcal{X}}\,\|\partial^{\boldsymbol{\nu}-\boldsymbol{m}}(u^{\boldsymbol{y}}-\widehat{u})\|_{\mathcal{X}} \\
&= 2\,\|u^{\boldsymbol{y}}-\widehat{u}\|_{\mathcal{X}}\,\|\partial^{\boldsymbol{\nu}}u^{\boldsymbol{y}}\|_{\mathcal{X}} + \sum_{\substack{\boldsymbol{m}\leqslant\boldsymbol{\nu}\\ \mathbf{0}\neq\boldsymbol{m}\neq\boldsymbol{\nu}}}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}\|\partial^{\boldsymbol{m}}u^{\boldsymbol{y}}\|_{\mathcal{X}}\,\|\partial^{\boldsymbol{\nu}-\boldsymbol{m}}u^{\boldsymbol{y}}\|_{\mathcal{X}}
\end{aligned}$$

$$\leqslant 2\left(\frac{\|f\|_{\mathcal{Y}'}}{\beta_1}+\|\widehat{u}\|_{\mathcal{X}}\right)\frac{\|f\|_{\mathcal{Y}'}}{\beta_1}|\boldsymbol{\nu}|!\,\boldsymbol{b}^{\boldsymbol{\nu}}+\left(\frac{\|f\|_{\mathcal{Y}'}}{\beta_1}\right)^2\boldsymbol{b}^{\boldsymbol{\nu}}\sum_{\substack{\boldsymbol{m}\leqslant\boldsymbol{\nu}\\ \boldsymbol{0}\neq\boldsymbol{m}\neq\boldsymbol{\nu}}}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}|\boldsymbol{m}|!\,|\boldsymbol{\nu}-\boldsymbol{m}|!,$$

where the sum over $\boldsymbol{m}$ can be rewritten as

$$\sum_{\ell=1}^{|\boldsymbol{\nu}|-1}\ell!\,(|\boldsymbol{\nu}|-\ell)!\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},\,|\boldsymbol{m}|=\ell}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}=\sum_{\ell=1}^{|\boldsymbol{\nu}|-1}\ell!\,(|\boldsymbol{\nu}|-\ell)!\binom{|\boldsymbol{\nu}|}{\ell}=|\boldsymbol{\nu}|!\,(|\boldsymbol{\nu}|-1),$$

where we used the identity (4.76). Combining the estimates yields the required result. □

We state a recursive form of Faà di Bruno's formula [141] for the exponential function.

Theorem 4.6.8. *Let $G:U\to\mathbb{R}$. For all $\boldsymbol{y}\in U$ and $\boldsymbol{\nu}\in\mathcal{F}\backslash\{\boldsymbol{0}\}$, we have*

$$\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}\exp(G(\boldsymbol{y}))=\exp(G(\boldsymbol{y}))\sum_{\lambda=1}^{|\boldsymbol{\nu}|}\alpha_{\boldsymbol{\nu},\lambda}(\boldsymbol{y}),$$

where the sequence $(\alpha_{\boldsymbol{\nu},\lambda}(\boldsymbol{y}))_{\boldsymbol{\nu}\in\mathcal{F},\lambda\in\mathbb{N}_0}$ is defined recursively by $\alpha_{\boldsymbol{\nu},0}(\boldsymbol{y})=\delta_{\boldsymbol{\nu},\boldsymbol{0}}$, $\alpha_{\boldsymbol{\nu},\lambda}(\boldsymbol{y})=0$ for $\lambda>|\boldsymbol{\nu}|$, and otherwise

$$\alpha_{\boldsymbol{\nu}+\boldsymbol{e}_j,\lambda}(\boldsymbol{y})=\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}(\partial^{\boldsymbol{\nu}-\boldsymbol{m}+\boldsymbol{e}_j}G)(\boldsymbol{y})\,\alpha_{\boldsymbol{m},\lambda-1}(\boldsymbol{y}),\qquad j\geqslant 1.$$

Proof. This is a special case of [141, Formulas (3,1) and (3.5)] in which f is the exponential function and $m=1$ so that λ is an integer. □

Lemma 4.6.9. *Let the sequence $(\mathbb{A}_{\boldsymbol{\nu},\lambda})_{\boldsymbol{\nu}\in\mathcal{F},\,\lambda\in\mathbb{N}_0}$ satisfy $\mathbb{A}_{\boldsymbol{\nu},0}=\delta_{\boldsymbol{\nu},\boldsymbol{0}}$, $\mathbb{A}_{\boldsymbol{\nu},\lambda}=0$ for $\lambda>|\boldsymbol{\nu}|$, and otherwise satisfy the recursion*

$$\mathbb{A}_{\boldsymbol{\nu}+\boldsymbol{e}_j,\lambda}\leqslant\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}c\,\boldsymbol{\rho}^{\boldsymbol{\nu}-\boldsymbol{m}+\boldsymbol{e}_j}\,(|\boldsymbol{\nu}|-|\boldsymbol{m}|+2)!\,\mathbb{A}_{\boldsymbol{m},\lambda-1},\qquad j\geqslant 1,\tag{4.87}$$

for some $c>0$ and a nonnegative sequence $\boldsymbol{\rho}$. Then for all $\boldsymbol{\nu}\neq\boldsymbol{0}$ and $1\leqslant\lambda\leqslant|\boldsymbol{\nu}|$ we have

$$\mathbb{A}_{\boldsymbol{\nu},\lambda}\leqslant c^{\lambda}\,\boldsymbol{\rho}^{\boldsymbol{\nu}}\sum_{k=1}^{\lambda}\frac{(-1)^{\lambda+k}\,(|\boldsymbol{\nu}|+2k-1)!}{(2k-1)!\,(\lambda-k)!\,k!}.\tag{4.88}$$

The result is sharp in the sense that both inequalities can be replaced by equalities.

Proof. We prove (4.88) for all $\boldsymbol{\nu}\neq\boldsymbol{0}$ and $1\leqslant\lambda\leqslant|\boldsymbol{\nu}|$ by induction on $|\boldsymbol{\nu}|$. The base case $\mathbb{A}_{\boldsymbol{e}_j,1}$ is easy to verify. Let $\boldsymbol{\nu}\neq\boldsymbol{0}$ and suppose that (4.88) holds for all multi-indices $\boldsymbol{m}$ of order $\leqslant|\boldsymbol{\nu}|$ and all $1\leqslant\lambda\leqslant|\boldsymbol{m}|$. The case $\mathbb{A}_{\boldsymbol{\nu}+\boldsymbol{e}_j,1}$ is also straightforward to verify. We consider therefore $2\leqslant\lambda\leqslant|\boldsymbol{\nu}|+1$. Using (4.87) and the induction hypothesis, we have

$$\begin{aligned}&\mathbb{A}_{\boldsymbol{\nu}+\boldsymbol{e}_j,\lambda}\\&\leqslant\sum_{\boldsymbol{0}\neq\boldsymbol{m}\leqslant\boldsymbol{\nu}}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}c\,\boldsymbol{\rho}^{\boldsymbol{\nu}-\boldsymbol{m}+\boldsymbol{e}_j}(|\boldsymbol{\nu}|-|\boldsymbol{m}|+2)!\\&\quad\times\left(c^{\lambda-1}\,\boldsymbol{\rho}^{\boldsymbol{m}}\sum_{k=1}^{\lambda-1}\frac{(-1)^{\lambda-1+k}\,(|\boldsymbol{m}|+2k-1)!}{(2k-1)!\,(\lambda-1-k)!\,k!}\right)\end{aligned}$$

$$= c^\lambda \, \boldsymbol{\rho}^{\boldsymbol{\nu}+\boldsymbol{e}_j} \sum_{\ell=1}^{|\boldsymbol{\nu}|} \sum_{\substack{\boldsymbol{m}\leqslant\boldsymbol{\nu} \\ |\boldsymbol{m}|=\ell}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \sum_{k=1}^{\lambda-1} \frac{(-1)^{\lambda-1+k}\,(|\boldsymbol{\nu}|-\ell+2)!\,(\ell+2k-1)!}{(2k-1)!\,(\lambda-1-k)!\,k!}$$

$$= c^\lambda \, \boldsymbol{\rho}^{\boldsymbol{\nu}+\boldsymbol{e}_j} \, \frac{2\,|\boldsymbol{\nu}|!\,(-1)^{\lambda-1}}{(\lambda-1)!} \underbrace{\sum_{k=1}^{\lambda-1} (-1)^k \binom{\lambda-1}{k} \sum_{\ell=1}^{|\boldsymbol{\nu}|} \binom{|\boldsymbol{\nu}|-\ell+2}{|\boldsymbol{\nu}|-\ell} \binom{\ell+2k-1}{\ell}}_{=:\,T}, \tag{4.89}$$

where we used (4.76) and then regrouped the factors as binomial coefficients. Next we take the binomial identity [128, Equation (5.6)]

$$\sum_{\ell=0}^{|\boldsymbol{\nu}|} \binom{|\boldsymbol{\nu}|-\ell+2}{|\boldsymbol{\nu}|-\ell} \binom{\ell+2k-1}{\ell} = \binom{|\boldsymbol{\nu}|+2k+2}{|\boldsymbol{\nu}|},$$

separate out the $\ell = 0$ term, and use $\sum_{k=1}^{\lambda-1}(-1)^k\binom{\lambda-1}{k} = \sum_{k=0}^{\lambda-1}(-1)^k\binom{\lambda-1}{k} - 1 = -1$, to rewrite T as

$$\begin{aligned} T &= \sum_{k=1}^{\lambda-1} (-1)^k \binom{\lambda-1}{k} \left[\binom{|\boldsymbol{\nu}|+2k+2}{|\boldsymbol{\nu}|} - \binom{|\boldsymbol{\nu}|+2}{|\boldsymbol{\nu}|} \right] \\ &= \sum_{k=1}^{\lambda-1} (-1)^k \binom{\lambda-1}{k} \binom{|\boldsymbol{\nu}|+2k+2}{|\boldsymbol{\nu}|} + \binom{|\boldsymbol{\nu}|+2}{|\boldsymbol{\nu}|} \\ &= \sum_{k=0}^{\lambda-1} (-1)^k \binom{\lambda-1}{k} \binom{|\boldsymbol{\nu}|+2k+2}{|\boldsymbol{\nu}|} = \sum_{k=1}^{\lambda} (-1)^{k-1} \binom{\lambda-1}{k-1} \binom{|\boldsymbol{\nu}|+2k}{|\boldsymbol{\nu}|}. \end{aligned}$$

Substituting this back into (4.89) and simplifying the factors, we obtain

$$\mathbb{A}_{\boldsymbol{\nu}+\boldsymbol{e}_j,\lambda} \leqslant c^\lambda \, \boldsymbol{\rho}^{\boldsymbol{\nu}+\boldsymbol{e}_j} \sum_{k=1}^{\lambda} \frac{(-1)^{\lambda+k}\,(|\boldsymbol{\nu}|+2k)!}{(2k-1)!\,(\lambda-k)!\,k!},$$

as required. □

Theorem 4.6.10. *Let $\theta > 0$, $\alpha_1, \alpha_2 \geqslant 0$, with $\alpha_1 + \alpha_2 > 0$. Let $f = (z, u_0) \in \mathcal{Y}'$ and $\widehat{u} \in \mathcal{X}$. For every $\boldsymbol{y} \in U$, let $u^{\boldsymbol{y}} \in \mathcal{X}$ be the solution of (4.20) and let $\Phi^{\boldsymbol{y}}$ be as in (4.37). Then for all $\boldsymbol{\nu} \in \mathcal{F}$ we have*

$$|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} \exp(\theta\,\Phi^{\boldsymbol{y}})| \leqslant e^{\max(\sigma,\,\sigma e^2+2\sigma-1)}\,|\boldsymbol{\nu}|!\,(e\boldsymbol{b})^{\boldsymbol{\nu}},$$

where the sequence $\boldsymbol{b} = (b_j)_{j\geqslant 1}$ is defined by (4.86) and σ is defined by (4.47).

Proof. For $\boldsymbol{\nu} = \boldsymbol{0}$ we have from (4.46) that $|\exp(\theta\,\Phi^{\boldsymbol{y}})| \leqslant e^\sigma$, which satisfies the required bound. For $\boldsymbol{\nu} \neq \boldsymbol{0}$, from Faà di Bruno's formula (Theorem 4.6.8) we have

$$|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} \exp(\theta\Phi^{\boldsymbol{y}})| \leqslant \exp(\theta\,\Phi^{\boldsymbol{y}}) \sum_{\lambda=1}^{|\boldsymbol{\nu}|} |\alpha_{\boldsymbol{\nu},\lambda}(\boldsymbol{y})|, \tag{4.90}$$

with $\alpha_{\boldsymbol{\nu},0}(\boldsymbol{y}) = \delta_{\boldsymbol{\nu},\boldsymbol{0}}$, $\alpha_{\boldsymbol{\nu},\lambda}(\boldsymbol{y}) = 0$ for $\lambda > |\boldsymbol{\nu}|$, and

$$|\alpha_{\boldsymbol{\nu}+\boldsymbol{e}_j,\lambda}(\boldsymbol{y})| \leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \theta\,|\partial_{\boldsymbol{y}}^{\boldsymbol{m}+\boldsymbol{e}_j}\Phi^{\boldsymbol{y}}|\,|\alpha_{\boldsymbol{\nu}-\boldsymbol{m},\lambda-1}(\boldsymbol{y})|$$

$$\leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \sigma\,(|\boldsymbol{m}|+2)!\,\boldsymbol{b}^{\boldsymbol{m}+\boldsymbol{e}_j}\,|\alpha_{\boldsymbol{\nu}-\boldsymbol{m},\lambda-1}(\boldsymbol{y})|,$$

where we used Lemma 4.6.7. Applying Lemma 4.6.9 we conclude that

$$|\alpha_{\boldsymbol{\nu},\lambda}(\boldsymbol{y})| \leqslant \sigma^\lambda\,\boldsymbol{b}^{\boldsymbol{\nu}} \sum_{k=1}^{\lambda} \frac{(-1)^{\lambda+k}\,(|\boldsymbol{\nu}|+2k-1)!}{(2k-1)!\,(\lambda-k)!\,k!}. \tag{4.91}$$

We have

$$\begin{aligned}
&\sum_{\lambda=1}^{|\boldsymbol{\nu}|} \sigma^\lambda \sum_{k=1}^{\lambda} \frac{(-1)^{\lambda+k}\,(|\boldsymbol{\nu}|+2k-1)!}{(2k-1)!\,(\lambda-k)!\,k!} = \sum_{k=1}^{|\boldsymbol{\nu}|} \frac{(|\boldsymbol{\nu}|+2k-1)!}{(2k-1)!\,k!} \sum_{\lambda=k}^{|\boldsymbol{\nu}|} \frac{(-1)^{\lambda+k}\,\sigma^\lambda}{(\lambda-k)!} \\
&= |\boldsymbol{\nu}|! \sum_{k=1}^{|\boldsymbol{\nu}|} \frac{\sigma^k}{k!} \binom{|\boldsymbol{\nu}|+2k-1}{2k-1} \sum_{\ell=0}^{|\boldsymbol{\nu}|-k} \frac{(-\sigma)^\ell}{\ell!} \leqslant |\boldsymbol{\nu}|! \sum_{k=1}^{|\boldsymbol{\nu}|} \frac{\sigma^k}{k!} e^{|\boldsymbol{\nu}|+2k-1} e^\sigma \\
&\leqslant |\boldsymbol{\nu}|!\, e^{|\boldsymbol{\nu}|+\sigma e^2+\sigma-1},
\end{aligned} \tag{4.92}$$

where we used $\binom{n}{m} \leqslant n^m/m! \leqslant e^n$. Combining (4.90), (4.91), (4.92) and (4.45) gives

$$|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} \exp(\theta\Phi^{\boldsymbol{y}})| \leqslant \exp(\sigma)\,\boldsymbol{b}^{\boldsymbol{\nu}}\,|\boldsymbol{\nu}|!\,e^{|\boldsymbol{\nu}|+\sigma e^2+\sigma-1} = e^{\sigma e^2+2\sigma-1}\,|\boldsymbol{\nu}|!\,(e\boldsymbol{b})^{\boldsymbol{\nu}},$$

as required. □

Remark 4.6.11. *In the proof of Theorem 4.6.10, a different manipulation of* (4.92) *can yield a different bound* $2c\,e^{|\boldsymbol{\nu}|+\sigma e^2+\sigma+1}(|\boldsymbol{\nu}|-1)!$ *for* $\boldsymbol{\nu}\neq\mathbf{0}$*, leading to a tighter upper bound for large* $|\boldsymbol{\nu}|$ *at the expense of a bigger constant,*

$$|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} \exp(\theta\,\Phi^{\boldsymbol{y}})| \leqslant 2\sigma\,\mathrm{e}^{\sigma e^2+2\sigma+1}\,(|\boldsymbol{\nu}|-1)!\,(e\boldsymbol{b})^{\boldsymbol{\nu}}.$$

This leads to a more complicated bound for Theorem 4.6.12 below. Hence we have chosen to present the current form of Theorem 4.6.10 to simplify our subsequent analysis.
Interestingly, the sum in (4.91) *can also be rewritten as a sum with only positive terms: denoting* $v=|\boldsymbol{\nu}|$*,*

$$\begin{aligned}
\sum_{k=1}^{\lambda} \frac{(-1)^{\lambda+k}(v+2k-1)!}{(2k-1)!(\lambda-k)!k!} &= \frac{v!}{\lambda!} \sum_{k=0}^{\lambda} \binom{\lambda}{k}\binom{v-1}{v-\lambda-k} 2^{\lambda-k} \\
&= 2^\lambda \binom{v-1}{v-\lambda} \sum_{k=0}^{\lambda} \frac{\binom{\lambda}{k}\binom{v-\lambda}{k}}{\binom{\lambda+k-1}{k}} 2^{-k},
\end{aligned}$$

which is identical to the sequence [17, Proposition 7] and the sequence A181289 in the OEIS (written in slightly different form). However, we were unable to find a closed form expression for the sum; neither [67] nor [128] were useful to us in this case. The hope is to obtain an alternative bound for (4.92) *that does not involve the factor* $e^{|\boldsymbol{\nu}|}$*, which remains open for future research.*

Theorem 4.6.12. *Let* $\theta>0$*,* $\alpha_1,\alpha_2\geqslant 0$*, with* $\alpha_1+\alpha_2>0$*. Let* $f=(z,u_0)\in\mathcal{Y}'$ *and* $\widehat{u}\in\mathcal{X}$*. For every* $\boldsymbol{y}\in U$*, let* $u^{\boldsymbol{y}}\in\mathcal{X}$ *be the solution of* (4.20) *and* $\Phi^{\boldsymbol{y}}$ *be as in* (4.37)*, and then let* $q^{\boldsymbol{y}}=(q_1^{\boldsymbol{y}},q_2^{\boldsymbol{y}})\in\mathcal{Y}$ *be the solution of* (4.32) *with* f_{dual} *given by* (4.40)*. Then for all* $\boldsymbol{\nu}\in\mathcal{F}$ *we have*

$$\big\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}\big(\exp(\theta\,\Phi^{\boldsymbol{y}})\,q_1^{\boldsymbol{y}}\big)\big\|_{L^2(V;I)} \leqslant \big\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}\big(\exp(\theta\,\Phi^{\boldsymbol{y}})\,q^{\boldsymbol{y}}\big)\big\|_{\mathcal{Y}} \leqslant \frac{\mu}{2}\,(|\boldsymbol{\nu}|+2)!\,(e\boldsymbol{b})^{\boldsymbol{\nu}},$$

where the sequence $\boldsymbol{b} = (b_j)_{j\geqslant 1}$ *is defined by* (4.86), σ *is defined by* (4.47) *and*

$$\mu := e^{\max(\sigma,\, \sigma e^2+2\sigma-1)} \Big(\frac{\alpha_1 + \alpha_2 \,\|E_T\|_{\mathcal{X}\to L^2(D)}}{\beta_1}\Big)\Big(\frac{\|f\|_{\mathcal{Y}'}}{\beta_1} + \|\widehat{u}\|_{\mathcal{X}}\Big).$$

Proof. Using the Leibniz product rule and Theorem 4.6.10 with Theorem 4.6.6, we obtain

$$\begin{aligned}
&\big\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}\big(\exp(\theta\,\Phi^{\boldsymbol{y}})\,q^{\boldsymbol{y}}\big)\big\|_{\mathcal{Y}} \leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \big|\partial_{\boldsymbol{y}}^{\boldsymbol{m}} \exp(\theta\Phi^{\boldsymbol{y}})\big|\,\big\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}-\boldsymbol{m}} q^{\boldsymbol{y}}\big\|_{\mathcal{Y}} \\
&\leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} e^{\max(\sigma,\sigma\, e^2+2\sigma-1)}\, |\boldsymbol{m}|!\,(e\boldsymbol{b})^{\boldsymbol{m}} \\
&\qquad \times \Big(\frac{\alpha_1 + \alpha_2 \,\|E_T\|^2_{\mathcal{X}\to L^2(D)}}{\beta_1}\Big)\Big(\frac{\|f\|_{\mathcal{Y}'}}{\beta_1} + \|\widehat{u}\|_{\mathcal{X}}\Big)\, \boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{m}}\,(|\boldsymbol{\nu}| - |\boldsymbol{m}| + 1)! \\
&\leqslant \mu\,(e\boldsymbol{b})^{\boldsymbol{\nu}} \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} |\boldsymbol{m}|!\,(|\boldsymbol{\nu}| - |\boldsymbol{m}| + 1)! \;=\; \mu\,(e\boldsymbol{b})^{\boldsymbol{\nu}}\, \frac{(|\boldsymbol{\nu}|+2)!}{2},
\end{aligned}$$

with the last equality due to [113, Formula (9.5)]. □

4.6.3 Analytic parametric linear operators

From Theorem 2.3.5 we know that the solution $u(\boldsymbol{y})$ of an analytic linear operator equation $A(\boldsymbol{y})u(\boldsymbol{y}) = f$ is again analytic. In the optimization problems studied in this manuscript, there arise operator equations in which also the right-hand side depends (analytically) on the parametric variables, see, e.g., the adjoint problem (4.55). The following result shows that in this case the dependence of $u(\boldsymbol{y})$ on the parameter sequence is again analytic, i.e., it generalized Theorem 2.3.5 to problems of the form $A(\boldsymbol{y})u(\boldsymbol{y}) = f(\boldsymbol{y})$, when f depends analytically on $\boldsymbol{y} \in U$.

Theorem 4.6.13. *Let the parametric family of operators* $\{A(\boldsymbol{y}) \in \mathcal{L}(X, Y') : \boldsymbol{y} \in U\}$ *satisfy Assumption 2.3.1 for some* $0 < p \leqslant 1$. *Then, for* $f(\boldsymbol{y}) \in Y'$, *with* $\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} f(\boldsymbol{y})\|_{Y'} \leqslant C|\boldsymbol{\nu}|!\boldsymbol{b}^{\boldsymbol{\nu}}/(\ln 2)^{|\boldsymbol{\nu}|}$ *for all finitely supported multiindices* $\boldsymbol{\nu} \in \mathcal{F}$, *and every* $\boldsymbol{y} \in U$ *there exists a unique solution* $u(\boldsymbol{y}) \in X$ *of the parametric operator equation*

$$A(\boldsymbol{y})u(\boldsymbol{y}) = f(\boldsymbol{y}) \tag{4.93}$$

and the parametric solution family $u(\boldsymbol{y})$ *depends analytically on the parameters* $\boldsymbol{y} \in U$, *with partial derivatives satisfying*

$$\sup_{\boldsymbol{y}\in U} \|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} u(\boldsymbol{y})\|_X \leqslant C(|\boldsymbol{\nu}| + 1)!\frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}}\,.$$

Proof. We prove the result by induction with respect to $|\boldsymbol{\nu}|$. If $|\boldsymbol{\nu}| = 0$, then $\boldsymbol{\nu} = \mathbf{0}$ and the result follows from Assumption 2.3.1 (i) and the a-priori bound (5.13). For $\mathbf{0} \neq \boldsymbol{\nu} \in \mathcal{F}$ we take the partial derivative $\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}$ of (4.94). By the Leibniz product rule we get

$$\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} (\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}-\boldsymbol{m}} A(\boldsymbol{y}))(\partial_{\boldsymbol{y}}^{\boldsymbol{m}} u(\boldsymbol{y})) = \partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} f(\boldsymbol{y})$$

Separating out the $\boldsymbol{\nu} = \boldsymbol{m}$ term, we obtain

$$A(\boldsymbol{y})(\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} u(\boldsymbol{y})) = -\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},\boldsymbol{m}\neq\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} (\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}-\boldsymbol{m}} A(\boldsymbol{y}))(\partial_{\boldsymbol{y}}^{\boldsymbol{m}} u(\boldsymbol{y})) + \partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} f(\boldsymbol{y})\,.$$

By Assumption 2.3.1 (i) we have

$$\begin{aligned}\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} u(\boldsymbol{y})\|_X &\leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},\boldsymbol{m}\neq\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \|A(\boldsymbol{y})^{-1}\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}-\boldsymbol{m}} A(\boldsymbol{y})\|_{\mathcal{L}(X)} \|\partial_{\boldsymbol{y}}^{\boldsymbol{m}} u(\boldsymbol{y})\|_X \\ &\quad + \|A(\boldsymbol{y})^{-1}\|_{\mathcal{L}(Y',X)} \|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} f(\boldsymbol{y})\|_{\mathcal{Y}'} \\ &\leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},\boldsymbol{m}\neq\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} C\boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{m}} \|\partial_{\boldsymbol{y}}^{\boldsymbol{m}} u(\boldsymbol{y})\|_X + C\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} f(\boldsymbol{y})\|_{Y'} .\end{aligned}$$

where we concluded from Assumption 2.3.1 that for all $\boldsymbol{\nu} \in \mathcal{F}$

$$\sup_{\boldsymbol{y}\in U} \|A(\boldsymbol{y})^{-1}\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} A(\boldsymbol{y})\|_{\mathcal{L}(X)} \leqslant \sup_{\boldsymbol{y}\in U} \|A(\boldsymbol{y})^{-1}A(\boldsymbol{0})\|_{\mathcal{L}(X)} \sup_{\boldsymbol{y}\in U} \|A(\boldsymbol{0})^{-1}\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} A(\boldsymbol{y})\|_{\mathcal{L}(X)} \leqslant C\boldsymbol{b}^{\boldsymbol{\nu}} .$$

From Lemma 4.6.1 we conclude that

$$\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} u(\boldsymbol{y})\|_X \leqslant C \sum_{\boldsymbol{k}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{k}} \frac{|\boldsymbol{k}|!}{(\ln 2)^{|\boldsymbol{k}|}} \boldsymbol{b}^{\boldsymbol{k}} \|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}-\boldsymbol{k}} f(\boldsymbol{y})\|_{Y'} .$$

By the assumption on f we obtain

$$\begin{aligned}\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} u(\boldsymbol{y})\|_X &\leqslant C \sum_{\boldsymbol{k}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{k}} \frac{|\boldsymbol{k}|!}{(\ln 2)^{|\boldsymbol{k}|}} \boldsymbol{b}^{\boldsymbol{k}} C \frac{\boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{k}}}{(\ln 2)^{|\boldsymbol{\nu}-\boldsymbol{k}|}} |\boldsymbol{\nu}-\boldsymbol{k}|! \\ &= C(|\boldsymbol{\nu}|+1)! \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} .\end{aligned}$$

□

Similar to Corollary 2.3.6 one can proof the following result for affine parametric linear operators.

Corollary 4.6.14. *Let the parametric family of operators $\{A(\boldsymbol{y}) \in \mathcal{L}(X,Y') : \boldsymbol{y} \in U\}$ satisfy Assumption 2.3.1 for some $0 < p \leqslant 1$ and in addition Assumption 2.3.2. Then, for $f(\boldsymbol{y}) \in Y'$, with $\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} f(\boldsymbol{y})\|_{Y'} \leqslant C|\boldsymbol{\nu}|!\boldsymbol{b}^{\boldsymbol{\nu}}$ for all finitely supported multiindices $\boldsymbol{\nu} \in \mathcal{F}$, and every $\boldsymbol{y} \in U$ there exists a unique solution $u(\boldsymbol{y}) \in X$ of the parametric operator equation*

$$A(\boldsymbol{y})u(\boldsymbol{y}) = f(\boldsymbol{y}) \tag{4.94}$$

and the parametric solution family $u(\boldsymbol{y})$ depends analytically on the parameters $\boldsymbol{y} \in U$, with partial derivatives satisfying

$$\sup_{\boldsymbol{y}\in U} \|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} u(\boldsymbol{y})\|_X \leqslant C(|\boldsymbol{\nu}|+1)!\boldsymbol{b}^{\boldsymbol{\nu}} .$$

In optimization problems, operator equations with analytic right-hand sides typically stem from the adjoint problem, see, e.g., (4.55). The following result shows that the dual operator (and hence the adjoint operator) admits the same parametric regularity as the operator itself.

Lemma 4.6.15. *The dual operator $A(\boldsymbol{y})'$ of an operator $A(\boldsymbol{y})$ that satisfies Assumption 2.3.1 admits the same regularity, i.e., for all $\boldsymbol{\nu} \in \mathcal{F}$ it holds that*

$$\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} A(\boldsymbol{y})\|_{\mathcal{L}(X,Y')} = \|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} A(\boldsymbol{y})'\|_{\mathcal{L}(Y,X')} . \tag{4.95}$$

Proof. We have

$$\begin{aligned}\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} A(\boldsymbol{y})\|_{\mathcal{L}(X,Y')} &= \sup_{v\in X}\sup_{w\in Y}\frac{|\langle \partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} A(\boldsymbol{y})v, w\rangle_{Y',Y}|}{\|v\|_X\|w\|_Y} = \sup_{v\in X}\sup_{w\in Y}\frac{|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}\langle A(\boldsymbol{y})v, w\rangle_{Y',Y}|}{\|v\|_X\|w\|_Y}\\ &= \sup_{v\in X}\sup_{w\in Y}\frac{|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}\langle v, A(\boldsymbol{y})'w\rangle_{X,X'}|}{\|v\|_X\|w\|_Y} = \sup_{v\in X}\sup_{w\in Y}\frac{|\langle v, \partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} A(\boldsymbol{y})'w\rangle_{X,X'}|}{\|v\|_X\|w\|_Y}\\ &= \sup_{w\in\mathcal{Y}}\frac{\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} A(\boldsymbol{y})'w\|_{X'}}{\|w\|_Y} = \|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} A(\boldsymbol{y})'\|_{\mathcal{L}(Y,X')}\,.\end{aligned}$$

□

In the case with the expected value as a risk measure, we have

$$q(\boldsymbol{y}) := (A^{-1}(\boldsymbol{y}))'\mathcal{Q}' R_{\mathfrak{J}}(A^{-1}(\boldsymbol{y})\mathcal{B}z - \widehat{u}) \in \mathcal{Y}, \tag{4.96}$$

From Corollary 2.3.6 we conclude that

$$\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}\mathcal{Q}u(\boldsymbol{y})\|_{\mathfrak{J}} = \|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}\mathcal{Q}(A^{-1}(\boldsymbol{y})\mathcal{B}z\|_{\mathfrak{J}} \leqslant C_{\mathcal{Q}}CC_{\mathcal{B}}\|z\|_{\mathcal{Z}}|\boldsymbol{\nu}|!\boldsymbol{b}^{\boldsymbol{\nu}}\,,$$

where C is defined in Assumption 2.3.2. Using this bound and Lemma 4.6.15, we can apply Corollary 4.6.14 and obtain

Lemma 4.6.16. *Let $A(\boldsymbol{y})$, $\boldsymbol{y}\in U$ satisfy Assumption 2.3.2. Then, under the assumptions in Section 4.3, it holds*

$$\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} q(\boldsymbol{y})\|_{\mathcal{Y}} \leqslant C_{\mathcal{Q}}CC_{\mathcal{B}}(|\boldsymbol{\nu}|+1)!\boldsymbol{b}^{\boldsymbol{\nu}}(\|z\|_{\mathcal{Z}} + \|\widehat{u}\|_{\mathfrak{J}}),$$

where the constant C is defined in Assumption 2.3.2.

The preceeding lemma holds with $\boldsymbol{b}^{\boldsymbol{\nu}}$ replaced by $\boldsymbol{b}^{\boldsymbol{\nu}}/(\ln 2)^{|\boldsymbol{\nu}|}$ for general non-affine operators that satisfy Assumption 2.3.1.

Regularity analysis for the entropic risk measure

We investigate the parametric regularity in the case of the entropic risk measure in conjunction with analytic parametric operator equations. To this end, we denote

$$\widetilde{\Phi}^{\boldsymbol{y}} := \|\mathcal{Q}A^{-1}(\boldsymbol{y})\mathcal{B}z - \widehat{u}\|_{\mathfrak{J}}^2. \tag{4.97}$$

Lemma 4.6.17. *Let $\widetilde{\Phi}^{\boldsymbol{y}}$ be as in (4.97). Then for all $\boldsymbol{\nu}\in\mathcal{F}$ we have*

$$|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}\widetilde{\Phi}^{\boldsymbol{y}}| \leqslant \left(C_{\mathcal{Q}}CC_{\mathcal{B}}\|z\|_{\mathcal{Z}} + \|\widehat{u}\|_{\mathcal{X}}\right)^2 (|\boldsymbol{\nu}|+1)!\,\boldsymbol{b}^{\boldsymbol{\nu}},$$

where the sequence $\boldsymbol{b} = (b_j)_{j\geqslant 1}$ is defined in Assumption 2.3.1.

Proof. Let $\boldsymbol{\nu} = \mathbf{0}$, then

$$|\widetilde{\Phi}^{\boldsymbol{y}}| \leqslant (\|\mathcal{Q}A^{-1}(\boldsymbol{y})\mathcal{B}z\|_{\mathfrak{J}} + \|\widehat{u}\|_{\mathfrak{J}})^2 \leqslant \left(C_{\mathcal{Q}}CC_{\mathcal{B}}\|z\|_{\mathcal{Z}} + \|\widehat{u}\|_{\mathfrak{J}}\right)^2,$$

where C is the constant from Assumption 2.3.1. Consider now $\boldsymbol{\nu}\neq\mathbf{0}$. We estimate the partial derivatives of $\widetilde{\Phi}^{\boldsymbol{y}}$ by differentiating under the integral sign and using the Leibniz product rule in conjunction with the Cauchy–Schwarz inequality to obtain

$$|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}\widetilde{\Phi}^{\boldsymbol{y}}| \leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}\|\partial^{\boldsymbol{m}}(\mathcal{Q}u^{\boldsymbol{y}} - \widehat{u})\|_{\mathfrak{J}}\,\|\partial^{\boldsymbol{\nu}-\boldsymbol{m}}(\mathcal{Q}u^{\boldsymbol{y}} - \widehat{u})\|_{\mathfrak{J}}.$$

Separating out the $\boldsymbol{m} = \mathbf{0}$ and $\boldsymbol{m} = \boldsymbol{\nu}$ terms and utilizing (4.85), we obtain

$$\begin{aligned}
&\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \|\partial^{\boldsymbol{m}}(\mathcal{Q}u^{\boldsymbol{y}} - \widehat{u})\|_{\mathfrak{J}} \, \|\partial^{\boldsymbol{\nu}-\boldsymbol{m}}(\mathcal{Q}u^{\boldsymbol{y}} - \widehat{u})\|_{\mathfrak{J}} \\
&= 2\,\|\mathcal{Q}u^{\boldsymbol{y}} - \widehat{u}\|_{\mathfrak{J}} \, \|\partial^{\boldsymbol{\nu}}\mathcal{Q}u^{\boldsymbol{y}}\|_{\mathfrak{J}} + \sum_{\substack{\boldsymbol{m}\leqslant\boldsymbol{\nu} \\ \mathbf{0}\neq\boldsymbol{m}\neq\boldsymbol{\nu}}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \|\partial^{\boldsymbol{m}}\mathcal{Q}u^{\boldsymbol{y}}\|_{\mathfrak{J}} \, \|\partial^{\boldsymbol{\nu}-\boldsymbol{m}}\mathcal{Q}u^{\boldsymbol{y}}\|_{\mathfrak{J}} \\
&\leqslant 2\,(C_{\mathcal{Q}}CC_{\mathcal{B}}\|z\|_{\mathcal{Z}} + \|\widehat{u}\|_{\mathcal{X}})C_{\mathcal{Q}}CC_{\mathcal{B}}\|z\|_{\mathcal{Z}}|\boldsymbol{\nu}|!\,\boldsymbol{b}^{\boldsymbol{\nu}} \\
&\quad + (C_{\mathcal{Q}}CC_{\mathcal{B}}\|z\|_{\mathcal{Z}})^2\boldsymbol{b}^{\boldsymbol{\nu}} \sum_{\substack{\boldsymbol{m}\leqslant\boldsymbol{\nu} \\ \mathbf{0}\neq\boldsymbol{m}\neq\boldsymbol{\nu}}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} |\boldsymbol{m}|!\,|\boldsymbol{\nu}-\boldsymbol{m}|!,
\end{aligned}$$

where the sum over $\boldsymbol{m}$ can be rewritten as

$$\sum_{\ell=1}^{|\boldsymbol{\nu}|-1} \ell!\,(|\boldsymbol{\nu}|-\ell)! \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},\,|\boldsymbol{m}|=\ell} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} = \sum_{\ell=1}^{|\boldsymbol{\nu}|-1} \ell!\,(|\boldsymbol{\nu}|-\ell)! \binom{|\boldsymbol{\nu}|}{\ell} = |\boldsymbol{\nu}|!\,(|\boldsymbol{\nu}|-1),$$

where we used the identity (4.76) again. □

Defining

$$\widetilde{\sigma} := \theta(C_{\mathcal{Q}}CC_{\mathcal{B}}\|z\|_{\mathcal{Z}} + \|\widehat{u}\|_{\mathfrak{J}})^2\,, \tag{4.98}$$

we obtain the following result.

Theorem 4.6.18. *Let $\theta > 0$. Let $\widetilde{\Phi}^{\boldsymbol{y}}$ be as in (4.97). Then for all $\boldsymbol{\nu} \in \mathcal{F}$ we have*

$$|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} \exp(\theta\,\widetilde{\Phi}^{\boldsymbol{y}})| \leqslant e^{\max(\widetilde{\sigma},\,\widetilde{\sigma}e^2+2\widetilde{\sigma}-1)}\,|\boldsymbol{\nu}|!\,(e\boldsymbol{b})^{\boldsymbol{\nu}},$$

where the sequence $\boldsymbol{b} = (b_j)_{j\geqslant 1}$ is defined in Assumption 2.3.1.

Proof. The steps of the proof are exactly the same is in Theorem 4.6.10. □

Theorem 4.6.19. *Let $\widetilde{\Phi}^{\boldsymbol{y}}$ be as in (4.97), and let $q^{\boldsymbol{y}} \in \mathcal{Y}$ be as in (4.96). Then for all $\boldsymbol{\nu} \in \mathcal{F}$ we have*

$$\big\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}\big(\exp(\theta\,\widetilde{\Phi}^{\boldsymbol{y}})\,q^{\boldsymbol{y}}\big)\big\|_{\mathcal{Y}} \leqslant \frac{\mu}{2}\,(|\boldsymbol{\nu}|+2)!\,(e\boldsymbol{b})^{\boldsymbol{\nu}},$$

where the sequence $\boldsymbol{b} = (b_j)_{j\geqslant 1}$ is defined in Assumption 2.3.1, $\widetilde{\sigma}$ is defined by (4.98) and

$$\mu := e^{\max(\widetilde{\sigma},\,\widetilde{\sigma}e^2+2\widetilde{\sigma}-1)}\,C_{\mathcal{Q}}CC_{\mathcal{B}}(\|z\|_{\mathcal{Z}} + \|\widehat{u}\|_{\mathfrak{J}}).$$

Proof. Using the Leibniz product rule and Theorem 4.6.18 with Lemma 4.6.16, we obtain

$$\begin{aligned}
\big\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}}\big(\exp(\theta\,\widetilde{\Phi}^{\boldsymbol{y}})\,q^{\boldsymbol{y}}\big)\big\|_{\mathcal{Y}} &\leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} |\partial_{\boldsymbol{y}}^{\boldsymbol{m}} \exp(\theta\widetilde{\Phi}^{\boldsymbol{y}})| \, \big\|\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}-\boldsymbol{m}} q^{\boldsymbol{y}}\big\|_{\mathcal{Y}} \\
&\leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} e^{\max(\widetilde{\sigma},\,\widetilde{\sigma}e^2+2\widetilde{\sigma}-1)}\,|\boldsymbol{m}|!\,(e\boldsymbol{b})^{\boldsymbol{m}} \\
&\qquad \times C_{\mathcal{Q}}CC_{\mathcal{B}}(\|z\|_{\mathcal{Z}} + \|\widehat{u}\|_{\mathfrak{J}})\,\boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{m}}\,(|\boldsymbol{\nu}| - |\boldsymbol{m}| + 1)! \\
&\leqslant \mu\,(e\boldsymbol{b})^{\boldsymbol{\nu}} \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} |\boldsymbol{m}|!\,(|\boldsymbol{\nu}| - |\boldsymbol{m}| + 1)! \;=\; \mu\,(e\boldsymbol{b})^{\boldsymbol{\nu}}\,\frac{(|\boldsymbol{\nu}|+2)!}{2},
\end{aligned}$$

with the last equality due to [113, Formula (9.5)]. □

5 Truncation of the parametric dimension

We have seen in Section 4.5 that the overall discretization error of the optimal control can be decomposed into dimension truncation error, cubature error, and spatial discretization error of the gradients of the objective function. In this chapter, which is strongly based on the joint work with Vesa Kaarnioja [74], we focus on the analysis of the dimension truncation error for integral quantities. The dimension truncation error is analyzed in an abstract setting, leading to results that apply to a wide range of problems in the field of uncertainty quantification, including optimal control problems under uncertainty. In fact, the contribution of dimension truncation error is independent of the numerical scheme chosen for the cubature operator or spatial discretization, which allows us to approach the problem from a general vantage point—for instance, we do not need to restrict our analysis to a specific numerical cubature method, spatial discretization scheme or even a specific mathematical model problem. Instead, we derive general conditions under which it is possible to derive explicit rates for the dimension truncation error. We begin this chapter with analyzing the dimension truncation error for abstract uncertainty quantification problems, and then apply the developed results to optimal control problems under uncertainty.

In the field of uncertainty quantification it is common to consider mathematical models where uncertain inputs are parameterized by infinite sequences of random variables. For instance, consider an abstract mathematical model $M\colon \mathcal{X} \times U \to \mathcal{Y}$ such that

$$M(g(\boldsymbol{y}), \boldsymbol{y}) = 0,$$

where $\mathcal{X}$ and $\mathcal{Y}$ are separable Banach spaces and U is a nonempty subset of an infinite-dimensional sequence space of parameters $\mathbb{R}^{\mathbb{N}}$. If there exists a solution $g(\boldsymbol{y}) \in \mathcal{X}$ for all $\boldsymbol{y} \in U$, then a natural quantity to investigate is the expected value

$$I(g) := \int_U g(\boldsymbol{y})\, \boldsymbol{\mu}(\mathrm{d}\boldsymbol{y}), \tag{5.1}$$

where $\boldsymbol{\mu}$ is a probability measure over U. In many applications, $\boldsymbol{\mu}$ is either chosen as the uniform probability measure over $U = [-1, 1]^{\mathbb{N}}$ or a Gaussian probability measure over $U = \mathbb{R}^{\mathbb{N}}$. In the setting of optimal control problems under uncertainty the mathematical model is the constraint and integrals of the form (5.1) appear in the objective function and the derivatives of the objective function, see Section 3.1.

For the numerical treatment of (5.1), a natural first step is to consider a dimensionally-truncated model $M_s\colon \mathcal{X} \times U_s \to \mathcal{Y}$ such that

$$M_s(g_s(\boldsymbol{y}_{\leqslant s}), \boldsymbol{y}_{\leqslant s}) = 0, \tag{5.2}$$

where $\varnothing \neq U_s \subseteq \mathbb{R}^s$ and $g_s(\boldsymbol{y}_{\leqslant s}) \in \mathcal{X}$ for all $\boldsymbol{y}_{\leqslant s} \in U_s$. The corresponding expected value in this case is then given by

$$I_s(g_s) := \int_{U_s} g_s(\boldsymbol{y}_{\leqslant s})\, \boldsymbol{\mu}_{\leqslant s}(\mathrm{d}\boldsymbol{y}_{\leqslant s}),$$

where $\boldsymbol{\mu}_{\leqslant s}$ denotes an appropriate probability measure on U_s. By considering $I_s(g_s)$ instead of $I(g)$, we have introduced a dimension truncation error

$$\|I(g) - I_s(g_s)\|_{\mathcal{X}}.$$

In many practical problems involving partial differential equations (PDEs), such as optimal control problems subject to PDEs with uncertain coefficients, there are also other sources of errors: for example, the integral operator I_s may need to be approximated by a cubature rule $Q_{s,n}$ with n nodes and, in practice, we may only have access to, e.g., a finite element approximation $g_{s,h}$ of the solution to (5.2) in some finite-dimensional subspace $\mathcal{X}_h$ of $\mathcal{X}$. The overall error can typically be estimated via an error decomposition of the form

$$\|I(g) - Q_{s,n}(g_{s,h})\|_{\mathcal{X}} \leqslant \|I(g) - I_s(g_s)\|_{\mathcal{X}} + \|I_s(g_s - g_{s,h})\|_{\mathcal{X}} + \|I_s(g_{s,h}) - Q_{s,n}(g_{s,h})\|_{\mathcal{X}},$$

where the last two terms correspond to finite element discretization error and cubature error, which are analyzed in the following chapters.

Dimension truncation error rates are typically studied in the setting of elliptic PDEs with random coefficients. In [116] the authors derive a dimension truncation rate for the elliptic PDE problem in conjunction with an affine parameterization of the uncertain diffusion coefficient, see Section 4.1. This result was improved by [53], where dimension truncation in the context of affine parametric operator equations is studied. Dimension truncation has also been analyzed for coupled PDE systems arising in optimal control problems under uncertainty [76] as well as in the context of the so-called "periodic model" of uncertainty quantification for both numerical integration [94] and kernel interpolation [95]. A common feature in these works is the use of Neumann series, which is a suitable tool for dimension truncation analysis provided that the uncertain parameters affinely enter the PDE. However, when the dependence of the PDE operator on the random variables is sufficiently nonlinear, the Neumann approach may produce only suboptimal dimension truncation rates: this is the case for lognormally parameterized diffusion coefficients for elliptic PDEs [70] or when the quantity of interest is a nonlinear functional of the PDE response [41, 86].

In contrast to the Neumann series approach, the use of Taylor series was considered in [60] to obtain dimension truncation error rates in the context of a spectral eigenvalue problem for an elliptic PDE with a random coefficient. The use of Taylor series allows to exploit the underlying parametric regularity of the model problem in order to derive dimension truncation rates, as opposed to Neumann series which is fundamentally dependent on the parametric structure of the PDE problem. Motivated by the paper [60], a similar Taylor series approach is used to derive a dimension truncation rate for a smooth nonlinear quantity of interest subject to an affine parametric parabolic PDE in [77].

The available literature provides some numerical evidence concerning dimension truncation rates for nonlinear parameterizations of PDE problems. For instance, [52] contains numerical experiments suggesting that the dimension truncation error rates for certain non-affine parametric PDE problems are significantly better than the theoretical bounds derived using the Neumann series approach.Furthermore, it is known in the context of

lognormal parameterizations for diffusion coefficients of parametric elliptic PDEs that the use of special Matérn covariances can yield even exponentially convergent dimension truncation errors (cf. [22, Section 7.2] and [51, 148]).

5.1 Problem setting

Let $g(\boldsymbol{y})$ be an element of a separable Banach space $\mathcal{X}$ for each $\boldsymbol{y} \in U_{\boldsymbol{\alpha}}$, where

$$U_{\boldsymbol{\alpha}} := \left\{ \boldsymbol{y} \in \mathbb{R}^{\mathbb{N}} : \sum_{j \geqslant 1} \alpha_j |y_j| < \infty \right\}$$

for a given sequence $\boldsymbol{\alpha} := (\alpha_j)_{j \geqslant 1} \in \ell^1(\mathbb{N})$ such that $\alpha_j \in [0, \infty)$ for all $j \in \mathbb{N}$.
Let us define $g_s(\boldsymbol{y}) := g(\boldsymbol{y}_{\leqslant s}, \boldsymbol{0}) := g(y_1, \ldots, y_s, 0, 0, \ldots)$. We consider the dimension truncation error

$$\left\| \int_{\mathbb{R}^{\mathbb{N}}} (g(\boldsymbol{y}) - g_s(\boldsymbol{y}))\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \right\|_{\mathcal{X}},$$

where

$$\boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) := \bigotimes_{j \geqslant 1} \mathcal{N}_\beta(0,1) \tag{5.3}$$

and $\mathcal{N}_\beta(0,1)$ denotes the univariate β-Gaussian distribution with density

$$\varphi_\beta(y) := \frac{1}{2\beta^{\frac{1}{\beta}}\Gamma(1+\frac{1}{\beta})} \mathrm{e}^{-\frac{|y|^\beta}{\beta}}, \quad y \in \mathbb{R},$$

where we restrict to the case $\beta \geqslant 1$. Importantly, in the case $\beta = 2$ the probability measure (5.3) is Gaussian and in the case $\beta = 1$ it corresponds to the Laplace distribution. Formally, the case $\beta = \infty$ corresponds to the uniform probability measure on $[-1,1]^{\mathbb{N}}$, which we denote by

$$\boldsymbol{\gamma}(\mathrm{d}\boldsymbol{y}) := \bigotimes_{j \geqslant 1} \frac{\mathrm{d}y}{2}.$$

To this end, we will consider dimension truncation subject to β-Gaussian probability measures and the uniform probability measure, equipped with their respective sets of assumptions.

β-Gaussian probability measures In the β-Gaussian setting, we will work under the following assumptions:

(A1) It holds for a.e. $\boldsymbol{y} \in U_{\boldsymbol{\alpha}}$ that

$$\|g(\boldsymbol{y}) - g_s(\boldsymbol{y})\|_{\mathcal{X}} \to 0 \quad \text{as } s \to \infty.$$

(A2) Let $(\Theta_k)_{k \geqslant 0}$ be a sequence of nonnegative numbers, let $\boldsymbol{b} := (b_j)_{j \geqslant 1} \in \ell^p(\mathbb{N})$ for some $p \in (0,1)$, and let $b_1 \geqslant b_2 \geqslant \cdots \geqslant 0$. We assume that the integrand g is continuously differentiable up to order $k+1$, with

$$\|\partial^{\boldsymbol{\nu}} g(\boldsymbol{y})\|_{\mathcal{X}} \leqslant \Theta_{|\boldsymbol{\nu}|} \boldsymbol{b}^{\boldsymbol{\nu}} \prod_{j \geqslant 1} \mathrm{e}^{\alpha_j |y_j|} < \infty$$

for all $\boldsymbol{y} \in U_{\boldsymbol{\alpha}}$ and all $\boldsymbol{\nu} \in \mathcal{F}_k := \{\boldsymbol{\nu} \in \mathbb{N}_0^{\mathbb{N}} : |\boldsymbol{\nu}| \leqslant k+1\}$, where $k := \lceil \frac{1}{1-p} \rceil$. In the case $\beta = 1$, we assume in addition that $\alpha_j < 1$ for all $j \in \mathbb{N}$.

Uniform probability measure In this setting, we suppose that $g(\boldsymbol{y}) \in \mathcal{X}$ for each $\boldsymbol{y} \in [-1,1]^{\mathbb{N}}$ and we will work under the following assumptions:

(A1') It holds for a.e. $\boldsymbol{y} \in [-1,1]^{\mathbb{N}}$ that

$$\|g(\boldsymbol{y}) - g_s(\boldsymbol{y})\|_{\mathcal{X}} \to 0 \quad \text{as } s \to \infty.$$

(A2') Let $(\Theta_k)_{k\geqslant 0}$ be a sequence of nonnegative numbers, let $\boldsymbol{b} := (b_j)_{j\geqslant 1} \in \ell^p(\mathbb{N})$ for some $p \in (0,1)$, and let $b_1 \geqslant b_2 \geqslant \cdots \geqslant 0$. We assume that the integrand g is continuously differentiable up to order $k+1$, with

$$\|\partial^{\boldsymbol{\nu}} g(\boldsymbol{y})\|_{\mathcal{X}} \leqslant \Theta_{|\boldsymbol{\nu}|} \boldsymbol{b}^{\boldsymbol{\nu}}$$

for all $\boldsymbol{y} \in [-1,1]^{\mathbb{N}}$ and all $\boldsymbol{\nu} \in \mathcal{F}_k := \{\boldsymbol{\nu} \in \mathbb{N}_0^{\mathbb{N}} : |\boldsymbol{\nu}| \leqslant k+1\}$, where $k := \lceil \frac{1}{1-p} \rceil$.

5.2 Infinite-dimensional integration

If (A2) holds, then $g(\boldsymbol{y}) \in \mathcal{X}$ for each $\boldsymbol{y} \in U_{\boldsymbol{\alpha}}$ and we infer that $\boldsymbol{y} \mapsto G(g(\boldsymbol{y}))$ for all $G \in \mathcal{X}'$ is continuous as a composition of continuous mappings. Hence $\boldsymbol{y} \mapsto G(g(\boldsymbol{y}))$ is measurable for all $G \in \mathcal{X}'$, i.e., $\boldsymbol{y} \mapsto g(\boldsymbol{y})$ is weakly measurable. Since $\mathcal{X}$ is assumed to be a separable Banach space, by Pettis' theorem (cf., e.g., [164, Chapter 4]) we obtain that $\boldsymbol{y} \mapsto g(\boldsymbol{y})$ is strongly measurable. The $\boldsymbol{\mu}_\beta$-integrability of the upper bound in (A2) is proved in [86, Proposition 3.2] for $\beta \in [1,2]$ and can be proved mutatis mutandis for $\beta > 2$. Thus we conclude from Bochner's theorem (cf., e.g., [164, Chapter 5]) and (A2) that g is $\boldsymbol{\mu}_\beta$-integrable over $U_{\boldsymbol{\alpha}}$. Bochner's theorem can also be used to ensure that a function $g(\boldsymbol{y}) \in \mathcal{X}$, $\boldsymbol{y} \in [-1,1]^{\mathbb{N}}$, is $\boldsymbol{\gamma}$-integrable provided that (A2') holds.
The following lemma has been adapted from [65, Lemma 2.28] to our setting.

Lemma 5.2.1. *It holds that $U_{\boldsymbol{\alpha}} \in \mathcal{B}(\mathbb{R}^{\mathbb{N}})$, where $\mathcal{B}$ denotes the Borel σ-algebra and $\boldsymbol{\mu}_\beta(U_{\boldsymbol{\alpha}}) = 1$.*

Proof. The first statement follows from

$$U_{\boldsymbol{\alpha}} = \bigcup_{N\geqslant 1} \bigcap_{M\geqslant 1} \left\{ \boldsymbol{y} \in \mathbb{R}^{\mathbb{N}} : \sum_{1\leqslant j\leqslant M} \alpha_j |y_j| \leqslant N \right\}.$$

By the monotone convergence theorem, we deduce that

$$\int_{\mathbb{R}^{\mathbb{N}}} \sum_{j\geqslant 1} \alpha_j |y_j| \, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) = \sum_{j\geqslant 1} \alpha_j \int_{\mathbb{R}^{\mathbb{N}}} |y_j| \, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) = \frac{\Gamma(\frac{2}{\beta})}{\beta^{1-\frac{1}{\beta}}\Gamma(1+\frac{1}{\beta})} \sum_{j\geqslant 1} \alpha_j < \infty,$$

for all $\beta > 0$, where we used [68, formula 3.326.2]. □

From the above lemma we conclude that we can restrict to $\boldsymbol{y} \in U_{\boldsymbol{\alpha}}$ since

$$\int_{\mathbb{R}^{\mathbb{N}}} g(\boldsymbol{y}) \, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) = \int_{U_{\boldsymbol{\alpha}}} g(\boldsymbol{y}) \, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}),$$

for any g satisfying (A2). Thus in the β-Gaussian setting, the domain of integration $\mathbb{R}^{\mathbb{N}}$ is interchangeable with $U_{\boldsymbol{\alpha}}$.

Lemma 5.2.2 ([118, Theorem 1] and [81, Section 26])**.** *From Lebesgue's dominated convergence theorem, we infer the following results.*

(i) Let $F\colon \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ be a $\boldsymbol{\mu}_\beta$-integrable function, which satisfies

$$\lim_{s\to\infty} F(\boldsymbol{y}_{\leqslant s}, \mathbf{0}) = F(\boldsymbol{y}) \quad \text{for a.e. } \boldsymbol{y}$$

and

$$|F(\boldsymbol{y}_{\leqslant s}, \mathbf{0})| \leqslant |h(\boldsymbol{y})| \quad \text{for a.e. } \boldsymbol{y}$$

for some $\boldsymbol{\mu}_\beta$-integrable h. Then

$$\lim_{s\to\infty} \int_{\mathbb{R}^s} F(\boldsymbol{y}_{\leqslant s}, \mathbf{0})\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}_{\leqslant s}) = \int_{\mathbb{R}^{\mathbb{N}}} F(\boldsymbol{y})\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}).$$

(ii) Let $F\colon [-1,1]^{\mathbb{N}} \to \mathbb{R}$ be a $\boldsymbol{\gamma}$-integrable function, which satisfies

$$\lim_{s\to\infty} F(\boldsymbol{y}_{\leqslant s}, \mathbf{0}) = F(\boldsymbol{y}) \quad \text{for a.e. } \boldsymbol{y}$$

and

$$|F(\boldsymbol{y}_{\leqslant s}, \mathbf{0})| \leqslant |h(\boldsymbol{y})| \quad \text{for a.e. } \boldsymbol{y}$$

for some $\boldsymbol{\gamma}$-integrable h. Then

$$\lim_{s\to\infty} \int_{[-1,1]^s} F(\boldsymbol{y}_{\leqslant s}, \mathbf{0})\, \boldsymbol{\gamma}(\mathrm{d}\boldsymbol{y}_{\leqslant s}) = \int_{[-1,1]^{\mathbb{N}}} F(\boldsymbol{y})\, \boldsymbol{\gamma}(\mathrm{d}\boldsymbol{y}).$$

In the case where $F(\boldsymbol{y}) := \langle G, g(\boldsymbol{y})\rangle_{\mathcal{X}',\mathcal{X}}$ for some G in the topological dual space of $\mathcal{X}$, it holds that

$$F(\boldsymbol{y}) - F(\boldsymbol{y}_{\leqslant s}, \mathbf{0}) = \langle G, g(\boldsymbol{y}) - g(\boldsymbol{y}_{\leqslant s}, \mathbf{0})\rangle_{\mathcal{X}',\mathcal{X}} \leqslant \|G\|_{\mathcal{X}'} \|g(\boldsymbol{y}) - g(\boldsymbol{y}_{\leqslant s}, \mathbf{0})\|_{\mathcal{X}},$$

and

$$|F(\boldsymbol{y}_{\leqslant s}, \mathbf{0})| \leqslant \|G\|_{\mathcal{X}'} \|g(\boldsymbol{y}_{\leqslant s}, \mathbf{0})\|_{\mathcal{X}},$$

which can be bounded by taking $\boldsymbol{\nu} = \mathbf{0}$ in (A2) or (A2'), respectively. Thus, the preceding result holds due to (A1) in the β-Gaussian setting and due to (A1') in the uniform setting.

5.3 Dimension truncation error

The following lemma is commonly used in the analysis of best N-term approximations (cf., e.g., [37]), and it will be highly useful in our treatment of the dimension truncation error.

Lemma 5.3.1 (Stechkin's lemma)**.** *Let $0 < p \leqslant q < \infty$ and let $(a_k)_{k\geqslant 1}$ be a sequence of real numbers ordered such that $|a_1| \geqslant |a_2| \geqslant \cdots$. Then*

$$\left(\sum_{k>N}^{\infty} |a_k|^q\right)^{\frac{1}{q}} \leqslant N^{-\frac{1}{p}+\frac{1}{q}} \left(\sum_{k\geqslant 1} |a_k|^p\right)^{\frac{1}{p}}.$$

Proof. For an elementary proof of this result, see, e.g., [110, Lemma 3.3]. □

The main result about the dimension truncation error is given below.

Theorem 5.3.2. *Suppose that assumptions (A1) and (A2) hold. Then*

$$\left\| \int_{\mathbb{R}^{\mathbb{N}}} (g(\boldsymbol{y}) - g_s(\boldsymbol{y})) \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \right\|_{\mathcal{X}} \leqslant C s^{-\frac{2}{p}+1},$$

where the constant $C > 0$ is independent of the dimension s.
Let $G \in \mathcal{X}'$ be arbitrary. Then

$$\left| \int_{\mathbb{R}^{\mathbb{N}}} G(g(\boldsymbol{y}) - g_s(\boldsymbol{y})) \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \right| \leqslant C s^{-\frac{2}{p}+1},$$

where the constant $C > 0$ is independent of the dimension s.

Proof. Let s^* be the smallest integer such that $\sum_{j>s^*} b_j \leqslant \frac{1}{2}$. Clearly,

$$\left\| \int_{\mathbb{R}^{\mathbb{N}}} (g(\boldsymbol{y}) - g_s(\boldsymbol{y})) \, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \right\|_{\mathcal{X}} \leqslant 2\Theta_0 \prod_{j\geqslant 1} \int_{\mathbb{R}} \mathrm{e}^{\alpha_j |y_j|} \varphi_\beta(y_j) \, \mathrm{d}y_j =: C_{\mathrm{init}} < \infty,$$

where we used Lemma 5.2.2 and Fubini's theorem. Therefore

$$\left\| \int_{\mathbb{R}^{\mathbb{N}}} (g(\boldsymbol{y}) - g_s(\boldsymbol{y})) \, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \right\|_{\mathcal{X}} \leqslant \frac{C_{\mathrm{init}}}{(s^*)^{-\frac{2}{p}+1}} s^{-\frac{2}{p}+1} \quad \text{for all } 1 \leqslant s \leqslant s^*.$$

Thus it is enough to prove the claim for sufficiently large s. In what follows, we assume that $s > s^*$, let $G \in \mathcal{X}'$ be arbitrary, and define

$$F(\boldsymbol{y}) := \langle G, g(\boldsymbol{y}) \rangle_{\mathcal{X}',\mathcal{X}} \quad \text{for all } \boldsymbol{y} \in U_{\boldsymbol{\alpha}}.$$

Let k be specified as in (A2) (note that it always holds that $k \geqslant 2$). Then

$$\partial^{\boldsymbol{\nu}} F(\boldsymbol{y}) = \langle G, \partial^{\boldsymbol{\nu}} g(\boldsymbol{y}) \rangle_{\mathcal{X}',\mathcal{X}} \quad \text{for all } \boldsymbol{\nu} \in \mathcal{F}_k \text{ and } \boldsymbol{y} \in U_{\boldsymbol{\alpha}}$$

and it follows from our assumptions that

$$|\partial^{\boldsymbol{\nu}} F(\boldsymbol{y})| \leqslant \|G\|_{\mathcal{X}'} \Theta_{|\boldsymbol{\nu}|} \boldsymbol{b}^{\boldsymbol{\nu}} \prod_{j\geqslant 1} \mathrm{e}^{\alpha_j |y_j|} \quad \text{for all } \boldsymbol{\nu} \in \mathcal{F}_k \text{ and } \boldsymbol{y} \in U_{\boldsymbol{\alpha}}.$$

Let $\boldsymbol{y} \in U_{\boldsymbol{\alpha}}$ be arbitrary. We develop the Taylor expansion around the point $(\boldsymbol{y}_{\leqslant s}, \boldsymbol{0})$, which yields

$$\begin{aligned} F(\boldsymbol{y}) = F(\boldsymbol{y}_{\leqslant s}, \boldsymbol{0}) &+ \sum_{\ell=1}^{k} \sum_{\substack{|\boldsymbol{\nu}|=\ell \\ \nu_j=0 \ \forall j\leqslant s}} \frac{\boldsymbol{y}^{\boldsymbol{\nu}}}{\boldsymbol{\nu}!} \partial^{\boldsymbol{\nu}} F(\boldsymbol{y}_{\leqslant s}, \boldsymbol{0}) \\ &+ \sum_{\substack{|\boldsymbol{\nu}|=k+1 \\ \nu_j=0 \ \forall j\leqslant s}} \frac{k+1}{\boldsymbol{\nu}!} \boldsymbol{y}^{\boldsymbol{\nu}} \int_0^1 (1-\tau)^k \partial^{\boldsymbol{\nu}} F(\boldsymbol{y}_{\leqslant s}, \tau \boldsymbol{y}_{>s}) \, \mathrm{d}\tau. \end{aligned}$$

Rearranging this equation and integrating both sides against the β-Gaussian product measure yields

$$\begin{aligned}
&\int_{\mathbb{R}^{\mathbb{N}}} (F(\boldsymbol{y}) - F(\boldsymbol{y}_{\leqslant s}, \boldsymbol{0}))\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \\
&= \sum_{\ell=1}^{k} \sum_{\substack{|\boldsymbol{\nu}|=\ell \\ \nu_j=0 \ \forall j\leqslant s}} \frac{1}{\boldsymbol{\nu}!} \int_{\mathbb{R}^{\mathbb{N}}} \boldsymbol{y}^{\boldsymbol{\nu}} \partial^{\boldsymbol{\nu}} F(\boldsymbol{y}_{\leqslant s}, \boldsymbol{0})\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \\
&\quad + \sum_{\substack{|\boldsymbol{\nu}|=k+1 \\ \nu_j=0 \ \forall j\leqslant s}} \frac{k+1}{\boldsymbol{\nu}!} \int_{\mathbb{R}^{\mathbb{N}}} \int_0^1 (1-\tau)^k \boldsymbol{y}^{\boldsymbol{\nu}} \partial^{\boldsymbol{\nu}} F(\boldsymbol{y}_{\leqslant s}, \tau \boldsymbol{y}_{>s})\, \mathrm{d}\tau\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}).
\end{aligned}$$

If there exists a single component $\nu_k = 1$ with $k > s$, then the summand in the first term vanishes since, by Lemma 5.2.2 and Fubini's theorem,

$$\begin{aligned}
&\int_{\mathbb{R}^{\mathbb{N}}} \boldsymbol{y}^{\boldsymbol{\nu}} \partial^{\boldsymbol{\nu}} F(\boldsymbol{y}_{\leqslant s}, \boldsymbol{0}) \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \\
&= \left(\int_{\mathbb{R}^s} \partial^{\boldsymbol{\nu}} F(\boldsymbol{y}_{\leqslant s}, \boldsymbol{0}) \prod_{j=1}^{s} \varphi_\beta(y_j)\, \mathrm{d}\boldsymbol{y}_{\leqslant s} \right) \underbrace{\left(\int_{\mathbb{R}} y_k \varphi_\beta(y_k)\, \mathrm{d}y_k \right)}_{=0} \left(\int_{\mathbb{R}^{\mathbb{N}}} \boldsymbol{y}^{\boldsymbol{\nu}} \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}_{\{s+1:\infty\}\setminus\{k\}}) \right) \\
&= 0.
\end{aligned}$$

Hence

$$\begin{aligned}
&\left| \int_{\mathbb{R}^{\mathbb{N}}} (F(\boldsymbol{y}) - F(\boldsymbol{y}_{\leqslant s}, \boldsymbol{0}))\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \right| \\
&\leqslant \sum_{\ell=2}^{k} \sum_{\substack{|\boldsymbol{\nu}|=\ell \\ \nu_j=0 \ \forall j\leqslant s \\ \nu_j\neq 1 \ \forall j>s}} \frac{1}{\boldsymbol{\nu}!} \int_{\mathbb{R}^{\mathbb{N}}} |\boldsymbol{y}^{\boldsymbol{\nu}}| \cdot |\partial^{\boldsymbol{\nu}} F(\boldsymbol{y}_{\leqslant s}, \boldsymbol{0})|\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \qquad (5.4) \\
&\quad + \sum_{\substack{|\boldsymbol{\nu}|=k+1 \\ \nu_j=0 \ \forall j\leqslant s}} \frac{k+1}{\boldsymbol{\nu}!} \int_{\mathbb{R}^{\mathbb{N}}} \int_0^1 (1-\tau)^k |\boldsymbol{y}^{\boldsymbol{\nu}}| \cdot |\partial^{\boldsymbol{\nu}} F(\boldsymbol{y}_{\leqslant s}, \tau \boldsymbol{y}_{>s})|\, \mathrm{d}\tau\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}). \qquad (5.5)
\end{aligned}$$

We start our estimation by splitting the terms in (5.4):

$$\begin{aligned}
&\int_{\mathbb{R}^{\mathbb{N}}} |\boldsymbol{y}^{\boldsymbol{\nu}}| \cdot |\partial^{\boldsymbol{\nu}} F(\boldsymbol{y}_{\leqslant s}, \boldsymbol{0})|\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \leqslant \|G\|_{\mathcal{X}'} \Theta_{|\boldsymbol{\nu}|} \boldsymbol{b}^{\boldsymbol{\nu}} \int_{\mathbb{R}^{\mathbb{N}}} |\boldsymbol{y}^{\boldsymbol{\nu}}| \prod_{j=1}^{s} \mathrm{e}^{\alpha_j |y_j|} \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \\
&\leqslant \|G\|_{\mathcal{X}'} \Theta_{|\boldsymbol{\nu}|} \boldsymbol{b}^{\boldsymbol{\nu}} \int_{\mathbb{R}^{\mathbb{N}}} |\boldsymbol{y}^{\boldsymbol{\nu}}| \prod_{j\geqslant 1} \mathrm{e}^{\alpha_j |y_j|} \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \\
&= \|G\|_{\mathcal{X}'} \Theta_{|\boldsymbol{\nu}|} \boldsymbol{b}^{\boldsymbol{\nu}} \Bigg(\underbrace{\prod_{j\in \mathrm{supp}(\boldsymbol{\nu})} \int_{\mathbb{R}} |y_j|^{\nu_j} \mathrm{e}^{\alpha_j |y_j|} \varphi_\beta(y_j) \mathrm{d}y_j}_{=:\mathrm{term}_1} \Bigg) \Bigg(\underbrace{\prod_{j\notin \mathrm{supp}(\boldsymbol{\nu})} \int_{\mathbb{R}} \mathrm{e}^{\alpha_j |y_j|} \varphi_\beta(y_j) \mathrm{d}y_j}_{=:\mathrm{term}_2} \Bigg),
\end{aligned}$$

where the final step follows from Lemma 5.2.2 and Fubini's theorem. In order to bound term_1, note that

$$C_{\alpha_j,\beta,\nu_j} := \int_{\mathbb{R}} |y_j|^{\nu_j} \mathrm{e}^{\alpha_j |y_j|} \varphi_\beta(y_j)\, \mathrm{d}y_j < \infty$$

since we assumed that $\beta \geqslant 1$ and $\alpha_j < 1$ in the case $\beta = 1$. We define an auxiliary constant $A_{\alpha_j,\beta,\ell} := \max_{1\leqslant k\leqslant \ell} C_{\alpha_j,\beta,k}$. Clearly, $C_{\alpha_j,\beta,\nu_j} \leqslant A_{\alpha_j,\beta,|\boldsymbol{\nu}|} < \infty$ and $C_{\alpha,\beta,\nu} \leqslant C_{\alpha',\beta,\nu}$ whenever $\alpha \leqslant \alpha'$ with β, ν fixed. In particular,

$$\begin{aligned}\mathrm{term}_1 &\leqslant \prod_{j\in\mathrm{supp}(\boldsymbol{\nu})} C_{\alpha_j,\beta,\nu_j} \leqslant \prod_{j\in\mathrm{supp}(\boldsymbol{\nu})} C_{\|\boldsymbol{\alpha}\|_\infty,\beta,\nu_j} \leqslant \prod_{j\in\mathrm{supp}(\boldsymbol{\nu})} A_{\|\boldsymbol{\alpha}\|_\infty,\beta,|\boldsymbol{\nu}|} \\ &\leqslant \max\{1, A_{\|\boldsymbol{\alpha}\|_\infty,\beta,|\boldsymbol{\nu}|}\}^{|\boldsymbol{\nu}|},\end{aligned}$$

where $\|\boldsymbol{\alpha}\|_\infty := \sup_{j\geqslant 1} |\alpha_j|$ is finite since $\boldsymbol{\alpha} \in \ell^1(\mathbb{N})$ by assumption. To bound term_2, we note that there is an index $j' \in \mathbb{N}$ such that $\alpha_j \leqslant \frac{1}{2}$ for all $j > j'$. Hence

$$\begin{aligned}\mathrm{term}_2 &= \Bigg(\prod_{\substack{j\notin\mathrm{supp}(\boldsymbol{\nu})\\ 1\leqslant j\leqslant j'}} \int_{\mathbb{R}} \mathrm{e}^{\alpha_j|y_j|}\varphi_\beta(y_j)\,\mathrm{d}y_j\Bigg)\Bigg(\prod_{\substack{j\notin\mathrm{supp}(\boldsymbol{\nu})\\ j>j'}} \int_{\mathbb{R}} \mathrm{e}^{\alpha_j|y_j|}\varphi_\beta(y_j)\,\mathrm{d}y_j\Bigg) \\ &\leqslant \max\{1, C_{\|\boldsymbol{\alpha}\|_\infty,\beta,0}\}^{j'}\Bigg(\prod_{\substack{j\notin\mathrm{supp}(\boldsymbol{\nu})\\ j>j'}} \int_{\mathbb{R}} \mathrm{e}^{\alpha_j|y_j|}\varphi_\beta(y_j)\,\mathrm{d}y_j\Bigg),\end{aligned}$$

using a similar argument as before. In order to ensure that the remaining factor is finite, we argue similarly to [86, Proposition 3.2].
Let us first consider the case $\beta > 1$. Young's inequality states, for all $x, y \geqslant 0$ and $\theta \in (0,1)$, that

$$xy = x^\theta x^{1-\theta} y \leqslant \frac{\beta-1}{\beta} x^{\theta\frac{\beta}{\beta-1}} + \frac{1}{\beta} x^{(1-\theta)\beta} y^\beta,$$

where $\beta > 1$. The special choice $\theta = \frac{\beta-1}{\beta}$ yields

$$xy \leqslant \frac{\beta-1}{\beta} x + \frac{1}{\beta} x y^\beta \quad \text{for all } x, y \geqslant 0.$$

Thereby

$$\begin{aligned}\int_{\mathbb{R}} \mathrm{e}^{\alpha_j|y_j|}\varphi_\beta(y_j)\,\mathrm{d}y_j &\leqslant \mathrm{e}^{\frac{\beta-1}{\beta}\alpha_j} \int_{\mathbb{R}} \mathrm{e}^{\frac{\alpha_j}{\beta}|y_j|^\beta}\varphi_\beta(y_j)\,\mathrm{d}y_j \\ &= \frac{\mathrm{e}^{\frac{\beta-1}{\beta}\alpha_j}}{2\beta^{\frac{1}{\beta}}\Gamma(1+\frac{1}{\beta})} \int_{\mathbb{R}} \mathrm{e}^{-(1-\alpha_j)\frac{|y_j|^\beta}{\beta}}\,\mathrm{d}y_j = \frac{\mathrm{e}^{\frac{\beta-1}{\beta}\alpha_j}}{(1-\alpha_j)^{\frac{1}{\beta}}},\end{aligned}$$

where we used $\int_{\mathbb{R}} \mathrm{e}^{-(1-\alpha_j)\frac{|y_j|^\beta}{\beta}}\,\mathrm{d}y_j = 2\beta^{\frac{1}{\beta}}(1-\alpha_j)^{-\frac{1}{\beta}}\Gamma(1+\frac{1}{\beta})$. Furthermore, since

$$\frac{1}{1-x} = 1 + \frac{x}{1-x} \leqslant \exp\left(\frac{x}{1-x}\right) \quad \text{for all } x \in [0,1),$$

we obtain

$$\int_{\mathbb{R}} \mathrm{e}^{\alpha_j|y_j|}\varphi_\beta(y_j)\,\mathrm{d}y_j \leqslant \exp\left(\frac{\beta-1}{\beta}\alpha_j\right)\exp\left(\frac{1}{\beta}\frac{\alpha_j}{1-\alpha_j}\right) \leqslant \exp\left(\frac{\beta+1}{\beta}\alpha_j\right)$$

since we assumed $\alpha_j \leqslant \frac{1}{2}$ for all $j > j'$. Therefore

$$\prod_{\substack{j\notin\mathrm{supp}(\boldsymbol{\nu})\\ j>j'}} \int_{\mathbb{R}} \mathrm{e}^{\alpha_j|y_j|}\varphi_\beta(y_j)\,\mathrm{d}y_j \leqslant \exp\left(\frac{\beta+1}{\beta}\sum_{\substack{j\notin\mathrm{supp}(\boldsymbol{\nu})\\ j>j'}} \alpha_j\right) \leqslant \exp\left(\frac{\beta+1}{\beta}\sum_{j\geqslant 1}\alpha_j\right) =: \widetilde{C}_\beta,$$

where $\widetilde{C}_\beta < \infty$ since $\boldsymbol{\alpha} \in \ell^1(\mathbb{N})$. The special case $\beta = 1$ follows since we assumed $\alpha_j < 1$ for all $j \geqslant 1$ and it holds that

$$\int_{\mathbb{R}} \mathrm{e}^{\alpha_j |y_j|} \varphi_1(y_j)\,\mathrm{d}y_j = \frac{1}{1-\alpha_j} \leqslant \exp\left(\frac{\alpha_j}{1-\alpha_j}\right) \leqslant \exp(2\alpha_j)$$

for $j > j'$. Thus

$$\prod_{\substack{j \notin \mathrm{supp}(\boldsymbol{\nu}) \\ j > j'}} \int_{\mathbb{R}} \mathrm{e}^{\alpha_j |y_j|} \varphi_1(y_j)\,\mathrm{d}y_j \leqslant \exp\left(2 \sum_{\substack{j \notin \mathrm{supp}(\boldsymbol{\nu}) \\ j > j'}} \alpha_j\right) \leqslant \exp\left(2 \sum_{j \geqslant 1} \alpha_j\right) =: \widetilde{C}_1 < \infty$$

since $\boldsymbol{\alpha} \in \ell^1(\mathbb{N})$. Combining the estimates for term_1 and term_2 gives

$$\begin{aligned}
&\int_{\mathbb{R}^{\mathbb{N}}} |\boldsymbol{y}^{\boldsymbol{\nu}}| \cdot |\partial^{\boldsymbol{\nu}} F(\boldsymbol{y}_{\leqslant s}, \mathbf{0})|\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \\
&\leqslant \widetilde{C}_\beta \|G\|_{\mathcal{X}'} \max\{1, C_{\|\boldsymbol{\alpha}\|_\infty,\beta,0}\}^{j'} \Theta_{|\boldsymbol{\nu}|} \boldsymbol{b}^{\boldsymbol{\nu}} \max\{1, A_{\|\alpha\|_\infty,\beta,|\boldsymbol{\nu}|}\}^{|\boldsymbol{\nu}|}.
\end{aligned}$$

Similarly we split the terms in (5.5),

$$\begin{aligned}
&\int_{\mathbb{R}^{\mathbb{N}}} \int_0^1 (1-\tau)^k |\boldsymbol{y}^{\boldsymbol{\nu}}| \cdot |\partial^{\boldsymbol{\nu}} F(\boldsymbol{y}_{\leqslant s}, \tau \boldsymbol{y}_{>s})|\,\mathrm{d}\tau\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \\
&\leqslant \|G\|_{\mathcal{X}'} \Theta_{|\boldsymbol{\nu}|} \boldsymbol{b}^{\boldsymbol{\nu}} \Bigg(\prod_{j \in \mathrm{supp}(\boldsymbol{\nu})} \int_{\mathbb{R}} |y_j|^{\nu_j} \mathrm{e}^{\alpha_j |y_j|} \varphi_\beta(y_j) \mathrm{d}y_j \Bigg) \Bigg(\prod_{j \notin \mathrm{supp}(\boldsymbol{\nu})} \int_{\mathbb{R}} \mathrm{e}^{\alpha_j |y_j|} \varphi_\beta(y_j) \mathrm{d}y_j \Bigg) \\
&\leqslant \widetilde{C}_\beta \|G\|_{\mathcal{X}'} \Theta_{|\boldsymbol{\nu}|} \max\{1, C_{\|\boldsymbol{\alpha}\|_\infty,\beta,0}\}^{j'} \boldsymbol{b}^{\boldsymbol{\nu}} \max\{1, A_{\|\boldsymbol{\alpha}\|_\infty,\beta,|\boldsymbol{\nu}|}\}^{|\boldsymbol{\nu}|},
\end{aligned}$$

where we used

$$\begin{aligned}
|\partial^{\boldsymbol{\nu}} F(\boldsymbol{y}_{\leqslant s}, \tau \boldsymbol{y}_{>s})| &\leqslant \|G\|_{\mathcal{X}'} \Theta_{|\boldsymbol{\nu}|} \boldsymbol{b}^{\boldsymbol{\nu}} \Bigg(\prod_{j=1}^{s} \mathrm{e}^{\alpha_j |y_j|} \Bigg) \Bigg(\prod_{j > s} \mathrm{e}^{\tau \alpha_j |y_j|} \Bigg) \\
&\leqslant \|G\|_{\mathcal{X}'} \Theta_{|\boldsymbol{\nu}|} \boldsymbol{b}^{\boldsymbol{\nu}} \Bigg(\prod_{j \geqslant 1} \mathrm{e}^{\alpha_j |y_j|} \Bigg).
\end{aligned}$$

These inequalities allow us to estimate

$$\begin{aligned}
&\left| \int_{\mathbb{R}^{\mathbb{N}}} (F(\boldsymbol{y}) - F(\boldsymbol{y}_{\leqslant s}, \mathbf{0}))\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \right| \\
&\leqslant \widetilde{C}_\beta \|G\|_{\mathcal{X}'} \max\{1, C_{\|\boldsymbol{\alpha}\|_\infty,\beta,0}\}^{j'} \sum_{\ell=2}^{k} \sum_{\substack{|\boldsymbol{\nu}|=\ell \\ \nu_j = 0 \ \forall j \leqslant s \\ \nu_j \neq 1 \ \forall j > s}} \frac{\Theta_\ell}{\boldsymbol{\nu}!} \boldsymbol{b}^{\boldsymbol{\nu}} \max\{1, A_{\|\boldsymbol{\alpha}\|_\infty,\beta,\ell}\}^{\ell} \\
&\quad + \widetilde{C}_\beta \|G\|_{\mathcal{X}'} \max\{1, C_{\|\boldsymbol{\alpha}\|_\infty,\beta,0}\}^{j'} \sum_{\substack{|\boldsymbol{\nu}|=k+1 \\ \nu_j = 0 \ \forall j \leqslant s}} \frac{k+1}{\boldsymbol{\nu}!} \Theta_{k+1} \boldsymbol{b}^{\boldsymbol{\nu}} \max\{1, A_{\|\boldsymbol{\alpha}\|_\infty,\beta,k+1}\}^{k+1} \\
&\leqslant \widetilde{C}_\beta \|G\|_{\mathcal{X}'} \max\{1, C_{\|\boldsymbol{\alpha}\|_\infty,\beta,0}\}^{j'} \big(\max_{2 \leqslant \ell \leqslant k} (\Theta_\ell \max\{1, A_{\|\boldsymbol{\alpha}\|_\infty,\beta,\ell}\}^{\ell}) \big) \sum_{\ell=2}^{k} \sum_{\substack{|\boldsymbol{\nu}|=\ell \\ \nu_j = 0 \ \forall j \leqslant s \\ \nu_j \neq 1 \ \forall j > s}} \boldsymbol{b}^{\boldsymbol{\nu}}
\end{aligned} \tag{5.6}$$

$$+ \widetilde{C}_\beta \|G\|_{\mathcal{X}'} \max\{1, C_{\|\boldsymbol{\alpha}\|_\infty,\beta,0}\}^{j'} \Theta_{k+1}(k+1) \max\{1, A_{\|\boldsymbol{\alpha}\|_\infty,\beta,k+1}\}^{k+1} \sum_{\substack{|\boldsymbol{\nu}|=k+1 \\ \nu_j=0 \ \forall j\leqslant s}} \boldsymbol{b}^{\boldsymbol{\nu}}. \quad (5.7)$$

To bound the term (5.6), we argue similarly to [53, Theorem 1] by noting that it follows from our definition of s^* that

$$\left(\sum_{k>s} b_k\right)^2 \leqslant \frac{1}{4} \quad \text{and} \quad b_j \leqslant \frac{1}{2} \quad \text{for all } j > s.$$

This leads us to estimate

$$\begin{aligned}
&\sum_{\ell=2}^{k} \sum_{\substack{|\boldsymbol{\nu}|=\ell \\ \nu_j=0 \ \forall j\leqslant s \\ \nu_j\neq 1 \ \forall j>s}} \boldsymbol{b}^{\boldsymbol{\nu}} = \sum_{\substack{2\leqslant|\boldsymbol{\nu}|\leqslant k \\ \nu_j=0 \ \forall j\leqslant s \\ \nu_j\neq 1 \ \forall j>s}} \boldsymbol{b}^{\boldsymbol{\nu}} \leqslant \sum_{\substack{0\neq|\boldsymbol{\nu}|_\infty\leqslant k \\ \nu_j=0 \ \forall j\leqslant s \\ \nu_j\neq 1 \ \forall j>s}} \boldsymbol{b}^{\boldsymbol{\nu}} = -1 + \prod_{j>s}\left(1 + \sum_{\ell=2}^{k} b_j^\ell\right) \\
&= -1 + \prod_{j>s}\left(1 + \frac{1-b_j^{k-1}}{1-b_j} b_j^2\right) \leqslant -1 + \prod_{j>s}\left(1 + 2b_j^2\right) \leqslant -1 + \exp\left(2\sum_{j>s} b_j^2\right) \\
&\leqslant 2(\mathrm{e}-1)\sum_{j>s} b_j^2,
\end{aligned}$$

where the final inequality is a consequence of Bernoulli's inequality $(1+x)^r \leqslant 1 + rx$ for all $0 \leqslant r \leqslant 1$ and $x \geqslant -1$. It is an immediate consequence of Lemma 5.3.1 that

$$\sum_{j>s} b_j^2 \leqslant s^{-\frac{2}{p}+1}\left(\sum_{j\geqslant 1} b_j^p\right)^{\frac{2}{p}} \quad (5.8)$$

since $\boldsymbol{b}$ was assumed to be a nonincreasing sequence such that $\boldsymbol{b} \in \ell^p(\mathbb{N})$ for some $p \in (0,1)$. We estimate the term (5.7) similarly to the approach taken in [60, Theorem 4.1]. By the trivial bound $\frac{|\boldsymbol{\nu}|!}{\boldsymbol{\nu}!} \geqslant 1$ and the multinomial theorem, we obtain

$$\sum_{\substack{|\boldsymbol{\nu}|=k+1 \\ \nu_j=0 \ \forall j\leqslant s}} \boldsymbol{b}^{\boldsymbol{\nu}} \leqslant \sum_{\substack{|\boldsymbol{\nu}|=k+1 \\ \nu_j=0 \ \forall j\leqslant s}} \frac{|\boldsymbol{\nu}|!}{\boldsymbol{\nu}!}\boldsymbol{b}^{\boldsymbol{\nu}} = \left(\sum_{j>s} b_j\right)^{k+1} \leqslant s^{(-\frac{1}{p}+1)(k+1)}\left(\sum_{j\geqslant 1} b_j^p\right)^{(k+1)/p}, \quad (5.9)$$

where the final inequality follows immediately from Lemma 5.3.1 and our assumption that $\boldsymbol{b}$ is a nonincreasing sequence such that $\boldsymbol{b} \in \ell^p(\mathbb{N})$ for some $p \in (0,1)$.
Putting the inequalities (5.8) and (5.9) together and utilizing $k = \lceil \frac{1}{1-p} \rceil$ we obtain

$$\left| \int_{\mathbb{R}^{\mathbb{N}}} (F(\boldsymbol{y}) - F(\boldsymbol{y}_{\leqslant s}, \boldsymbol{0}))\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \right| \leqslant C(s^{-\frac{2}{p}+1} + s^{(-\frac{1}{p}+1)(k+1)}) \leqslant C s^{-\frac{2}{p}+1}$$

for some constant $C > 0$ independent of s. Finally, by recalling that $F(\boldsymbol{y}) = \langle G, g(\boldsymbol{y})\rangle_{\mathcal{X}',\mathcal{X}}$ and $G \in \mathcal{X}'$ was arbitrary, we can take the supremum over $\{G \in \mathcal{X}' : \|G\|_{\mathcal{X}'} \leqslant 1\}$ to obtain

$$\begin{aligned}
&\sup_{G\in\mathcal{X}':\|G\|_{\mathcal{X}'}\leqslant 1} \left| \int_{\mathbb{R}^{\mathbb{N}}} (F(\boldsymbol{y}) - F(\boldsymbol{y}_{\leqslant s}, \boldsymbol{0}))\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \right| \\
&= \sup_{G\in\mathcal{X}':\|G\|_{\mathcal{X}'}\leqslant 1} \left| \int_{\mathbb{R}^{\mathbb{N}}} \langle G, g(\boldsymbol{y}) - g(\boldsymbol{y}_{\leqslant s}, \boldsymbol{0})\rangle_{\mathcal{X}',\mathcal{X}}\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \right|
\end{aligned}$$

$$= \sup_{G\in\mathcal{X}':\|G\|_{\mathcal{X}'}\leqslant 1} \left| \left\langle G, \int_{\mathbb{R}^{\mathbb{N}}} (g(\boldsymbol{y}) - g(\boldsymbol{y}_{\leqslant s}, \mathbf{0}))\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \right\rangle_{\mathcal{X}',\mathcal{X}} \right|$$
$$= \left\| \int_{\mathbb{R}^{\mathbb{N}}} (g(\boldsymbol{y}) - g_s(\boldsymbol{y}))\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \right\|_{\mathcal{X}} \leqslant C s^{-\frac{2}{p}+1}$$

as desired. □

We also state the corresponding dimension truncation result in the uniform case formally corresponding to $\beta = \infty$.

Theorem 5.3.3. *Suppose that assumptions (A1') and (A2') hold. Then*

$$\left\| \int_{[-1,1]^{\mathbb{N}}} (g(\boldsymbol{y}) - g_s(\boldsymbol{y}))\, \boldsymbol{\gamma}(\mathrm{d}\boldsymbol{y}) \right\|_{\mathcal{X}} \leqslant C s^{-\frac{2}{p}+1},$$

where the constant $C > 0$ is independent of the dimension s.
Let $G \in \mathcal{X}'$ be arbitrary. Then

$$\left| \int_{[-1,1]^{\mathbb{N}}} G(g(\boldsymbol{y}) - g_s(\boldsymbol{y}))\, \boldsymbol{\gamma}(\mathrm{d}\boldsymbol{y}) \right| \leqslant C s^{-\frac{2}{p}+1},$$

where the constant $C > 0$ is independent of the dimension s.

Proof. The steps are completely analogous to Theorem 5.3.2 in the special case $\alpha_j = 0$ for all $j \geqslant 1$ and by restricting the domain of integration to $[-1,1]^{\mathbb{N}}$. In the special case $\Theta_{|\boldsymbol{\nu}|} = (|\boldsymbol{\nu}| + r_1)!$, $r_1 \in \mathbb{N}_0$, the proof works as in [77, Theorem 6.2]. □

Remark. The conditions (A2) and (A2') are formulated as sufficient conditions. The form in which the regularity bounds are postulated in (A2) and (A2') is an important ingredient for the Taylor series argument. However, it is known that (A2') is not a necessary condition: an example is given in [94, Lemma 2.4], where the authors obtain the dimension truncation rate $\mathcal{O}(s^{-\frac{2}{p}+1})$ for a problem which satisfies a more general parametric regularity bound than (A2').

5.4 Application to parametric PDEs and optimal control

In this section, we illustrate how to apply the main dimension truncation results proved in Section 5.3 to parametric elliptic PDE model problems. We consider uniform and affine (see Section 4.1) as well as lognormal parameterizations of the input random field. The rate we obtain for the uniform and affine model coincides with the well-known rate in the literature [53] and is not a new result, however, we present it for completeness. Remarkably, the dimension truncation rate we obtain for the lognormal model using our method improves the rates in the existing literature (cf., e.g., [70, 113]). Finally, we give an example on how our results can be applied to assess dimension truncation rates corresponding to PDE solutions composed with nonlinear quantities of interest.
Similar to Section 4.1, we consider the problem of finding $u : D \times U \to \mathbb{R}$ such that

$$\begin{aligned} -\nabla \cdot (a(\boldsymbol{x},\boldsymbol{y})\nabla u(\boldsymbol{x},\boldsymbol{y})) &= f(\boldsymbol{x}), \quad \boldsymbol{x} \in D,\ \boldsymbol{y} \in U, \\ u(\boldsymbol{x},\boldsymbol{y}) &= 0, \qquad \boldsymbol{x} \in \partial D,\ \boldsymbol{y} \in U, \end{aligned} \tag{5.10}$$

in some bounded Lipschitz domain $D \subset \mathbb{R}^d$, $d \in \{1,2,3\}$, for some given source term $f\colon D \to \mathbb{R}$ and diffusion coefficient $a\colon D \times U \to \mathbb{R}$. The parameter set U is assumed to be a nonempty subset of $\mathbb{R}^{\mathbb{N}}$.
The relevant function spaces for the elliptic PDE problem (5.10) are $\mathcal{X} := H_0^1(D)$ and its dual $\mathcal{X}' = H^{-1}(D)$, which we understand with respect to the pivot space $H := L^2(D)$. The space H is identified with its own dual and we set $\|v\|_{\mathcal{X}} := \|\nabla v\|_H$ for $v \in \mathcal{X}$, as we did in Section 4.1. The weak formulation of (5.10) is to find, for all $\boldsymbol{y} \in U$, a solution $u(\cdot, \boldsymbol{y}) \in \mathcal{X}$ such that

$$\int_D a(\boldsymbol{x}, \boldsymbol{y}) \nabla u(\boldsymbol{x}, \boldsymbol{y}) \cdot \nabla v(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = \langle f, v \rangle_{\mathcal{X}',\mathcal{X}} \quad \text{for all } v \in \mathcal{X}, \tag{5.11}$$

where $f \in \mathcal{X}'$ and $\langle \cdot, \cdot \rangle_{\mathcal{X}',\mathcal{X}}$ denotes the duality pairing between elements of $\mathcal{X}'$ and $\mathcal{X}$.
Note that (5.11) is more general than the elliptic PDE in Section 4.1: in (5.11) we do not restrict to an affine parameter dependece (AE3) – (AE4), we allow other than uniform distributions of the parameters (AE2), and we do not restrict to uniform bounded diffusion coeffiecients (AE5). In particular, (M2) below coincides with the assumptions in Section 4.1.
The following lemma collects basic, well-known results about the existence of a unique solution to (5.11) (Lax–Milgram lemma), the continuity of the PDE solution with respect to the right-hand side of (5.11) (a priori bound), and the continuity of the PDE solution with respect to the diffusion coefficient (the second Strang lemma).

Lemma 5.4.1. *Let $D \subset \mathbb{R}^d$, $d \in \{1,2,3\}$, be a bounded Lipschitz domain, $\varnothing \neq U \subseteq \mathbb{R}^{\mathbb{N}}$, $f \in \mathcal{X}'$, and suppose that there exist $a_{\min}(\boldsymbol{y}) := \min_{\boldsymbol{x} \in \overline{D}} a(\boldsymbol{x}, \boldsymbol{y}) \in L^\infty_+(D)$ and $a_{\max}(\boldsymbol{y}) := \max_{\boldsymbol{x} \in \overline{D}} a(\boldsymbol{x}, \boldsymbol{y}) \in L^\infty_+(D)$ such that*

$$0 < a_{\min}(\boldsymbol{y}) \leqslant a(\boldsymbol{x}, \boldsymbol{y}) \leqslant a_{\max}(\boldsymbol{y}) < \infty \quad \text{for all } \boldsymbol{x} \in D \text{ and } \boldsymbol{y} \in U, \tag{5.12}$$

where $a(\cdot, \boldsymbol{y}) \in L^\infty_+(D)$, $\boldsymbol{y} \in U$, is the diffusion coefficient in (5.11). We define $a_s(\cdot, \boldsymbol{y}) := a(\cdot, (\boldsymbol{y}_{\leqslant s}, \boldsymbol{0}))$, $a^s_{\min}(\boldsymbol{y}) := a_{\min}(\boldsymbol{y}_{\leqslant s}, \boldsymbol{0})$, and $u_s(\cdot, \boldsymbol{y}) := u(\cdot, (\boldsymbol{y}_{\leqslant s}, \boldsymbol{0}))$. Then there exists a unique solution to (5.11) such that

$$\|u(\cdot, \boldsymbol{y})\|_{\mathcal{X}} \leqslant \frac{\|f\|_{\mathcal{X}'}}{a_{\min}(\boldsymbol{y})} \quad \text{for all } \boldsymbol{y} \in U \tag{5.13}$$

and

$$\|u(\cdot, \boldsymbol{y}) - u_s(\cdot, \boldsymbol{y})\|_{\mathcal{X}} \leqslant \frac{1}{a_{\min}(\boldsymbol{y}) a^s_{\min}(\boldsymbol{y})} \|a(\cdot, \boldsymbol{y}) - a_s(\cdot, \boldsymbol{y})\|_{L^\infty(D)} \|f\|_{\mathcal{X}'} \quad \text{for all } \boldsymbol{y} \in U. \tag{5.14}$$

Proof. The existence of a unique solution to (5.11) is an immediate consequence of the Lax–Milgram lemma due to the ellipticity assumption (5.12), while the bound (5.13) follows from

$$\begin{aligned} a_{\min}(\boldsymbol{y}) \|u(\cdot, \boldsymbol{y})\|^2_{\mathcal{X}} \leqslant \int_D a(\boldsymbol{x}, \boldsymbol{y}) \nabla u(\boldsymbol{x}, \boldsymbol{y}) \cdot \nabla u(\boldsymbol{x}, \boldsymbol{y}) \,\mathrm{d}\boldsymbol{x} &= \langle f, u(\cdot, \boldsymbol{y}) \rangle_{\mathcal{X}',\mathcal{X}} \\ &\leqslant \|f\|_{\mathcal{X}'} \|u(\cdot, \boldsymbol{y})\|_{\mathcal{X}} \end{aligned}$$

for all $\boldsymbol{y} \in U$.

To prove (5.14), we let $\boldsymbol{y} \in U$ and begin by observing that

$$\int_D a(\boldsymbol{x},\boldsymbol{y})\nabla u(\boldsymbol{x},\boldsymbol{y}) \cdot \nabla v(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = \int_D f(\boldsymbol{x})v(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x},$$
$$\int_D a_s(\boldsymbol{x},\boldsymbol{y})\nabla u_s(\boldsymbol{x},\boldsymbol{y}) \cdot \nabla v(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = \int_D f(\boldsymbol{x})v(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x},$$

for all $v \in \mathcal{X}$. Taking the difference of these two equations, we are left with

$$\begin{aligned}
0 &= \int_D (a(\boldsymbol{x},\boldsymbol{y})\nabla u(\boldsymbol{x},\boldsymbol{y}) - a_s(\boldsymbol{x},\boldsymbol{y})\nabla u_s(\boldsymbol{x},\boldsymbol{y})) \cdot \nabla v(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \\
&= \int_D (a(\boldsymbol{x},\boldsymbol{y}) - a_s(\boldsymbol{x},\boldsymbol{y}))\nabla u(\boldsymbol{x},\boldsymbol{y}) \cdot \nabla v(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \\
&\quad + \int_D a_s(\boldsymbol{x},\boldsymbol{y})\nabla (u(\boldsymbol{x},\boldsymbol{y}) - u_s(\boldsymbol{x},\boldsymbol{y})) \cdot \nabla v(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}.
\end{aligned}$$

Rearranging this equation and setting $v = u(\cdot,\boldsymbol{y}) - u_s(\cdot,\boldsymbol{y}) \in \mathcal{X}$ yields

$$\begin{aligned}
&a^s_{\min}(\boldsymbol{y})\|u(\cdot,\boldsymbol{y}) - u_s(\cdot,\boldsymbol{y})\|^2_{\mathcal{X}} \\
&\leqslant \int_D a_s(\boldsymbol{x},\boldsymbol{y})|\nabla (u(\boldsymbol{x},\boldsymbol{y}) - u_s(\boldsymbol{x},\boldsymbol{y}))|^2 \,\mathrm{d}\boldsymbol{x} \\
&= -\int_D (a(\boldsymbol{x},\boldsymbol{y}) - a_s(\boldsymbol{x},\boldsymbol{y}))\nabla u(\boldsymbol{x},\boldsymbol{y}) \cdot \nabla (u(\boldsymbol{x},\boldsymbol{y}) - u_s(\boldsymbol{x},\boldsymbol{y})) \,\mathrm{d}\boldsymbol{x} \\
&\leqslant \|a(\cdot,\boldsymbol{y}) - a_s(\cdot,\boldsymbol{y})\|_{L^\infty(D)}\|u(\cdot,\boldsymbol{y})\|_{\mathcal{X}}\|u(\cdot,\boldsymbol{y}) - u_s(\cdot,\boldsymbol{y})\|_{\mathcal{X}} \\
&\leqslant \frac{\|a(\cdot,\boldsymbol{y}) - a_s(\cdot,\boldsymbol{y})\|_{L^\infty(D)}}{a_{\min}(\boldsymbol{y})}\|u(\cdot,\boldsymbol{y}) - u_s(\cdot,\boldsymbol{y})\|_{\mathcal{X}}\|f\|_{\mathcal{X}'},
\end{aligned}$$

where we used the a priori bound (5.13) established above. The claim directly follows. □

Studies in uncertainty quantification for PDEs typically consider one of the following two models for the input random field.

(M1) The diffusion coefficient is parameterized by

$$a(\boldsymbol{x},\boldsymbol{y}) = a_0(\boldsymbol{x}) \exp\Bigg(\sum_{j=1}^{\infty} y_j \psi_j(\boldsymbol{x})\Bigg), \quad y_j \in \mathbb{R},$$

for $a_0 \in L^\infty_+(D)$, $\psi_j \in L^\infty(D)$ for all $j \geqslant 1$ with $(\|\psi_j\|_{L^\infty}) \in \ell^p(\mathbb{N})$ for some $p \in (0,1)$, and $U = \mathbb{R}^{\mathbb{N}}$, such that

$$0 < a_{\min}(\boldsymbol{y}) \leqslant a(\boldsymbol{x},\boldsymbol{y}) \leqslant a_{\max}(\boldsymbol{y}) < \infty \quad \text{for all } \boldsymbol{x} \in D \text{ and } \boldsymbol{y} \in U,$$

where $a_{\min}(\boldsymbol{y}) = \min_{x \in \overline{D}} a(\boldsymbol{x},\boldsymbol{y})$ and $a_{\max}(\boldsymbol{y}) = \max_{x \in \overline{D}} a(\boldsymbol{x},\boldsymbol{y})$.

(M2) The diffusion coefficient is parameterized by

$$a(\boldsymbol{x},\boldsymbol{y}) = a_0(\boldsymbol{x}) + \sum_{j=1}^{\infty} y_j \psi_j(\boldsymbol{x}), \quad y_j \in [-1,1],$$

for $a_0 \in L^\infty(D)$, $\psi_j \in L^\infty(D)$ for all $j \geqslant 1$ with $(\|\psi_j\|_{L^\infty}) \in \ell^p(\mathbb{N})$ for some $p \in (0,1)$, and $U = [-1,1]^{\mathbb{N}}$, such that

$$0 < a_{\min} \leqslant a(\boldsymbol{x},\boldsymbol{y}) \leqslant a_{\max} < \infty \quad \text{for all } \boldsymbol{x} \in D \text{ and } \boldsymbol{y} \in U,$$

for some constants $a_{\max} \geqslant a_{\min} > 0$ independent of $\boldsymbol{x} \in D$ and $\boldsymbol{y} \in U$.

Let $\boldsymbol{b} := (b_j)_{j\geqslant 1}$ with $b_j := \|\psi_j\|_{L^\infty}$. In addition, we assume that $b_1 \geqslant b_2 \geqslant \cdots$.
Recall that $u_s(\cdot, \boldsymbol{y}) = u(\cdot, (\boldsymbol{y}_{\leqslant s}, \boldsymbol{0}))$ for $\boldsymbol{y} \in U$. In the context of high-dimensional numerical integration, it is germane in the setting (M1) to quantify the dimension truncation error

$$\left\| \int_U (u(\cdot, \boldsymbol{y}) - u_s(\cdot, \boldsymbol{y}))\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \right\|_{\mathcal{X}}, \quad \text{with } U = \mathbb{R}^{\mathbb{N}},$$

and in the setting (M2), the dimension truncation error

$$\left\| \int_U (u(\cdot, \boldsymbol{y}) - u_s(\cdot, \boldsymbol{y}))\, \boldsymbol{\gamma}(\mathrm{d}\boldsymbol{y}) \right\|_{\mathcal{X}}, \quad \text{with } U = [-1, 1]^{\mathbb{N}}.$$

Lognormal model and its generalizations The model (M1) is the lognormal model (cf., e.g., [63, 70, 69, 72, 84, 119, 148]) when the uncertain parameter $\boldsymbol{y} \in \mathbb{R}^{\mathbb{N}}$ is endowed with the β-Gaussian probability measure with $\beta = 2$ and $\alpha_j = b_j$ for all $j \geqslant 1$. However, the dimension truncation analysis in Section 5.3 covers the more general setting where we have arbitrary $\boldsymbol{\alpha} \in \ell^1(\mathbb{N})$ and either $\beta \in (1, \infty)$ or $\beta = 1$ with $\alpha_j < 1$ for all $j \geqslant 1$. We remark that the latter case corresponds to random variables distributed according to the Laplace distribution. By (5.14), condition (A1) holds because

$$\|a(\cdot, \boldsymbol{y}) - a_s(\cdot, \boldsymbol{y})\|_{L^\infty(D)} \xrightarrow{s\to\infty} 0 \quad \text{for all } \boldsymbol{y} \in U_{\boldsymbol{b}}.$$

On the other hand, condition (A2) holds due to the well-known parametric regularity bound

$$\|\partial^{\boldsymbol{\nu}} u(\cdot, \boldsymbol{y})\|_{\mathcal{X}} \leqslant \frac{\|f\|_{\mathcal{X}'}}{\min_{\boldsymbol{x}\in\overline{D}} a_0(\boldsymbol{x})} \frac{|\boldsymbol{\nu}|!}{(\log 2)^{|\boldsymbol{\nu}|}} \boldsymbol{b}^{\boldsymbol{\nu}} \prod_{j\geqslant 1} \mathrm{e}^{b_j |y_j|} \quad \text{for all } \boldsymbol{y} \in U_{\boldsymbol{b}},\ \boldsymbol{\nu} \in \mathcal{F}.$$

Especially, this corresponds to our setting with the special choice $\alpha_j = b_j$ for all $j \geqslant 1$. By Theorem 5.3.2, we obtain that

$$\left\| \int_{\mathbb{R}^{\mathbb{N}}} (u(\cdot, \boldsymbol{y}) - u_s(\cdot, \boldsymbol{y}))\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \right\|_{\mathcal{X}} \leqslant C s^{-\frac{2}{p}+1},$$

where the implied coefficient is independent of the dimension s.
Let D be a convex and bounded polyhedron and suppose that $\{X_h\}_h$ is a family of conforming finite element subspaces $X_h \subset X$, indexed by the mesh size $h > 0$ (see Section 7.1 for more details on the finite element method). Let $u_h(\cdot, \boldsymbol{y}) \in \mathcal{X}_h$ and $u_{s,h}(\cdot, \boldsymbol{y}) \in \mathcal{X}_h$ denote the finite element discretized solutions corresponding to $u(\cdot, \boldsymbol{y})$ and $u_s(\cdot, \boldsymbol{y})$, respectively. Then it also holds that

$$\left\| \int_{\mathbb{R}^{\mathbb{N}}} (u_h(\cdot, \boldsymbol{y}) - u_{s,h}(\cdot, \boldsymbol{y}))\, \boldsymbol{\mu}_\beta(\mathrm{d}\boldsymbol{y}) \right\|_{\mathcal{X}} \leqslant C s^{-\frac{2}{p}+1},$$

where the implied coefficient is again independent of the dimension s.

Uniform and affine model The model (M2) is known as the uniform and affine model (cf., e.g., [31, 40, 39, 54, 65, 116, 117, 146]) when the uncertain parameter $\boldsymbol{y} \in [-1, 1]^{\mathbb{N}}$ is endowed with the uniform probability measure. By (5.14), condition (A1') holds since

$$\|a(\cdot, \boldsymbol{y}) - a_s(\cdot, \boldsymbol{y})\|_{L^\infty(D)} \xrightarrow{s\to\infty} 0 \quad \text{for all } \boldsymbol{y} \in [-1, 1]^{\mathbb{N}}.$$

Moreover, condition (A2') holds due to the well-known parametric regularity bound

$$\|\partial^{\boldsymbol{\nu}} u(\cdot,\boldsymbol{y})\|_{\mathcal{X}} \leqslant \frac{\|f\|_{\mathcal{X}'}}{a_{\min}^{|\boldsymbol{\nu}|+1}} |\boldsymbol{\nu}|! \boldsymbol{b}^{\boldsymbol{\nu}} \quad \text{for all } \boldsymbol{y} \in [-1,1]^{\mathbb{N}},\ \boldsymbol{\nu} \in \mathcal{F}. \tag{5.15}$$

It follows from Theorem 5.3.3 that

$$\left\| \int_{[-1,1]^{\mathbb{N}}} (u(\cdot,\boldsymbol{y}) - u_s(\cdot,\boldsymbol{y}))\, \boldsymbol{\gamma}(\mathrm{d}\boldsymbol{y}) \right\|_{\mathcal{X}} \leqslant C s^{-\frac{2}{p}+1},$$

where the implied coefficient is independent of the dimension s. The same result holds if u and u_s are replaced by finite element solutions belonging to a conforming finite element subspace of $\mathcal{X}$ (see Section 7.1 for more details on the finite element method).
Finally, we present an example illustrating how our results can be applied to nonlinear quantities of interest of the PDE response.

Example 5.4.2. *Consider the uniform and affine model (M2) with $U = \left[-\frac{1}{2},\frac{1}{2}\right]^{\mathbb{N}}$ and suppose that we are interested in analyzing*

$$\left| \int_{\left[-\frac{1}{2},\frac{1}{2}\right]^{\mathbb{N}}} (G(u(\cdot,\boldsymbol{y})) - G(u_s(\cdot,\boldsymbol{y})))\, \mathrm{d}\boldsymbol{y} \right|, \tag{5.16}$$

where u and u_s denote the parametric PDE solution and its dimension truncation in $X = H_0^1(D)$, respectively. Suppose that the quantity of interest is the nonlinear functional

$$G(v) := \exp\left(\int_D v(\boldsymbol{x})^2\, \mathrm{d}\boldsymbol{x} \right), \quad v \in \mathcal{X}. \tag{5.17}$$

Let $\boldsymbol{\nu} \in \mathcal{F}$ and $\boldsymbol{y} \in [-\frac{1}{2},\frac{1}{2}]^{\mathbb{N}}$. It follows by an application of the Leibniz product rule and the regularity bound (5.15) *that*

$$\begin{aligned}
\|\partial^{\boldsymbol{\nu}} u(\cdot,\boldsymbol{y})^2\|_{\mathcal{X}} &= \left\| \sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \partial^{\boldsymbol{m}} u(\cdot,\boldsymbol{y}) \cdot \partial^{\boldsymbol{\nu}-\boldsymbol{m}} u(\cdot,\boldsymbol{y}) \right\|_{\mathcal{X}} \\
&\leqslant \frac{\|f\|_{\mathcal{X}'}^2}{a_{\min}^{|\boldsymbol{\nu}|+2}} \boldsymbol{b}^{\boldsymbol{\nu}} \sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} |\boldsymbol{m}|!\, |\boldsymbol{\nu}-\boldsymbol{m}|! \\
&= \frac{\|f\|_{\mathcal{X}'}^2}{a_{\min}^{|\boldsymbol{\nu}|+2}} \boldsymbol{b}^{\boldsymbol{\nu}} \sum_{\ell=0}^{|\boldsymbol{\nu}|} \ell!\, (|\boldsymbol{\nu}|-\ell)! \sum_{\substack{\boldsymbol{m} \leqslant \boldsymbol{\nu} \\ |\boldsymbol{m}|=\ell}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \\
&= \frac{\|f\|_{\mathcal{X}'}^2}{a_{\min}^{|\boldsymbol{\nu}|+2}} (|\boldsymbol{\nu}|+1)!\, \boldsymbol{b}^{\boldsymbol{\nu}},
\end{aligned}$$

where we used the generalized Vandermonde identity $\sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu},\, |\boldsymbol{m}|=\ell} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} = \binom{|\boldsymbol{\nu}|}{\ell}$. In complete analogy with the regularity analysis presented in [77, Section 5], which can be found also in Section 4.6, it follows that

$$|\partial^{\boldsymbol{\nu}} G(u(\cdot,\boldsymbol{y}))| \lesssim \left(\frac{\mathrm{e}}{a_{\min}} \right)^{|\boldsymbol{\nu}|} |\boldsymbol{\nu}|!\, \boldsymbol{b}^{\boldsymbol{\nu}},$$

where the implied coefficient only depends on $\|f\|_{\mathcal{X}}$ and $a_{\min}$. Moreover, we have

$$\int_{\left[-\frac{1}{2},\frac{1}{2}\right]^{\mathbb{N}}} (G(u(\cdot,\boldsymbol{y})) - G(u_s(\cdot,\boldsymbol{y})))\, \mathrm{d}\boldsymbol{y} = \int_{[-1,1]^{\mathbb{N}}} (G(u(\cdot,\tfrac{1}{2}\boldsymbol{y})) - G(u_s(\cdot,\tfrac{1}{2}\boldsymbol{y})))\, \boldsymbol{\gamma}(\mathrm{d}\boldsymbol{y})$$

with

$$|\partial^{\boldsymbol{\nu}} G(u(\cdot, \tfrac{1}{2}\boldsymbol{y}))| \lesssim \left(\frac{\mathrm{e}}{2a_{\min}}\right)^{|\boldsymbol{\nu}|} |\boldsymbol{\nu}|!\, \boldsymbol{b}^{\boldsymbol{\nu}}.$$

It follows from Theorem 5.3.3 that the term (5.16) *decays according to* $\mathcal{O}(s^{-\frac{2}{p}+1})$.

Elliptic optimal control problem The problem (4.1) – (4.4) in Section 4.1 is formulated under the assumptions of model (M2). Hence, by (5.14), condition (A1') holds for the state PDE since

$$\|a(\cdot, \boldsymbol{y}) - a_s(\cdot, \boldsymbol{y})\|_{L^\infty(D)} \xrightarrow{s\to\infty} 0 \quad \text{for all } \boldsymbol{y} \in [-\tfrac{1}{2}, \tfrac{1}{2}]^{\mathbb{N}}.$$

Similarly condition (A1') can be verified for the adjoint PDE. Moreover, condition (A2') holds due to the parametric regularity bounds obtained in Lemma 4.6.2 and Lemma 4.6.3 for the state and the adjoint PDE, respectively. Setting $b_j := \|\psi_j\|_{L^\infty}/a_{\min}$ and assuming that $b_1 \geqslant b_2 \geqslant \ldots$, it follows from Theorem 5.3.3 that

$$\left\| \int_{[-\frac{1}{2},\frac{1}{2}]^{\mathbb{N}}} (u(\cdot, \boldsymbol{y}) - u_s(\cdot, \boldsymbol{y}))\, \mathrm{d}\boldsymbol{y} \right\|_V \leqslant C s^{-\frac{2}{p}+1},$$

and for the first term in (4.65) that

$$\left\| \int_{[-\frac{1}{2},\frac{1}{2}]^{\mathbb{N}}} (q(\cdot, \boldsymbol{y}) - q_s(\cdot, \boldsymbol{y}))\, \mathrm{d}\boldsymbol{y} \right\|_V \leqslant C s^{-\frac{2}{p}+1},$$

where the implied coefficient is independent of the dimension s. The same result holds if u and u_s as well as q and q_s are replaced by finite element solutions belonging to a conforming finite element subspace of $\mathcal{X}$ (see Section 7.1 for more details on the finite element method).

Parabolic optimal control problem The optimal control problem (4.38) in Section 4.2 is formulated in the affine and uniform setting. Similar to the elliptic case one can show the continuity of the PDE solution $u^{\boldsymbol{y}}$ with respect to the diffusion coefficient, i.e., verify (A1') using

$$\sup_{t\in I} \|a(\cdot, \boldsymbol{y}) - a_s(\cdot, \boldsymbol{y})\|_{L^\infty(D)} \xrightarrow{s\to\infty} 0 \quad \text{for all } \boldsymbol{y} \in [-\tfrac{1}{2}, \tfrac{1}{2}]^{\mathbb{N}}.$$

By a similar argument (A1') can be verified for the adjoint PDE and by continuity also for S and T, which are defined in Section 4.2.
Moreover, condition (A2') holds due to the parametric regularity bounds obtained in Lemma 4.6.4, Theorem 4.6.6, Theorem 4.6.12 and Theorem 4.6.10 for the state PDE $u^{\boldsymbol{y}}$, the adjoint PDE $q^{\boldsymbol{y}}$, and S, and T, respectively. The following theorem then follows immediately from Theorem 5.3.3.

Theorem 5.4.3. *Let* $f = (z, u_0) \in \mathcal{Y}'$. *For every* $\boldsymbol{y} \in U$, *let* $u^{\boldsymbol{y}} \in \mathcal{X}$ *be the solution of* (4.20) *and* $\Phi^{\boldsymbol{y}}$ *be as in* (4.37), *and then let* $q^{\boldsymbol{y}} \in \mathcal{Y}$ *be the solution of* (4.32) *with* f_{dual} *given by* (4.40). *Suppose the sequence* $\boldsymbol{b} = (b_j)_{j\geqslant 1}$ *defined by* (4.86) *satisfies* $b_1 \geqslant b_2 \geqslant \cdots$. *Then for every* $s \in \mathbb{N}$, *the truncated solutions* $u_s^{\boldsymbol{y}}$, $q_s^{\boldsymbol{y}}$ *and* $\Phi_s^{\boldsymbol{y}}$ *satisfy*

$$\left\| \int_U (u^{\boldsymbol{y}} - u_s^{\boldsymbol{y}})\, \mathrm{d}\boldsymbol{y} \right\|_{\mathcal{X}} \leqslant C\, s^{-\frac{2}{p}+1},$$

$$\left\| \int_U (q^{\boldsymbol{y}} - q_s^{\boldsymbol{y}}) \, \mathrm{d}\boldsymbol{y} \right\|_{\mathcal{Y}} \leqslant C \, s^{-\frac{2}{p}+1},$$

$$\|S - S_s\|_{L^2(V;I)} = \left\| \int_U \left(\exp\left(\theta \, \Phi^{\boldsymbol{y}}\right) q^{\boldsymbol{y}} - \exp\left(\theta \, \Phi_s^{\boldsymbol{y}}\right) q_s^{\boldsymbol{y}} \right) \mathrm{d}\boldsymbol{y} \right\|_{\mathcal{Y}} \leqslant C \, s^{-\frac{2}{p}+1},$$

$$|T - T_s| = \left| \int_U \left(\exp\left(\theta \, \Phi^{\boldsymbol{y}}\right) - \exp\left(\theta \, \Phi_s^{\boldsymbol{y}}\right) \right) \mathrm{d}\boldsymbol{y} \right| \leqslant C \, s^{-\frac{2}{p}+1}.$$

In each case we have a generic constant $C > 0$ independent of s, but depending on z, u_0, $\widehat{u}$ and other constants as appropriate.

In particular, the above theorem provides bounds for the first terms in (4.66), (4.69), and (4.70).

Optimal control problem subject to analytic linear operator equations By continuity we get that $\|A(\boldsymbol{y}_{\leqslant s}) - A(\boldsymbol{y})\|_{\mathcal{L}(X,Y')} \to 0$ as $s \to \infty$. Using a similar strategy as in the Strang lemma (see Lemma 5.4.1) one can verify (A1') for the state. Similarly, (A1') can be verified for the adjoint state and S and T as defined in Section 4.3. Furthermore, (A2') holds due to Corollary 4.6.14, Lemma 4.6.15, Lemma 4.6.16, Theorem 4.6.18, and Theorem 4.6.19. The following theorem then follows immediately from Theorem 5.3.3.

Theorem 5.4.4. *Let $f = \mathcal{B}z \in \mathcal{Y}'$. For every $\boldsymbol{y} \in U$, let $u^{\boldsymbol{y}} \in \mathcal{X}$ be the solution of (3.24) and $\widetilde{\Phi}^{\boldsymbol{y}}$ be as in (4.97), and then let $q^{\boldsymbol{y}} \in \mathcal{Y}$ be the solution of (4.54). Suppose the sequence $\boldsymbol{b} = (b_j)_{j \geqslant 1}$ defined in Corollary 2.3.4 satisfies $b_1 \geqslant b_2 \geqslant \cdots$. Then for every $s \in \mathbb{N}$, the truncated solutions $u_s^{\boldsymbol{y}}$, $q_s^{\boldsymbol{y}}$ and $\widetilde{\Phi}_s^{\boldsymbol{y}}$ satisfy*

$$\left\| \int_U (u^{\boldsymbol{y}} - u_s^{\boldsymbol{y}}) \, \mathrm{d}\boldsymbol{y} \right\|_{\mathcal{X}} \leqslant C \, s^{-\frac{2}{p}+1},$$

$$\left\| \int_U (q^{\boldsymbol{y}} - q_s^{\boldsymbol{y}}) \, \mathrm{d}\boldsymbol{y} \right\|_{\mathcal{Y}} \leqslant C \, s^{-\frac{2}{p}+1},$$

$$\|S - S_s\|_{\mathcal{Y}} = \left\| \int_U \left(\exp\left(\theta \, \Phi^{\boldsymbol{y}}\right) q^{\boldsymbol{y}} - \exp\left(\theta \, \Phi_s^{\boldsymbol{y}}\right) q_s^{\boldsymbol{y}} \right) \mathrm{d}\boldsymbol{y} \right\|_{\mathcal{Y}} \leqslant C \, s^{-\frac{2}{p}+1},$$

$$|T - T_s| = \left| \int_U \left(\exp\left(\theta \, \Phi^{\boldsymbol{y}}\right) - \exp\left(\theta \, \Phi_s^{\boldsymbol{y}}\right) \right) \mathrm{d}\boldsymbol{y} \right| \leqslant C \, s^{-\frac{2}{p}+1}.$$

In each case we have a generic constant $C > 0$ independent of s, but depending on z, $C_{\mathcal{B}}$, $\widehat{u}$ and other constants as appropriate.

In particular, the above theorem provides bounds for the first terms in (4.71), (4.74), and (4.75).

5.5 Numerical experiments

In this section we verify numerically the dimension truncation error rates for in the log-normal setting, for a smooth, nonlinear quantity of interest applied to the PDE solution in the affine and uniform setting, and for the state PDEs and derivatives of the optimal control problems in Section 4.1 and Section 4.2. This section includes computations using the computational cluster Katana supported by Research Technology Services at UNSW Sydney [101].

5.5.1 Lognormal input random field

We consider the PDE problem (5.10) over the spatial domain $D = (0,1)^2$ with the source term $f(\boldsymbol{x}) = x_2$. The PDE (5.10) is discretized spatially using a finite element method with piecewise linear basis functions and mesh size $h = 2^{-5}$ (see Section 7.1 for details on the finite element method). We let $U = \mathbb{R}^{\mathbb{N}}$ and endow the PDE problem (5.10) with the lognormal diffusion coefficient

$$a(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(\sum_{j \geqslant 1} y_j j^{-\vartheta} \sin(j\pi x_1) \sin(j\pi x_2)\right), \quad \boldsymbol{x} \in D,\ \boldsymbol{y} \in \mathbb{R}^{\mathbb{N}},\ \vartheta > 1.$$

To estimate the dimension truncation error, we compute the quantity

$$\begin{aligned} &\left\| \int_{\mathbb{R}^{s'}} (u_{s'}(\cdot, \boldsymbol{y}) - u_s(\cdot, \boldsymbol{y})) \prod_{j=1}^{s} \varphi_2(y_j) \, \mathrm{d}\boldsymbol{y} \right\|_{H_0^1(D)} \\ &= \left\| \int_{(0,1)^{s'}} (u_{s'}(\cdot, \Phi^{-1}(\boldsymbol{t})) - u_s(\cdot, \Phi^{-1}(\boldsymbol{t}))) \, \mathrm{d}\boldsymbol{t} \right\|_{H_0^1(D)}, \end{aligned} \tag{5.18}$$

where $s' \gg s$ and Φ^{-1} is the inverse cumulative distribution function of $\prod_{j=1}^{s'} \varphi_2(y_j)$. The high-dimensional integral appearing in (5.18) was approximated by using a randomly shifted rank-1 lattice rule (see Chapter 6 belows for more details on the quasi-Monte Carlo method) with 2^{20} cubature nodes and a single random shift. The integration lattice was tailored for each value of the decay parameter ϑ by using the QMC4PDE software [114, 113] and the same random shift was used for each ϑ. As the reference, we use the solution corresponding to $s' = 2^{11}$. The numerical results are displayed in Figure 5.1 for dimensions $s \in \{2^k : k \in \{1, \ldots, 9\}\}$ and decay rates $\vartheta \in \{1.5, 2.0, 3.0\}$. The corresponding theoretical convergence rates are -2.0, -3.0, and -5.0, respectively, and they are displayed alongside the numerical results. The observed dimension truncation rates corresponding to $\vartheta \in$

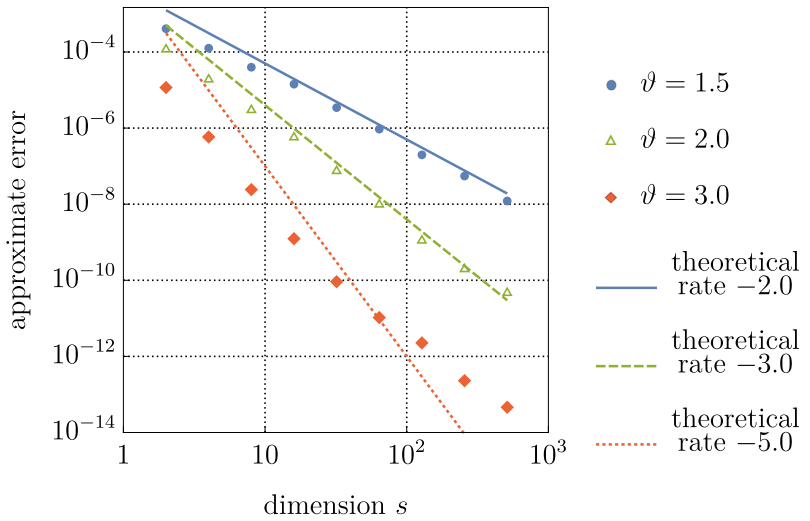


Figure 5.1: The dimension truncation errors corresponding to a lognormally parameterized input random field with decay parameters $\vartheta \in \{1.5, 2.0, 3.0\}$. The expected dimension truncation error rates are -2.0, -3.0, and -5.0, respectively.

$\{1.5, 2.0\}$ start off with slower convergence rates and reach the theoretically predicted convergence rates approximately when $s > 16$. This behavior appears to be the most extreme in the experiment with $\vartheta = 3.0$, where the initial convergence rate for small values of s is slightly slower than the theoretically predicted rate. Moreover, we note that the dimension truncation convergence rate appears to degenerate for large values of s in the case $\vartheta = 3.0$, which may be attributed to cubature error when approximating the high-dimensional integral in (5.18).

5.5.2 Nonlinear quantity of interest

We consider again the PDE problem (5.10) over the spatial domain $D = (0,1)^2$ with the source term $f(\boldsymbol{x}) = x_2$ with the same spatial discretization as in the lognormal case. We revisit Example 5.4.2: let $U = [-\frac{1}{2}, \frac{1}{2}]^{\mathbb{N}}$ and endow the PDE problem (5.10) with the affine random coefficient

$$a(\boldsymbol{x}, \boldsymbol{y}) = \frac{3}{2} + \sum_{j \geqslant 1} y_j j^{-\vartheta} \sin(j\pi x_1) \sin(j\pi x_2), \quad \boldsymbol{x} \in D,\ \boldsymbol{y} \in [-\tfrac{1}{2}, \tfrac{1}{2}]^{\mathbb{N}},\ \vartheta > 1.$$

To estimate the dimension truncation error, we compute

$$\left| \int_{[-\frac{1}{2}, \frac{1}{2}]^{s'}} (G(u_{s'}(\cdot, \boldsymbol{y})) - G(u_s(\cdot, \boldsymbol{y}))) \, \mathrm{d}\boldsymbol{y} \right|,$$

where $s' \gg s$ and G is the nonlinear quantity of interest defined by (5.17). The high-dimensional integrals were again approximated using tailored randomly shifted rank-1 lattice rules with 2^{20} cubature nodes and a single random shift, with the same random shift used for each ϑ. The solution corresponding to $s' = 2^{11}$ was used as the reference. The numerical results are displayed in Figure 5.2 for dimensions $s \in \{2^k : k \in \{1, \dots, 9\}\}$ and decay rates $\vartheta \in \{1.5, 2.0, 3.0\}$ alongside their respective theoretical convergence rates

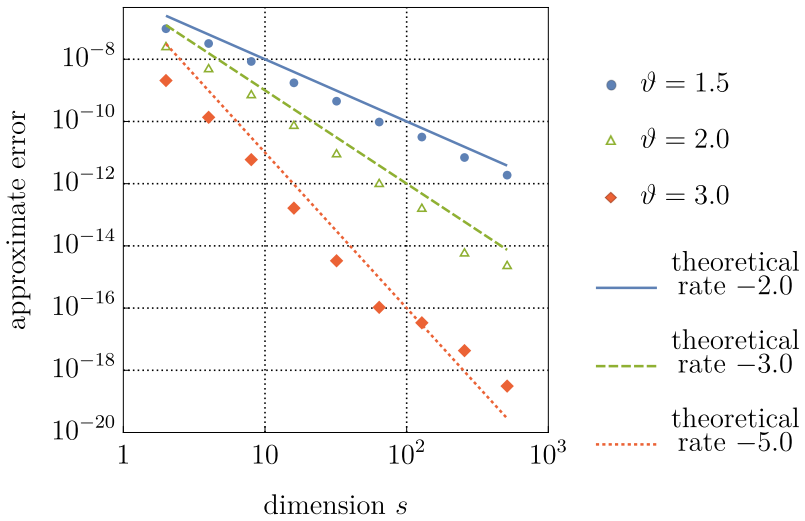


Figure 5.2: The dimension truncation errors corresponding to a nonlinear quantity of interest with decay parameters $\vartheta \in \{1.5, 2.0, 3.0\}$. The expected dimension truncation error rates are -2.0, -3.0, and -5.0, respectively.

-2.0, -3.0, and -5.0. The obtained results agree nicely with the theory, and the numerical results corresponding to $\vartheta = 3.0$ exhibit a saturation effect similar to the one observed in the lognormal setting for larger values of s.

5.5.3 Elliptic optimal control problem

We consider the elliptic state PDE (4.2) – (4.3) and adjoint PDE (4.17) from Section 4.1 in the two-dimensional physical domain $D = (0,1)^2$ equipped with the diffusion coefficient (4.5). We set $a_0(\boldsymbol{x}) \equiv 1$ as the mean field and use the parameterized family of fluctuations (4.62), as we did in the numerical experiment Section 4.4.1.We fix the source term $z(\boldsymbol{x}) = x_2$ and set $\widetilde{u}(\boldsymbol{x}) = x_1^2 - x_2^2$ for $\boldsymbol{x} = (x_1, x_2) \in D$.

The dimension truncation error was estimated by approximating the quantities

$$\left\| \int_U (u(\cdot, \boldsymbol{y}, z) - u_s(\cdot, \boldsymbol{y}, z)) \, \mathrm{d}\boldsymbol{y} \right\|_{L^2(D)} \quad \text{and} \quad \left\| \int_U (q(\cdot, \boldsymbol{y}, z) - q_s(\cdot, \boldsymbol{y}, z)) \, \mathrm{d}\boldsymbol{y} \right\|_{L^2(D)}$$

using a lattice quadrature rule (see Chapter 6) with $n = 2^{15}$ nodes and a single fixed random shift to evaluate the parametric integrals. The coupled PDE system was discretized using the mesh width $h = 2^{-5}$ and, as the reference solutions u and q, we used the FE solutions corresponding to the parameters $s = 2^{11}$ and $h = 2^{-5}$. The obtained results are displayed in Figure 5.5 for the fluctuation operators $(\psi_j)_{j \geqslant 1}$ corresponding to the decay rates $\vartheta \in \{1.5, 2.0\}$ and dimensions $s \in \{2^k : k \in \{1, \ldots, 9\}\}$. The numerical results are accompanied by the corresponding theoretical rates, which are $\mathcal{O}(s^{-2})$ for $\vartheta = 1.5$ and $\mathcal{O}(s^{-3})$ for $\vartheta = 2.0$ according to Theorem 5.3.2.

In all cases, we find that the observed rates tend toward the expected rates as s increases. In particular, by carrying out a least squares fit for the data points corresponding to the values $s \in \{2^5, \ldots, 2^9\}$, the calculated dimension truncation error rate for the state PDE is $\mathcal{O}(s^{-2.00315})$ (corresponding to the decay rate $\vartheta = 1.5$) and $\mathcal{O}(s^{-2.83015})$ (corresponding to the decay rate $\vartheta = 2.0$). For the adjoint PDE, the corresponding rates are $\mathcal{O}(s^{-2.0065})$ and $\mathcal{O}(s^{-2.72987})$, respectively. The discrepancy between the obtained rate and the expected rate in the case of the decay parameter $\vartheta = 2.0$ may be explained by two factors: the lattice quadrature error rate is at best linear, so the quadrature error is likely not completely eliminated with $n = 2^{15}$ lattice quadrature points. Moreover, the rate obtained in Theorem 5.3.3 is sharp only for potentially high values of s. This phenomenon may also be observed in the slight curvature of the data presented in Figure 5.3.

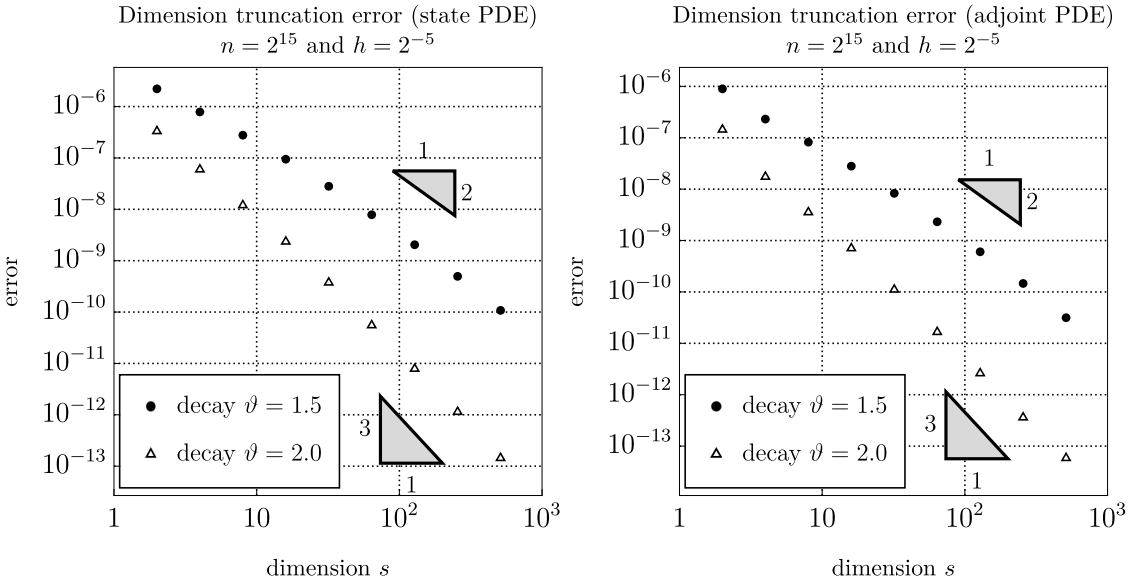


Figure 5.3: The computed dimension truncation errors displayed against the expected rates.

5.5.4 Parabolic optimal control problem

We consider the optimal control problem in Section 4.2, that is we aim to minimize (4.38), i.e., the state PDE $u^{\boldsymbol{y}}$ and adjoint PDE $q^{\boldsymbol{y}}$ are given by (4.20) and (4.43), respectively. We fix the physical domain $D = (0,1)^2$ and the terminal time $T = 1$. The uncertain diffusion coefficient, defined as in (4.21), is independent of t, and parameterized in all experiments with mean field $a_0(\boldsymbol{x}) \equiv 1$ and the fluctuations as in Section 4.4.1.
The initial state, the target state, the objective functional, as well as the FE method are chosen as in Section 4.4.1.
The dimension truncation errors in the parabolic optimal control problem under uncertainty are estimated by approximating the quantities

$$\left\| \int_{U_{s'}} (u^{\boldsymbol{y}}_{s'} - u^{\boldsymbol{y}}_{s}) \, \mathrm{d}\boldsymbol{y} \right\|_{L^2(V;I)} \quad \text{and} \quad \left\| \int_{U_{s'}} (q^{\boldsymbol{y}}_{s'} - q^{\boldsymbol{y}}_{s}) \, \mathrm{d}\boldsymbol{y} \right\|_{L^2(V;I)}$$

as well as

$$\|S_{s'} - S_s\|_{L^2(V;I)} \quad \text{and} \quad |T_{s'} - T_s|$$

for $s' \gg s$, by using a tailored lattice cubature rule generated using the fast CBC algorithm with $n = 2^{15}$ nodes and a single fixed random shift to compute the high-dimensional parametric integrals. The obtained results are displayed in Figures 5.4 and 5.5 for the fluctuations $(\psi_j)_{j \geqslant 1}$ corresponding to decay rates $\vartheta \in \{1.3, 2.6\}$ and dimensions $s \in \{2^k \mid k \in \{1, \dots, 9\}\}$. We use $\theta = 10$ in the computations corresponding to S_s and T_s. As the reference solution, we use the solutions corresponding to dimension $s' = 2048 = 2^{11}$.
The theoretical dimension truncation rate is readily observed in the case $\vartheta = 1.3$. We note in the case $\vartheta = 2.6$ that the dimension truncation convergence rates degenerate for large values of s, which is possibly due to the fact that the QMC cubature with $n = 2^{15}$ nodes has an error around 10^{-8} (see Figure 6.2 in Section 6.3.2). For smaller values of s, the higher order convergence is also apparent in the case $\vartheta = 2.6$.

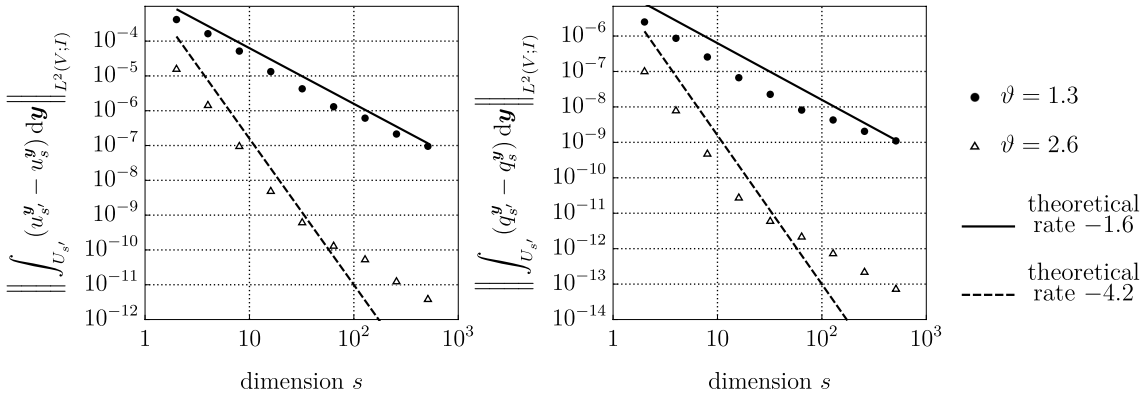


Figure 5.4: The approximate dimension truncation errors corresponding to the state and adjoint PDEs.

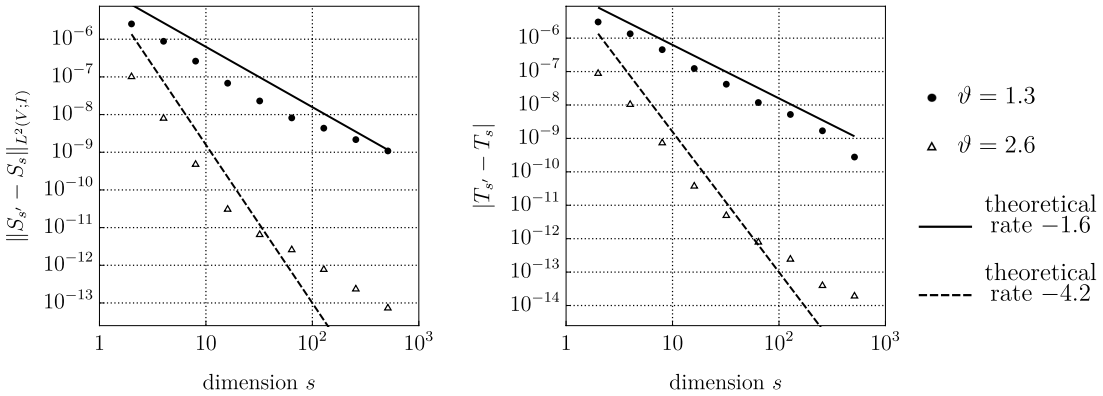


Figure 5.5: The approximate dimension truncation errors corresponding to $\|S_{s'} - S_s\|_{L^2(V;I)}$ and $|T_{s'} - T_s|$.

6 Quasi-Monte Carlo methods

We are interested in computing s-dimensional Bochner integrals of the form

$$I_s(g) := \int_{U_s} g(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y},$$

where $g(\boldsymbol{y})$ is an element of a separable Banach space Z for each $\boldsymbol{y} \in U_s := \left[-\frac{1}{2}, \frac{1}{2}\right]^{\mathbb{N}}$. As our estimator of $I_s(g)$, we use a cubature rule of the form

$$Q_{s,n}(g) := \sum_{i=1}^{n} \alpha_i \, g(\boldsymbol{y}^{(i)}) \,,$$

with weights $\alpha_i \in \mathbb{R}$ and cubature points $\boldsymbol{y}^{(i)} \in U_s$. In particular, we are interested in QMC rules (see, e.g., [38, 113]), which are cubature rules characterized by equal weights $\alpha_i = 1/n$ and carefully chosen (deterministic) points $\boldsymbol{y}^{(i)}$ for $i = 1, \ldots, n$.
We shall see that for sufficiently smooth integrands, randomly shifted rank-1 lattice rules, which are particular QMC rules, lead faster convergence rates compared to Monte Carlo methods. Moreover, under moderate assumptions on the anistropy of the problem with respect to the integration variables, the convergence rate is not dependent on the dimension of the parameter space. In our applications we will hence focus on randomly shifted rank-1 lattice rules.

6.1 Randomly shifted rank-1 lattice rules for real-valued functions

Randomly shifted rank-1 lattice rules are cubature rules over the s-dimensional unit cube $\widetilde{U}_s = [0,1]^s$ with cubature points

$$\widetilde{\boldsymbol{y}}_{\boldsymbol{\Delta}}^{(i)} := \mathrm{frac}\Big(\frac{i\boldsymbol{z}}{n} + \boldsymbol{\Delta}\Big), \quad i = 1, \ldots, n \,,$$

where $\boldsymbol{z} \in \mathbb{N}^s$ is known as the generating vector, $\boldsymbol{\Delta} \in [0,1]^s$ is the random shift and $\mathrm{frac}(\cdot)$ returns the fractional part of each component in the vector. For integration over $U_s = \left[-\frac{1}{2}, \frac{1}{2}\right]^s$, consider the obvious adjustment

$$\boldsymbol{y}_{\boldsymbol{\Delta}}^{(i)} := \mathrm{frac}\Big(\frac{i\boldsymbol{z}}{n} + \boldsymbol{\Delta}\Big) - \Big(\frac{1}{2}, \ldots, \frac{1}{2}\Big), \quad i = 1, \ldots, n \,. \tag{6.1}$$

Integration over different domains and with respect to different measures is possible. For example, by the change of variables $\boldsymbol{\xi} = \Phi^{-1}(\boldsymbol{y})$ for $\boldsymbol{y} \in (0,1)^s$, and the $\Phi(\cdot)$ the element-wise cumulative normal distribution, one can obtain a QMC approximation of the integral

over $\mathbb{R}^s$ with respect to a Gaussian measure, see Section 7.2.2. To keep the analysis in this section simple, we restrict to the case with uniformly distributed parameters y_j over $[-\frac{1}{2}, \frac{1}{2}]$.
In order to get an unbiased estimator, in practice we take the mean over R uniformly drawn random shifts, i.e., we estimate $I_s(g)$ using

$$\overline{Q}_{s,n}(g) := \frac{1}{R}\sum_{r=1}^{R} Q_{s,n}^{(r)}(g), \quad \text{with} \quad Q_{s,n}^{(r)}(g) := \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{y}_{\boldsymbol{\Delta}^{(r)}}^{(i)}). \tag{6.2}$$

In order to quantify the quality of an equal weighted cubature rule, which is determined by its point set, we define the worst-case error of a QMC rule and a point set $P = \{\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(n)}\}$ in a normed space Z by

$$e_{n,s}(P; Z) := \sup_{\|g\|_Z \leqslant 1} |I_s(g) - Q_{s,n}^{(r)}(g)|.$$

For any function $g \in Z$, we have by linearity

$$|I_s(g) - Q_{s,n}^{(r)}(g)| \leqslant e_{n,s}(P; Z)\|g\|_Z \,,$$

and

$$e_{s,n}^{\mathrm{sh}} := \sqrt{\mathbb{E}_{\boldsymbol{\Delta}}|I_s(g) - Q_{s,n}^{(r)}(g)|^2} \leqslant \sqrt{\mathbb{E}_{\boldsymbol{\Delta}}[e_{n,s}^2(P + \boldsymbol{\Delta}; Z)]}\|g\|_Z \,, \tag{6.3}$$

where $P + \boldsymbol{\Delta} = \{\mathrm{frac}(\boldsymbol{y}^{(i)} + \boldsymbol{\Delta}) : i = 1, \dots, n\}$. The shift-averaged worst case error $e_{s,n}^{\mathrm{sh}}$ serves as a quality measure for the QMC rules.
Worst-case errors are in general hard to compute, however for certain function spaces, such as reproducing kernel Hilbert spaces (RKHS) (see [38, Theorem 5.3]), there are explicit formulas for the shift-averaged worst case error.
Let us consider real-valued functions

$$g : \left[-\tfrac{1}{2}, \tfrac{1}{2}\right]^s \to \mathbb{R}$$

that belong to the weighted Sobolev spaces $\mathcal{W}_{s,\boldsymbol{\gamma}}$ with square integrable mixed first derivatives, and which is equipped with norm

$$\|g\|_{\mathcal{W}_{s,\boldsymbol{\gamma}}}^2 = \sum_{\mathfrak{u}\subseteq\{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} \int_{[0,1]^{|\mathfrak{u}|}} \left| \int_{[0,1]^{s-|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|} g}{\partial \boldsymbol{y}_{\mathfrak{u}}}(\boldsymbol{y}_{\mathfrak{u}}; \boldsymbol{y}_{\{1:s\}\setminus\mathfrak{u}}) \, \mathrm{d}\boldsymbol{y}_{\{1:s\}\setminus\mathfrak{u}} \right|^2 \mathrm{d}\boldsymbol{y}_{\mathfrak{u}} \,.$$

From [38, Lemma 5.5] we know that the (squared) shift-averaged worst case error for a rank-1 lattice rule in $\mathcal{W}_{s,\boldsymbol{\gamma}}$ is given by

$$e_{s,n,\boldsymbol{\gamma}}^{\mathrm{sh}}(\boldsymbol{z})^2 = \frac{1}{n}\sum_{k=0}^{n-1} \sum_{\varnothing\neq\mathfrak{u}\subseteq\{1:s\}} \gamma_{\mathfrak{u}} \prod_{j\in\mathfrak{u}} \omega\Big(\mathrm{frac}\Big(\tfrac{kz_j}{n}\Big)\Big), \quad \text{where} \quad \omega(x) = x^2 - x + \frac{1}{6} \,.$$

Considering (6.1), for given s and n, the entire point set is thus determined by the generating vector $\boldsymbol{z}$. Hence, finding a good generating vector is essential to construct good lattice rules. Thus we aim to minimize the the quality measure of the QMC rule, which is the shift-averaged worst case error, with respect to the generating vector.

Note that the components of the generating vector $\boldsymbol{z}$ can be restricted to the set

$$\mathbb{U}_n := \{z \in \mathbb{Z} : 1 \leqslant z \leqslant n-1 \quad \text{and} \quad \gcd(z,n) = 1\},$$

which has cardinality given by the Euler totient function $\phi_{\text{tot}}(n) := |\mathbb{U}_n|$. Here gcd stands for greatest common divisor. When n is prime, then $\phi_{\text{tot}}(n)$ takes its largest value $n-1$, and hence there are up to $(n-1)^s$ possible choices for $\boldsymbol{z}$, which is too many for an exhaustive search for the best generating vector $\boldsymbol{z}$. The component-by-component (CBC) construction is a feasible method to find a good generating vector.

Algorithm 5 CBC construction

Input: n, $s_{\max}$, and weights $\boldsymbol{\gamma}_{\mathfrak{u}}$.

1: Set $z_1 = 1$.
2: **for** For $s = 2, 3, \ldots, s_{\max}$ **do**
3: Choose z_s in $\mathbb{U}_n$ to minimize $e^{\text{sh}}_{s,n,\boldsymbol{\gamma}}(z_1, \ldots, z_s)^2$
4: **end for**

For general weights $\boldsymbol{\gamma}_{\mathfrak{u}}$, the cost of the CBC algorithm is prohibitively expensive, hence we shall be interested in special structures of the weights, such as product weights, or product and order dependent (POD) ones. Since the CBC construction is not the focus of this work, we refer the interested reader for efficient implementations of the CBC construction to [33, 93, 113, 124, 125].
The weights $\boldsymbol{\gamma}_{\mathfrak{u}}$ are said to be of product and order dependent form if they can be written as

$$\boldsymbol{\gamma}_{\mathfrak{u}} = \Gamma_{|\mathfrak{u}|} \prod_{j \in \mathfrak{u}} \gamma_j \,,$$

for two sequences $\gamma_1, \gamma_2, \ldots$ and $\Gamma_1, \Gamma_2, \ldots$ of nonnegative numbers. We have seen in Section 4.6 that POD weights arise naturally in the regularity analysis of PDEs with random coefficients. The computational cost of the fast CBC construction with POD weights is of order $\mathcal{O}(sn \log n + s^2 n)$.

Theorem 6.1.1 ([38, Theorem 5.8])**.** *The generating vector $\boldsymbol{z} \in \mathbb{U}_n^s$ constructed by the CBC algorithm, with the squared shift-averaged worst case error $e^{\text{sh}}_{s,n,\boldsymbol{\gamma}}(\boldsymbol{z})^2$ for the weighted Sobolev space $\mathcal{W}_{s,\boldsymbol{\gamma}}$, satisfies*

$$e^{\text{sh}}_{s,n,\boldsymbol{\gamma}}(\boldsymbol{z})^2 \leqslant \left(\frac{1}{\phi_{\text{tot}}(n)} \sum_{\varnothing \neq \mathfrak{u} \subseteq \{1:s\}} \boldsymbol{\gamma}_{\mathfrak{u}}^{\lambda} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}} \right)^{|\mathfrak{u}|} \right)^{\frac{1}{\lambda}},$$

for all $\lambda \in (\frac{1}{2}, 1]$, where $\zeta(x) := \sum_{h=1}^{\infty} \frac{1}{h^x}$ for $x > 1$ is the Riemann zeta function, and $\phi_{\text{tot}}(n) = |\mathbb{U}_n^s|$ is the Euler totient function.

This directly leads to

Theorem 6.1.2 ([38, Theorem 5.10])**.** *Let $g \in \mathcal{W}_{s,\boldsymbol{\gamma}}$. Then a generating vector $\boldsymbol{z}$ can be constructed using a CBC algorithm such that, for all $\lambda \in (\frac{1}{2}, 1]$,*

$$\sqrt{\mathbb{E}_{\boldsymbol{\Delta}} |I_s(g) - Q^{(r)}_{s,n}(g)|^2} \leqslant \left(\frac{1}{\phi_{\text{tot}}(n)} \sum_{\varnothing \neq \mathfrak{u} \subseteq \{1:s\}} \boldsymbol{\gamma}_{\mathfrak{u}}^{\lambda} \left(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}} \right)^{|\mathfrak{u}|} \right)^{\frac{1}{2\lambda}} \|g\|_{\mathcal{W}_{s,\boldsymbol{\gamma}}} \,,$$

where $\zeta(x) = \sum_{h=1}^{\infty} \frac{1}{h^x}$ *for* $x > 1$ *is the Riemann zeta function, and* $\phi_{\mathrm{tot}}(n) = |\mathbb{U}_n^s|$ *is the Euler totient function.*

In PDE-constrained optimization problems that are subject to uncertainty, the integrals appearing in the objective function and derivatives are typically high-dimensional integrals over Banach space-valued functions. In the next section we thus generalize the well-known results presented in this section to integrals over Banach space-valued functions, i.e., Bochner integrals.

6.2 Randomly shifted rank-1 lattice rules for Bochner integrals

In this section we generalize the results from the previous section to Bochner integrals and apply them to bound the cubature errors in Section 4.5. We first prove a new general result which holds for any cubature rule in a separable Banach space setting.

Theorem 6.2.1. *Let* $U_s = [-\frac{1}{2}, \frac{1}{2}]^s$ *and let* $\mathcal{W}_s$ *be a Banach space of functions* $F : U_s \to \mathbb{R}$*, which is continuously embedded in the space of continuous functions. Consider an* n*-point cubature rule with weights* $\alpha_i \in \mathbb{R}$ *and points* $\boldsymbol{y}^{(i)} \in U_s$*, given by*

$$I_s(F) := \int_{U_s} F(\boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} \approx \sum_{i=1}^{n} \alpha_i \, F(\boldsymbol{y}^{(i)}) =: Q_{s,n}(F),$$

and define the worst case error of $Q_{s,n}$ *in* $\mathcal{W}_s$ *by*

$$e^{\mathrm{wor}}(Q_{s,n}; \mathcal{W}_s) := \sup_{\substack{F \in \mathcal{W}_s, \\ \|F\|_{\mathcal{W}_s} \leqslant 1}} |I_s(F) - Q_{s,n}(F)|.$$

Let Z *be a separable Banach space and let* Z' *denote its dual space. Let* $g : \boldsymbol{y} \mapsto g(\boldsymbol{y})$ *be continuous and* $g(\boldsymbol{y}) \in Z$ *for all* $\boldsymbol{y} \in U_s$*. Then*

$$\Big\| \int_{U_s} g(\boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} - \sum_{i=1}^{n} \alpha_i \, g(\boldsymbol{y}^{(i)}) \Big\|_Z \leqslant e^{\mathrm{wor}}(Q_{s,n}; \mathcal{W}_s) \sup_{\substack{G \in Z', \\ \|G\|_{Z'} \leqslant 1}} \|G(g)\|_{\mathcal{W}_s}. \tag{6.4}$$

Proof. From the separability of Z and the continuity of $g(\boldsymbol{y})$ we get strong measurability of $g(\boldsymbol{y})$. Moreover, from the compactness of U_s and the continuity of $\boldsymbol{y} \mapsto g(\boldsymbol{y})$ we conclude that $\sup_{\boldsymbol{y} \in U_s} \|g(\boldsymbol{y})\|_Z < \infty$ and hence $\int_{U_s} \|g(\boldsymbol{y})\|_Z \,\mathrm{d}\boldsymbol{y} < \infty$, which in turn implies $\| \int_{U_s} g(\boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} \|_Z < \infty$. Thus $g(\boldsymbol{y})$ is Bochner integrable.
Furthermore, for every normed space Z, its dual space Z' is a Banach space equipped with the norm $\|G\|_{Z'} := \sup_{g \in Z, \, \|g\|_Z \leqslant 1} |\langle G, g \rangle_{Z',Z}|$. Then it holds for every $g \in Z$ that $\|g\|_Z = \sup_{G \in Z', \, \|G\|_{Z'} \leqslant 1} |\langle G, g \rangle_{Z',Z}|$. This follows from the Hahn–Banach Theorem, see, e.g., [135, Theorem 4.3].
Thus we have

$$\begin{aligned} \Big\| \int_{U_s} g(\boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} - \sum_{i=1}^{n} \alpha_i \, g(\boldsymbol{y}^{(i)}) \Big\|_Z &= \sup_{\substack{G \in Z', \\ \|G\|_{Z'} \leqslant 1}} \Big| \Big\langle G, \int_{U_s} g(\boldsymbol{y}) \,\mathrm{d}\boldsymbol{y} - \sum_{i=1}^{n} \alpha_i \, g(\boldsymbol{y}^{(i)}) \Big\rangle_{Z',Z} \Big| \\ &= \sup_{\substack{G \in Z', \\ \|G\|_{Z'} \leqslant 1}} \Big| \int_{U_s} \langle G, g(\boldsymbol{y}) \rangle_{Z',Z} \,\mathrm{d}\boldsymbol{y} - \sum_{i=1}^{n} \alpha_i \, \langle G, g(\boldsymbol{y}^{(i)}) \rangle_{Z',Z} \Big|, \end{aligned} \tag{6.5}$$

where we used the linearity of G and the fact that for Bochner integrals we can swap the integral with the linear functional, see (2.16).
From the definition of the worst case error of $Q_{s,n}$ in $\mathcal{W}_s$ it follows that for any $F \in \mathcal{W}_s$ we have

$$|I_s(F) - Q_{s,n}(F)| \leqslant e^{\mathrm{wor}}(Q_{s,n}; \mathcal{W}_s)\, \|F\|_{\mathcal{W}_s}.$$

Applying this to the special case $F(\boldsymbol{y}) = G(g(\boldsymbol{y})) = \langle G, g(\boldsymbol{y})\rangle_{Z',Z}$ in (6.5) yields (6.4). □

Theorem 6.2.2. *Let the assumptions of the preceding Theorem hold. In addition, suppose there exist constants $C_0 > 0$, $r_1 \geqslant 0$, $r_2 > 0$ and a positive sequence $\boldsymbol{\rho} = (\rho_j)_{j\geqslant 1}$ such that for all $\mathfrak{u} \subseteq \{1:s\}$ and for all $\boldsymbol{y} \in U_s$ we have*

$$\Big\| \frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{y}_{\mathfrak{u}}} g(\boldsymbol{y}) \Big\|_Z \leqslant C_0\, (|\mathfrak{u}| + r_1)! \prod_{j\in\mathfrak{u}} (r_2\, \rho_j). \tag{6.6}$$

Then, a randomly shifted rank-1 lattice rule can be constructed using a CBC algorithm such that

$$\mathbb{E}_{\boldsymbol{\Delta}} \Big\| \int_{U_s} g(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y} - \frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{y}^{(i)}) \Big\|_Z^2 \leqslant C_{s,\boldsymbol{\gamma},\lambda}\, [\phi_{\mathrm{tot}}(n)]^{-1/\lambda} \ \text{ for all } \ \lambda \in (\tfrac{1}{2}, 1],$$

where $\phi_{\mathrm{tot}}(n)$ is the Euler totient function, with $1/\phi_{\mathrm{tot}}(n) \leqslant 2/n$ when n is a prime power, and

$$C_{s,\boldsymbol{\gamma},\lambda} := C_0^2 \Bigg(\sum_{\varnothing \neq \mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}}^{\lambda} \Big(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}} \Big)^{|\mathfrak{u}|} \Bigg)^{\frac{1}{\lambda}} \Bigg(\sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{[(|\mathfrak{u}| + r_1)!]^2 \prod_{j\in\mathfrak{u}} (r_2\, \rho_j)^2}{\gamma_{\mathfrak{u}}} \Bigg). \tag{6.7}$$

Proof. We consider randomly shifted rank-1 lattice rules in the unanchored weighted Sobolev space $\mathcal{W}_{s,\boldsymbol{\gamma}}$ with norm

$$\begin{aligned} \|F\|_{\mathcal{W}_{s,\boldsymbol{\gamma}}}^2 &:= \sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} \int_{[-\frac{1}{2},\frac{1}{2}]^{|\mathfrak{u}|}} \Big| \int_{[-\frac{1}{2},\frac{1}{2}]^{s-|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{y}_{\mathfrak{u}}} F(\boldsymbol{y}_{\mathfrak{u}}; \boldsymbol{y}_{\{1:s\}\setminus\mathfrak{u}})\, \mathrm{d}\boldsymbol{y}_{\{1:s\}\setminus\mathfrak{u}} \Big|^2 \mathrm{d}\boldsymbol{y}_{\mathfrak{u}} \\ &\leqslant \sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} \int_{U_s} \Big| \frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{y}_{\mathfrak{u}}} F(\boldsymbol{y}) \Big|^2 \mathrm{d}\boldsymbol{y}. \end{aligned}$$

We have seen in the preceding section that CBC construction yields a lattice generating vector satisfying

$$\mathbb{E}_{\boldsymbol{\Delta}}[e^{\mathrm{wor}}(Q_{s,n}; \mathcal{W})]^2 \leqslant \Big(\frac{1}{\phi_{\mathrm{tot}}(n)} \sum_{\varnothing \neq \mathfrak{u} \subseteq \{1:s\}} \gamma_{\mathfrak{u}}^{\lambda} \Big(\frac{2\zeta(2\lambda)}{(2\pi^2)^{\lambda}} \Big)^{|\mathfrak{u}|} \Big)^{\frac{1}{\lambda}} \text{ for all } \lambda \in (\tfrac{1}{2}, 1].$$

We have from (6.4) that

$$\mathbb{E}_{\boldsymbol{\Delta}} \Big\| \int_{U_s} g(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y} - \frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{y}^{(i)}) \Big\|_Z^2 \leqslant \mathbb{E}_{\boldsymbol{\Delta}}[e^{\mathrm{wor}}(Q_{s,n}; \mathcal{W})]^2 \sup_{\substack{G\in Z' \\ \|G\|_{Z'}\leqslant 1}} \|G(g)\|_{\mathcal{W}_{s,\boldsymbol{\gamma}}}^2.$$

Using the definition of the $\mathcal{W}_{s,\boldsymbol{\gamma}}$-norm, we have

$$\|G(g)\|_{\mathcal{W}_{s,\boldsymbol{\gamma}}}^2 \leqslant \sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} \int_{U_s} \Big| \frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{y}_{\mathfrak{u}}} G(g(\boldsymbol{y})) \Big|^2 \mathrm{d}\boldsymbol{y}$$

$$= \sum_{\mathrm{u}\subseteq\{1:s\}} \frac{1}{\gamma_{\mathrm{u}}} \int_{U_s} \left| G\Big(\frac{\partial^{|\mathrm{u}|}}{\partial \boldsymbol{y}_{\mathrm{u}}} g(\boldsymbol{y})\Big)\right|^2 \mathrm{d}\boldsymbol{y} \leqslant \sum_{\mathrm{u}\subseteq\{1:s\}} \frac{1}{\gamma_{\mathrm{u}}} \int_{U_s} \|G\|_{Z'}^2 \left\| \frac{\partial^{|\mathrm{u}|}}{\partial \boldsymbol{y}_{\mathrm{u}}} g(\boldsymbol{y})\right\|_Z^2 \mathrm{d}\boldsymbol{y}\,.$$

We can now use the assumption (6.6) and combine all of the estimates to arrive at the required bound. □

We now apply this general result to bound the cubature errors in Section 4.5. We start with the elliptic example, i.e., to the third term in (4.65).

Theorem 6.2.3. *Let $z \in L^2(D)$. For every $\boldsymbol{y} \in U$ and $s \in \mathbb{N}$, let $u_s^{\boldsymbol{y}} \in H_0^1(D)$ be the truncated solution of (4.8), and then let $q_s^{\boldsymbol{y}} \in H_0^1(D)$ be the truncated solution of (4.17). Then a randomly shifted rank-1 lattice rule can be constructed using a CBC algorithm such that for all $\lambda \in (\frac{1}{2}, 1]$ we have*

$$\mathbb{E}_{\boldsymbol{\Delta}} \left\| \int_{U_s} u_s^{\boldsymbol{y}}\, \mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^n u_s^{\boldsymbol{y}^{(i)}} \right\|_{L^2(D)}^2 \leqslant C_{s,\boldsymbol{\gamma},\lambda}\, [\phi_{\mathrm{tot}}(n)]^{-1/\lambda}\,,$$
$$\mathbb{E}_{\boldsymbol{\Delta}} \left\| \int_{U_s} q_s^{\boldsymbol{y}}\, \mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^n q_s^{\boldsymbol{y}^{(i)}} \right\|_{L^2(D)}^2 \leqslant C_{s,\boldsymbol{\gamma},\lambda}\, [\phi_{\mathrm{tot}}(n)]^{-1/\lambda}\,,$$

where $\phi_{\mathrm{tot}}(n)$ is the Euler totient function, with $1/\phi_{\mathrm{tot}}(n) \leqslant 2/n$ when n is a prime power. Here $C_{s,\boldsymbol{\gamma},\lambda}$ is given by (6.7), with $r_1 = 1$, $r_2 = 1$, $\rho_j = \|\psi_j\|_{L^\infty(D)}/a_{\min}$, and $C_0 > 0$ is independent of s, n, λ and weights $\boldsymbol{\gamma}$ but depends on z, $\widehat{u}$, $C_{\mathcal{B}}$, $C_{\mathcal{Q}}$, and other constants.

Remark 6.2.4. *For conforming FE methods, i.e., when the FE spaces V_h are subspaces of the solution space V, and the FE error bounds are independent of the parameters $\boldsymbol{y} \in U$, which is the case in the uniform setting considered in this section, then the Banach space QMC error bound directly applies to the FE discretizations. In Theorem 6.2.3 we can replace $u_s^{\boldsymbol{y}}$ and $q_s^{\boldsymbol{y}}$ by their FE approximations $u_{s,h}^{\boldsymbol{y}}$ and $q_{s,h}^{\boldsymbol{y}}$.*

We next apply the general result Theorem 6.2.2 to the second terms in (4.66), (4.69) and (4.70).

Theorem 6.2.5. *Let $f = (z, u_0) \in \mathcal{Y}'$ and $\widehat{u} \in \mathcal{X}$. For every $\boldsymbol{y} \in U$ and $s \in \mathbb{N}$, let $u_s^{\boldsymbol{y}} \in \mathcal{X}$ be the truncated solution of (4.20) and $\Phi_s^{\boldsymbol{y}}$ be as in (4.37), and then let $q_s^{\boldsymbol{y}} \in \mathcal{Y}$ be the truncated solution of (4.32). Then a randomly shifted rank-1 lattice rule can be constructed using a CBC algorithm such that for all $\lambda \in (\frac{1}{2}, 1]$ we have*

$$\mathbb{E}_{\boldsymbol{\Delta}} \left\| \int_{U_s} u_s^{\boldsymbol{y}}\, \mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^n u_s^{\boldsymbol{y}^{(i)}} \right\|_{\mathcal{X}}^2 \leqslant C_{s,\boldsymbol{\gamma},\lambda}\, [\phi_{\mathrm{tot}}(n)]^{-1/\lambda}, \tag{6.8}$$
$$\mathbb{E}_{\boldsymbol{\Delta}} \left\| \int_{U_s} q_s^{\boldsymbol{y}}\, \mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^n q_s^{\boldsymbol{y}^{(i)}} \right\|_{\mathcal{Y}}^2 \leqslant C_{s,\boldsymbol{\gamma},\lambda}\, [\phi_{\mathrm{tot}}(n)]^{-1/\lambda}, \tag{6.9}$$
$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Delta}} \|S_s - S_{s,n}\|_{L^2(V;I)}^2 &\leqslant \mathbb{E}_{\boldsymbol{\Delta}} \left\| \int_{U_s} \exp(\theta\, \Phi_s^{\boldsymbol{y}})\, q_s^{\boldsymbol{y}}\, \mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^n \exp(\theta\, \Phi_s^{\boldsymbol{y}^{(i)}})\, q_s^{\boldsymbol{y}^{(i)}} \right\|_{\mathcal{Y}}^2 \\ &\leqslant C_{s,\boldsymbol{\gamma},\lambda}\, [\phi_{\mathrm{tot}}(n)]^{-1/\lambda}, \end{aligned} \tag{6.10}$$
$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Delta}} |T_s - T_{s,n}|^2 &\leqslant \mathbb{E}_{\boldsymbol{\Delta}} \left| \int_{U_s} \exp(\theta\, \Phi_s^{\boldsymbol{y}})\, \mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^n \exp(\theta\, \Phi_s^{\boldsymbol{y}^{(i)}}) \right|^2 \\ &\leqslant C_{s,\boldsymbol{\gamma},\lambda}\, [\phi_{\mathrm{tot}}(n)]^{-1/\lambda}, \end{aligned} \tag{6.11}$$

where $\phi_{\mathrm{tot}}(n)$ is the Euler totient function, with $1/\phi_{\mathrm{tot}}(n) \leqslant 2/n$ when n is a prime power. Here $C_{s,\boldsymbol{\gamma},\lambda}$ is given by (6.7), with $r_1 = 2$, $r_2 = e$, $\rho_j = b_j$ defined in (4.86), and $C_0 > 0$ is independent of s, n, λ and weights $\boldsymbol{\gamma}$ but depends on u_0, z, $\widehat{u}$, and other constants.

Proof. This follows directly from Theorem 6.2.2 by applying the regularity bounds in Lemma 4.6.4, Theorem 4.6.6, Theorem 4.6.12 and Theorem 4.6.10. For simplicity we set C_0, r_1 and r_2 to be the largest values arising from the four results. □

We get an analogous result for the optimal control problem with parametric linear operator constraints, which can be applied to the second terms in (4.71), (4.74) and (4.75).

Theorem 6.2.6. *Let $\mathcal{B}z \in \mathcal{Y}'$. For every $\boldsymbol{y} \in U$ and $s \in \mathbb{N}$, let $u_s^{\boldsymbol{y}} \in \mathcal{X}$ be the truncated solution of (3.24) and $\widetilde{\Phi}_s^{\boldsymbol{y}}$ be as in (4.97), and then let $q_s^{\boldsymbol{y}} \in \mathcal{Y}$ be the truncated solution of (4.54). Then a randomly shifted rank-1 lattice rule can be constructed using a CBC algorithm such that for all $\lambda \in (\frac{1}{2}, 1]$ we have*

$$\mathbb{E}_{\boldsymbol{\Delta}}\left\| \int_{U_s} u_s^{\boldsymbol{y}} \,\mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^{n} u_s^{\boldsymbol{y}^{(i)}} \right\|_{\mathcal{X}}^2 \leqslant C_{s,\boldsymbol{\gamma},\lambda}\,[\phi_{\mathrm{tot}}(n)]^{-1/\lambda}, \tag{6.12}$$

$$\mathbb{E}_{\boldsymbol{\Delta}}\left\| \int_{U_s} q_s^{\boldsymbol{y}} \,\mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^{n} q_s^{\boldsymbol{y}^{(i)}} \right\|_{\mathcal{Y}}^2 \leqslant C_{s,\boldsymbol{\gamma},\lambda}\,[\phi_{\mathrm{tot}}(n)]^{-1/\lambda}, \tag{6.13}$$

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\Delta}}\|S_s - S_{s,n}\|_{\mathcal{Y}}^2 &\leqslant \mathbb{E}_{\boldsymbol{\Delta}}\left\| \int_{U_s} \exp(\theta\,\Phi_s^{\boldsymbol{y}})\, q_s^{\boldsymbol{y}} \,\mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^{n} \exp(\theta\,\Phi_s^{\boldsymbol{y}^{(i)}})\, q_s^{\boldsymbol{y}^{(i)}} \right\|_{\mathcal{Y}}^2 \\
&\leqslant C_{s,\boldsymbol{\gamma},\lambda}\,[\phi_{\mathrm{tot}}(n)]^{-1/\lambda}, \qquad (6.14)\\
\mathbb{E}_{\boldsymbol{\Delta}}|T_s - T_{s,n}|^2 &\leqslant \mathbb{E}_{\boldsymbol{\Delta}}\left| \int_{U_s} \exp(\theta\,\Phi_s^{\boldsymbol{y}}) \,\mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^{n} \exp(\theta\,\Phi_s^{\boldsymbol{y}^{(i)}}) \right|^2 \\
&\leqslant C_{s,\boldsymbol{\gamma},\lambda}\,[\phi_{\mathrm{tot}}(n)]^{-1/\lambda}, \qquad (6.15)
\end{aligned}$$

where $\phi_{\mathrm{tot}}(n)$ is the Euler totient function, with $1/\phi_{\mathrm{tot}}(n) \leqslant 2/n$ when n is a prime power. Here $C_{s,\boldsymbol{\gamma},\lambda}$ is given by (6.7), with $r_1 = 2$, $r_2 = e$, $\rho_j = b_j$ defined in Corollary 2.3.4, and $C_0 > 0$ is independent of s, n, λ and weights $\boldsymbol{\gamma}$ but depends on z, $\widehat{u}$, $C_{\mathcal{B}}$, $C_{\mathcal{Q}}$, and other constants.

Proof. This follows directly from Theorem 6.2.2 by applying the regularity bounds in Corollary 4.6.14, Lemma 4.6.15, Lemma 4.6.16, Theorem 4.6.18, and Theorem 4.6.19. For simplicity we set C_0, r_1 and r_2 to be the largest values arising from the four results. □

Theorem 6.2.7. *With the choices*

$$\lambda = \begin{cases} \frac{1}{2-2\delta} \text{ for all } \delta \in (0,1) & \text{if } p \in (0, \frac{2}{3}], \\ \frac{p}{2-p} & \text{if } p \in (\frac{2}{3}, 1), \end{cases}$$

$$\gamma_{\mathfrak{u}} = \gamma_{\mathfrak{u}}^* := \left((|\mathfrak{u}| + r_1)! \prod_{j \in \mathfrak{u}} \frac{r_2\,\rho_j}{\sqrt{2\zeta(2\lambda)/(2\pi^2)^{\lambda}}} \right)^{2/(1+\lambda)},$$

we have that $C_{s,\boldsymbol{\gamma}^*,\lambda}$ is bounded independently of s. (However, $C_{s,\boldsymbol{\gamma}^*,\frac{1}{2-2\delta}} \to \infty$ as $\delta \to 0$ and $C_{s,\boldsymbol{\gamma}^*,\frac{p}{2-p}} \to \infty$ as $p \to (2/3)^+$.) Consequently, under the assumption $b_1 \geqslant b_2 \geqslant \ldots$,

the above three mean-square errors in Theorem 6.2.3, Theorem 6.2.5, and Theorem 6.2.6 are of order

$$\begin{cases} [\phi_{\mathrm{tot}}(n)]^{-(2-2\delta)} & \text{for all } \delta \in (0,1) \quad \text{if } p \in (0, \frac{2}{3}]\,, \\ [\phi_{\mathrm{tot}}(n)]^{-(2/p-1)} & \text{if } p \in (\frac{2}{3}, 1)\,. \end{cases} \tag{6.16}$$

Proof. We know from [116, Lemma 6.2] that for any λ, $C_{s,\boldsymbol{\gamma},\lambda}$ defined in (6.7) is minimized by $\gamma_{\mathrm{u}} = \gamma_{\mathrm{u}}^*$. By inserting $\boldsymbol{\gamma}^*$ into $C_{s,\boldsymbol{\gamma},\lambda}$ we can then derive the condition $p < \frac{2\lambda}{1+\lambda} < 1$ for which $C_{s,\boldsymbol{\gamma}^*,\lambda}$ is bounded independently of s. This condition on λ, together with $\lambda \in (\frac{1}{2}, 1]$ and $p \in (0,1)$ yields the result. □

6.3 Numerical experiments

In this section we numerically verify the theoretical QMC error rates for the optimal control problems with the elliptic and the parabolic PDE constraints which we obtained in Theorem 6.2.3 and Theorem 6.2.5.

6.3.1 Elliptic optimal control problem

We assess the rate in Theorem 6.2.3 by using the root-mean-square approximation

$$\sqrt{\mathbb{E}_{\boldsymbol{\Delta}}\left\| \int_{U_s} q_{s,h}(\cdot, \boldsymbol{y}_{\{1:s\}}, z)\, \mathrm{d}\boldsymbol{y}_{\{1:s\}} - \frac{1}{n}\sum_{i=1}^{n} q_{s,h}(\cdot, \{\boldsymbol{t}^{(i)} + \boldsymbol{\Delta}\} - \tfrac{\mathbf{1}}{\mathbf{2}}, z)\right\|_{L^2(D)}^2}$$
$$\approx \sqrt{\frac{1}{R(R-1)}\sum_{r=1}^{R} \left\|\overline{Q}_{s,n} - Q_{s,n}^{(r)}\right\|_{L^2(D)}^2}\,,$$

where $Q_{s,n}^{(r)} := \frac{1}{n}\sum_{i=1}^{n} q_{s,h}(\cdot, \{\boldsymbol{t}^{(i)} + \boldsymbol{\Delta}^{(r)}\} - \frac{\mathbf{1}}{\mathbf{2}}, z)$ and $\overline{Q}_{s,n} = \frac{1}{R}\sum_{r=1}^{R} Q_{s,n}^{(r)}$, for a randomly shifted rank-1 lattice rule with $n = 2^m$, $m \in \{7, \dots, 15\}$, lattice points $(\boldsymbol{t}^{(i)})_{i=1}^{n}$ in $[0,1]^s$ and $R = 16$ random shifts $\boldsymbol{\Delta}^{(r)}$ drawn from the uniform distribution $\mathcal{U}([0,1]^s)$ with $s = 100$. The FE solutions were computed using the mesh width $h = 2^{-6}$, see Section 7.1 for the details. The results are displayed in Figure 6.1. For our experiments we fix the source term $z(\boldsymbol{x}) = x_2$ and otherwise the setup is the same as in Section 4.4.1. In both cases, the theoretical rate is $\mathcal{O}(n^{-1+\delta})$, $\delta > 0$. For the decay rate $\vartheta = 1.5$ (see (4.62)), we observe the rates $\mathcal{O}(n^{-0.984193})$ for the state PDE and $\mathcal{O}(n^{-0.987608})$ for the adjoint PDE. When the decay rate is $\vartheta = 2.0$, we obtain the rates $\mathcal{O}(n^{-1.01080})$ and $\mathcal{O}(n^{-1.012258})$ for the state and adjoint PDE, respectively.

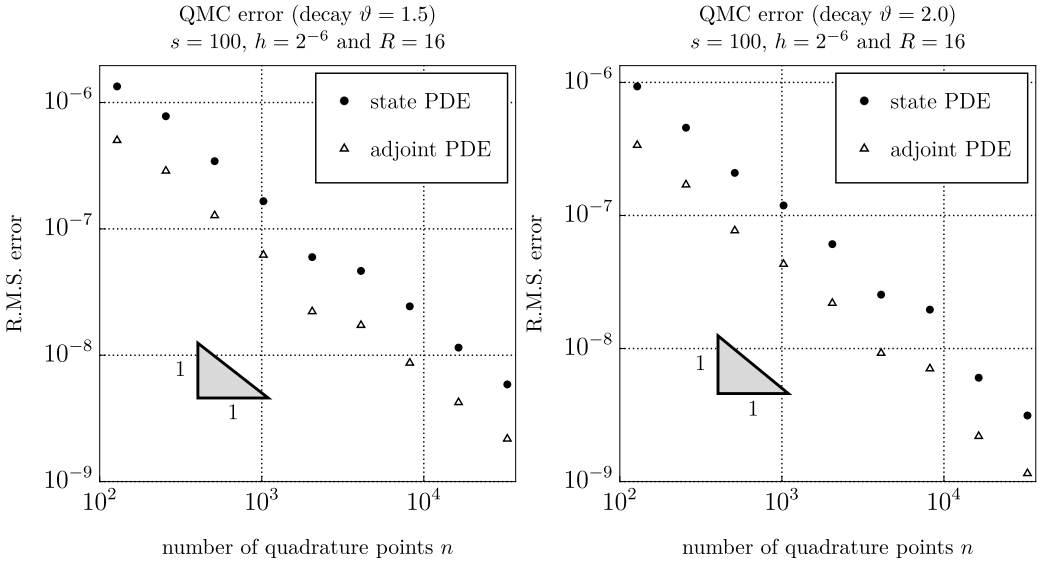


Figure 6.1: The computed root-mean-square errors for the randomly shifted rank-1 lattice rules.

6.3.2 Parabolic optimal control problem

We investigate the QMC error rate obtained in Theorem 6.2.5 by computing the root-mean-square approximations

$$\sqrt{\frac{1}{R(R-1)}\sum_{r=1}^{R}\|(\overline{Q}_{s,n}-Q_{s,n}^{(r)})(u_s^{\cdot})\|_{L^2(V;I)}^2},$$
$$\sqrt{\frac{1}{R(R-1)}\sum_{r=1}^{R}\|(\overline{Q}_{s,n}-Q_{s,n}^{(r)})(q_s^{\cdot})\|_{L^2(V;I)}^2},$$
$$\sqrt{\frac{1}{R(R-1)}\sum_{r=1}^{R}\|(\overline{Q}_{s,n}-Q_{s,n}^{(r)})(\exp(\Phi_s^{\cdot})\,q_s^{\cdot})\|_{L^2(V;I)}^2},$$
$$\sqrt{\frac{1}{R(R-1)}\sum_{r=1}^{R}|(\overline{Q}_{s,n}-Q_{s,n}^{(r)})(\exp(\Phi_s^{\cdot}))|^2},$$

corresponding to (6.8) – (6.11), where $\overline{Q}_{s,n}$ and $Q^{(r)}$ are as in (6.2) for a randomly shifted rank-1 lattice rule with cubature nodes (6.1), where the random shift $\boldsymbol{\Delta}$ is drawn from the uniform distribution $\mathcal{U}([0,1]^s)$. As the generating vector, we use lattice rules constructed using the fast CBC algorithm with $n = 2^m$, $m \in \{4,\dots,15\}$, lattice points and $R = 16$ random shifts, and $s = 100$. We carry out the experiments using two different decay rates $\vartheta \in \{1.3, 2.6\}$ for the input random field, and fix the source term $z(\boldsymbol{x},t) = 10x_1(1-x_1)x_2(1-x_2)$. Otherwise the setup is the same as in Section 4.4.1. The results are displayed in Figure 6.2. The root-mean-square error converges at a linear rate in all experiments, which is consistent with the theory.

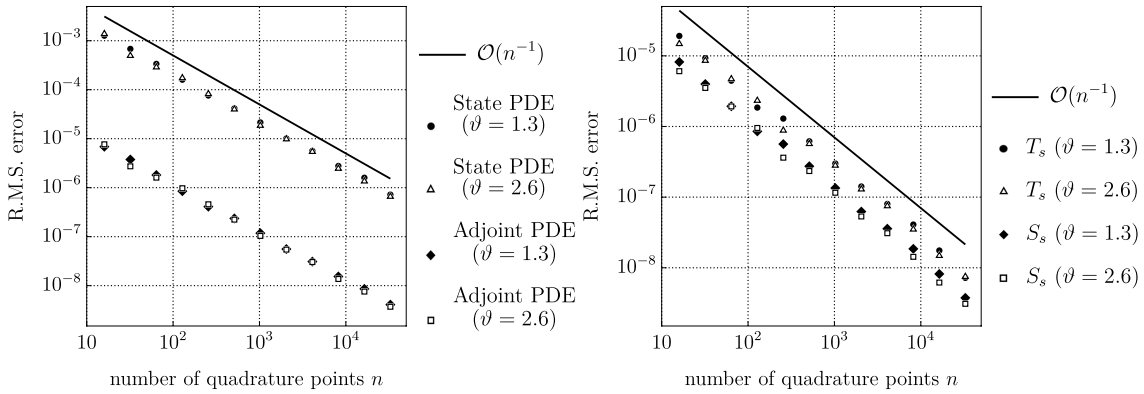


Figure 6.2: Left: The approximate root-mean-square error for QMC approximation of the integrals $\int_{U_s} u_s^{\boldsymbol{y}} \, \mathrm{d}\boldsymbol{y}$ and $\int_{U_s} q_s^{\boldsymbol{y}} \, \mathrm{d}\boldsymbol{y}$. Right: The approximate root-mean-square error for QMC approximation of quantities S_s and T_s. All computations were carried out using $R = 16$ random shifts, $n = 2^m$, $m \in \{4, \dots, 15\}$, and dimension $s = 100$.

7 Discretization and multilevel methods

In this section we briefly present a finite element method. Particular we focus on piecewise linear finite elements and establish error bounds and a convergence rate for the optimal control problem with elliptic PDE constraint.
The second part of this chapter is devoted to a multilevel QMC (MLQMC) estimator of the gradient in the optimal control problem with elliptic PDE constraint. We show that the proposed MLQMC method outperforms a multilevel Monte-Carlo (MLMC) method and the (single level) QMC method.

7.1 Finite element discretization

In this section we apply a basic finite element method to the elliptic example (4.1) – (4.4). In particular, we derive an error bound and a convergence rate for the second term in (4.65). To this end, we briefly recall the basic concept of finite element (FE) methods.
FE methods are numerical schemes for solving PDEs in the spatial variable $\boldsymbol{x} \in D$. To solve a PDE problem, the domain D is divided into subdomains, so-called finite elements. Each subdomain is represented by a set of equations, which are combined to approximate the global solution of the PDE on the domain D.
We present the application to the elliptic PDE problem here for two reasons: we want to complete the error analysis, see (4.65), and we want to introduce the method to better understand how a multilevel method (see Section 7.2 below) uses a discretization scheme to reduce the computational complexity of the problem.
However, since the FE approximation is not the focus of this work and requires mainly well-known results from the literature, we only present the FE error analysis the elliptic PDE example (4.1) – (4.4).
In order to keep the analysis simple and obtain convergence rates of the finite element solutions we make the following additional assumptions (cp. [116, 76])

(AE6) $D \subset \mathbb{R}^d$ is convex bounded polyhedron with plane faces

(AE7) $a_0 \in W^{1,\infty}(D)\,, \quad \sum_{j\geqslant 1} \|\psi_j\|_{W^{1,\infty}(D)} < \infty,$

where $\|v\|_{W^{1,\infty}(D)} := \max\{\|v\|_{L^\infty(D)}\,, \|\nabla v\|_{L^\infty(D)}\}$. The assumption that the geometry of the computational domain D is approximated exactly by the FE mesh simplifies the forthcoming analysis, however, this assumption can substantially be relaxed. For example, standard results on FE analysis as, e.g., in [27] will imply corresponding results for domains D with curved boundaries.
By $\{V_h\}_h$ we denote a family of subspaces $V_h \subset V$ of dimensions $M_h < \infty$, where M_h is of order h^{-d}, with $d \in \{1, 2, 3\}$ denoting the dimension of D. We think of the spaces V_h as spaces spanned by continuous, piecewise linear finite element basis functions on a

sequence of regular, simplicial meshes in D obtained from an initial, regular triangulation of D by recursive, uniform bisection of simplices. Then it is well known (see details, e.g., in [59, 116]) that for functions $v \in V \cap H^2(D)$ there exists a constant $C > 0$, such that as $h \to 0$

$$\inf_{v_h \in V_h} \|v - v_h\|_V \leqslant C\, h\, \|v\|_{V \cap H^2(D)}\,, \tag{7.1}$$

where $\|v\|_{V\cap H^2(D)} := (\|v\|^2_{L^2(D)} + \|\Delta v\|^2_{L^2(D)})^{1/2}$. Note that we need the additional regularity to derive the asymptotic convergence rate as $h \to 0$. For any $\boldsymbol{y} \in U$ and every $z \in L^2(D)$ (and $E_1 z \in V'$, see (4.6)), we define the parametric finite element approximations $u_h(\cdot, \boldsymbol{y}, z) \in V_h$ and $q_h(\cdot, \boldsymbol{y}, z) \in V_h$ by

$$\int_D a(\boldsymbol{x}, \boldsymbol{y}) \nabla u_h(\boldsymbol{x}, \boldsymbol{y}, z) \cdot \nabla v_h(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{x} = \langle E_1 z, v_h \rangle \quad \forall v_h \in V_h\,, \tag{7.2}$$

and then

$$\int_D a(\boldsymbol{x}, \boldsymbol{y}) \nabla q_h(\boldsymbol{x}, \boldsymbol{y}, z) \cdot \nabla w_h(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{x} = \langle u_h(\cdot, \boldsymbol{y}, z) - \widehat{u}, w_h \rangle \quad \forall w_h \in V_h\,, \tag{7.3}$$

We note that the FE approximation (7.2) and (7.3) are defined pointwise with respect to $\boldsymbol{y} \in U$ so that the application of a QMC rule to the FE approximation is well defined. To stress the dependence on s for truncated parameters $\boldsymbol{y} = (y_1, \ldots, y_s, 0, 0, \ldots) \in U$ we write $u_{s,h}$ and $q_{s,h}$ instead of u_h and q_h, respectively.
More precisely, let $(\phi_i)_{i=1}^{M_h}$ be a basis of V_h. Substituting

$$u_h(\boldsymbol{x}, \boldsymbol{y}, z) = \sum_{i=1}^{M_h} u_i(\boldsymbol{y}, z) \phi_i(\boldsymbol{x})$$

into (7.2) with $v_h = \phi_j$ gives

$$\sum_{i=1}^{M_h} u_i \int_D a(\boldsymbol{x}, \boldsymbol{y}) \nabla \phi_i(\boldsymbol{x}) \cdot \nabla \phi_j(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \int_D E_1 z(\boldsymbol{x})\, \phi_j(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}\,,$$

which is equivalent to solving the system of linear equations

$$K\boldsymbol{u} = \boldsymbol{z}\,,$$

where the stiffness matrix K and the load vector $\boldsymbol{z}$ are given as

$$K_{i,j} := \int_D a(\boldsymbol{x}, \boldsymbol{y}) \nabla \phi_i(\boldsymbol{x}) \cdot \nabla \phi_j(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \quad \text{and} \quad \boldsymbol{z} := \int_D E_1 z(\boldsymbol{x}) \cdot \phi_j(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}\,.$$

This system of linear equations can be solved efficiently if the stiffness matrix is sparse. Choosing basis functions with small support $\operatorname{supp}(\phi_j) := \overline{\{x \in D : \phi_j \neq 0\}}$ favors sparsity in the stiffness matrix K. Let $\tau \in T$ be a triangle and let T be a set of disjoint triangles such that $\overline{D} = \bigcup_{\tau \in T} \tau$. Denoting by $\mathrm{vert}_1, \ldots, \mathrm{vert}_{M_h}$ the vertices of the triangles we uniquely define the piecewise linear basis functions as

$$\phi_j(\mathrm{vert}_i) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases}$$

and being linear in each triangle $\tau \in T$. In this case it holds that $K_{i,j} \neq 0$ if and only if vert_i and vert_j are neighbouring vertices. The vertices vert_i and vert_j are neighbouring if there exists a $\tau \in T$ such that $\mathrm{vert}_i, \mathrm{vert}_j \in \tau$.
We obtain the following result for the second term in the error expansion (4.65).

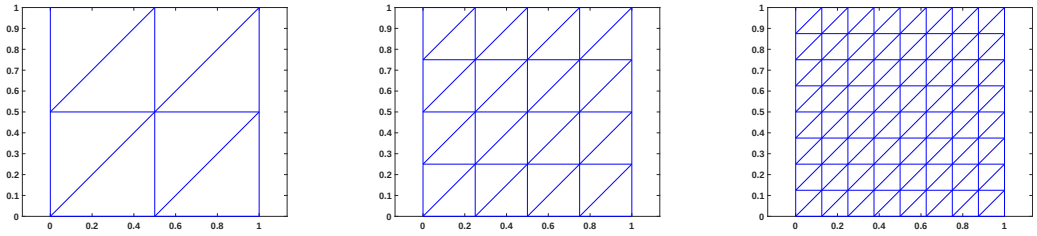


Figure 7.1: From left to right: sequence of regular, triangular meshes in $D = [0,1]^2$ obtained by recursive, uniform bisection of simplices of an initial, regular triangulation of D.

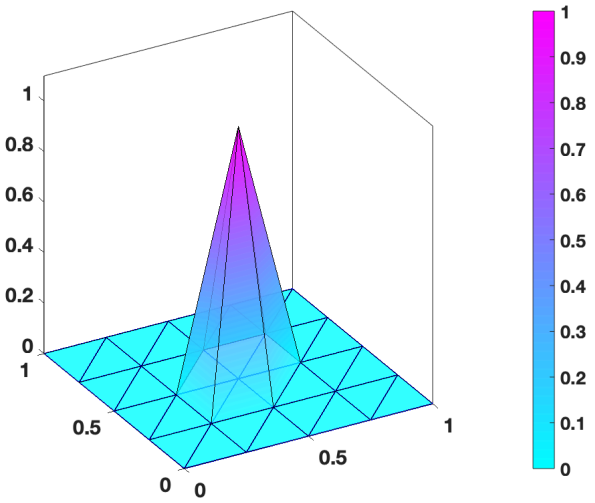


Figure 7.2: Basis function ϕ_j for $\text{vert}_j = (0.5, 0.5)$

Theorem 7.1.1. *Under Assumptions (AE6) and (AE7), for $z \in \mathcal{Z}$, there holds the asymptotic convergence estimate as $h \to 0$*

$$\sup_{\boldsymbol{y}\in U} \|q(\cdot,\boldsymbol{y},z) - q_h(\cdot,\boldsymbol{y},z)\|_{L^2(D)} \leqslant Ch^2 \left(\|z\|_{L^2(D)} + \|\widehat{u}\|_{L^2(D)}\right) ,$$

and

$$\left\| \int_U (q(\cdot,\boldsymbol{y},z) - q_h(\cdot,\boldsymbol{y},z)) \, \mathrm{d}\boldsymbol{y} \right\|_{L^2(D)} \leqslant Ch^2 \left(\|z\|_{L^2(D)} + \|\widehat{u}\|_{L^2(D)}\right) ,$$

where $C > 0$ is independent of h, z and $\widehat{u}$ and $\boldsymbol{y}$.

For truncated $\boldsymbol{y} = (y_1, \ldots, y_s, 0, 0, \ldots) \in U$, the result of Theorem 7.1.1 clearly holds with q and q_h replaced by q_s and $q_{s,h}$, respectively.

Proof. Let $S_{\boldsymbol{y},h}$ be the self-adjoint solution operator defined analogously to (4.12); which for every $\boldsymbol{y} \in U$ assigns to each function $f \in L^2(D)$ the unique solution $g_h(\cdot, \boldsymbol{y}) \in V_h \subset V \subset L^2(D)$. In particular $S_{\boldsymbol{y},h}$ is the solution operator of the problem: find $g_h \in V_h$ such that $b(\boldsymbol{y}; g_h, v_h) = \langle f, v_h \rangle \; \forall v_h \in V_h$. Note that $S_{\boldsymbol{y},h}$ is a bounded and linear operator for given $\boldsymbol{y} \in U$. For every $\boldsymbol{y} \in U$, we can thus estimate

$$\begin{aligned}
\|q(\cdot, \boldsymbol{y}, z) - q_h(\cdot, \boldsymbol{y}, z)\|_{L^2(D)} &= \|S_{\boldsymbol{y}}(u(\cdot, \boldsymbol{y}, z) - \widehat{u}) - S_{\boldsymbol{y},h}(u_h(\cdot, \boldsymbol{y}, z) - \widehat{u})\|_{L^2(D)} \\
&\leqslant \|S_{\boldsymbol{y}}(u(\cdot, \boldsymbol{y}, z) - \widehat{u}) - S_{\boldsymbol{y},h}(u(\cdot, \boldsymbol{y}, z) - \widehat{u})\|_{L^2(D)} \\
&\quad + \|S_{\boldsymbol{y},h} u(\cdot, \boldsymbol{y}, z) - S_{\boldsymbol{y},h} u_h(\cdot, \boldsymbol{y}, z)\|_{L^2(D)} \\
&\leqslant \|(S_{\boldsymbol{y}} - S_{\boldsymbol{y},h})(u(\cdot, \boldsymbol{y}, z) - \widehat{u})\|_{L^2(D)} \\
&\quad + \frac{c_1 c_2}{a_{\min}} \|u(\cdot, \boldsymbol{y}, z) - u_h(\cdot, \boldsymbol{y}, z)\|_{L^2(D)} \,. \qquad (7.4)
\end{aligned}$$

The last step is true because (4.10) holds for all $v \in V$ and therefore it holds in particular for $u_h \in V_h \subset V$. Hence we can bound $\|S_{\boldsymbol{y},h}\|_{\mathcal{L}(L^2(D))} \leqslant \frac{c_1 c_2}{a_{\min}}$. We can now apply the Aubin–Nitsche duality argument (see, e.g., [59]) to bound (7.4): for $w \in L^2(D)$ it holds that

$$\|w\|_{L^2(D)} = \sup_{g \in L^2(D) \setminus \{0\}} \frac{\langle g, w \rangle}{\|g\|_{L^2(D)}} \,. \qquad (7.5)$$

From (4.8) and (7.2) follows the Galerkin orthogonality: $b(\boldsymbol{y}; u(\cdot, \boldsymbol{y}, z) - u_h(\cdot, \boldsymbol{y}, z), v_h) = 0$ for all $v_h \in V_h$, where the parametric bilinear form $b(\boldsymbol{y}; \cdot, \cdot)$ is defined in (4.9). Further, given $g \in L^2(D)$, we define $u_g(\cdot, \boldsymbol{y})$ for every $\boldsymbol{y} \in U$ as the unique solution of the problem: find $u_g(\cdot, \boldsymbol{y}) \in V$ such that

$$b(\boldsymbol{y}; u_g(\cdot, \boldsymbol{y}), w) = \langle g, w \rangle \quad \forall w \in V \,,$$

which leads together with the choice $w := u - u_h$ and the Galerkin orthogonality of the FE discretization to

$$\begin{aligned}
\langle g, u(\cdot, \boldsymbol{y}, z) - u_h(\cdot, \boldsymbol{y}, z) \rangle &= b(\boldsymbol{y}; u_g(\cdot, \boldsymbol{y}), u(\cdot, \boldsymbol{y}, z) - u_h(\cdot, \boldsymbol{y}, z)) \\
&= b(\boldsymbol{y}; u_g(\cdot, \boldsymbol{y}) - v_h, u(\cdot, \boldsymbol{y}, z) - u_h(\cdot, \boldsymbol{y}, z)) \\
&\leqslant a_{\max} \|u_g(\cdot, \boldsymbol{y}) - v_h\|_V \|u(\cdot, \boldsymbol{y}, z) - u_h(\cdot, \boldsymbol{y}, z)\|_V \,.
\end{aligned}$$

With (7.5) we get for every $\boldsymbol{y} \in U$ that

$$\begin{aligned}
\|u(\cdot, \boldsymbol{y}, z) - u_h(\cdot, \boldsymbol{y}, z)\|_{L^2(D)} &= \sup_{g \in L^2(D) \setminus \{0\}} \frac{\langle g, u(\cdot, \boldsymbol{y}, z) - u_h(\cdot, \boldsymbol{y}, z) \rangle}{\|g\|_{L^2(D)}} \\
&\leqslant a_{\max} \|u(\cdot, \boldsymbol{y}, z) - u_h(\cdot, \boldsymbol{y}, z)\|_V \sup_{g \in L^2(D) \setminus \{0\}} \left\{ \inf_{v_h \in V} \frac{\|u_g(\cdot, \boldsymbol{y}) - v_h\|_V}{\|g\|_{L^2(D)}} \right\} .
\end{aligned}$$

Now from (7.1) we infer for every $\boldsymbol{y} \in U$ that

$$\inf_{v_h \in V} \|u_g(\cdot, \boldsymbol{y}) - v_h\|_V \leqslant C_3 \, h \, \|u_g(\cdot, \boldsymbol{y})\|_{V \cap H^2(D)} \leqslant C_4 C_3 \, h \, \|g\|_{L^2(D)} \,,$$

where C_3 is the constant in (7.1). The last step follows from [116, Theorem 4.1] with $t = 1$, and C_4 is the constant in that theorem. For every $\boldsymbol{y} \in U$, we further obtain with Céa's lemma, (7.1) and [116, Theorem 4.1]

$$\|u(\cdot, \boldsymbol{y}, z) - u_h(\cdot, \boldsymbol{y}, z)\|_V \leqslant \frac{a_{\max}}{a_{\min}} \inf_{v_h \in V} \|u(\cdot, \boldsymbol{y}, z) - v_h\|_V$$

$$\leqslant \frac{a_{\max}}{a_{\min}} C_3\, h\, \|u(\cdot,\boldsymbol{y},z)\|_{V\cap H^2(D)} \leqslant \frac{a_{\max}}{a_{\min}} C_4 C_3\, h\, \|z\|_{L^2(D)}\,.$$

Thus for every $\boldsymbol{y}\in U$ it holds that

$$\|u(\cdot,\boldsymbol{y},z)-u_h(\cdot,\boldsymbol{y},z)\|_{L^2(D)} \leqslant \frac{a_{\max}^2}{a_{\min}} C_4^2 C_3^2\, h^2\, \|z\|_{L^2(D)}\,. \tag{7.6}$$

By the same argument we get for every $\boldsymbol{y}\in U$ that

$$\|(S_{\boldsymbol{y}}-S_{\boldsymbol{y},h})(u(\cdot,\boldsymbol{y},z)-\widehat{u})\|_{L^2(D)} \leqslant \frac{a_{\max}^2}{a_{\min}} C_4^2 C_3^2\, h^2 \left(\frac{c_1c_2}{a_{\min}}\|z\|_{L^2(D)}+\|\widehat{u}\|_{L^2(D)}\right). \tag{7.7}$$

Combining (7.6) and (7.7) in (7.4) leads for every $\boldsymbol{y}\in U$ to

$$\|q(\cdot,\boldsymbol{y},z)-q_h(\cdot,\boldsymbol{y},z)\|_{L^2(D)} \leqslant \frac{a_{\max}^2}{a_{\min}} C_4^2 C_3^2\, h^2 \left(\frac{2\,c_1c_2}{a_{\min}}\|z\|_{L^2(D)}+\|\widehat{u}\|_{L^2(D)}\right).$$

The second result easily follows from the first result since

$$\left\| \int_U (q(\cdot,\boldsymbol{y},z)-q_h(\cdot,\boldsymbol{y},z))\,\mathrm{d}\boldsymbol{y}\right\|^2_{L^2(D)} \leqslant \int_U \|q(\cdot,\boldsymbol{y},z)-q_h(\cdot,\boldsymbol{y},z)\|^2_{L^2(D)}\,\mathrm{d}\boldsymbol{y}\,.$$

□

Remark 7.1.2. *In this work the optimal control z^* will always be implicitly discretized in terms of the FE discretization $S_{\boldsymbol{y},h}$ of the solution operator $S_{\boldsymbol{y}}$, see [88].*

Numerical experiments

We validate the FE error bounds given in Theorem 7.1.1 numerically. To this end, we consider the coupled PDE system (7.2) – (7.3) in the two-dimensional physical domain $D=(0,1)^2$ equipped with the diffusion coefficient eq. (4.5), chosen as in Section 4.4.1. We fix the source term $z(\boldsymbol{x})=x_2$ and set $\widehat{u}(\boldsymbol{x})=x_1^2-x_2^2$ for $\boldsymbol{x}=(x_1,x_2)\in D$.
Two numerical experiments were carried out:

(a) The L^2 errors $\|u_s(\cdot,\boldsymbol{y},z)-u_{s,h}(\cdot,\boldsymbol{y},z)\|_{L^2(D)}$ and $\|q_s(\cdot,\boldsymbol{y},z)-q_{s,h}(\cdot,\boldsymbol{y},z)\|_{L^2(D)}$ of the FE solutions to the state and adjoint PDEs, respectively, were computed using the parameters $s=100$ and $h\in\{2^{-k}:k\in\{2,\dots,9\}\}$ for a single realization of the parametric vector $\boldsymbol{y}\in[-1/2,1/2]^{100}$ drawn from $U([-1/2,1/2]^{100})$.

(b) The terms $\left\|\int_{\mathbb{U}_s}(u_s(\cdot,\boldsymbol{y},z)-u_{s,h}(\cdot,\boldsymbol{y},z))\,\mathrm{d}\boldsymbol{y}\right\|_{L^2(D)}$ and $\left\|\int_{\mathbb{U}_s}(q_s(\cdot,\boldsymbol{y},z)-q_{s,h}(\cdot,\boldsymbol{y},z))\,\mathrm{d}\boldsymbol{y}\right\|_{L^2(D)}$ were approUmated by using a lattice rule with a single fixed random shift to evaluate the parametric integrals with dimensionality $s=100$, $n=2^{15}$ nodes and mesh width $h\in\{2^{-k}:k\in\{2,\dots,6\}\}$.

The value $\vartheta=2.0$ was used in both experiments as the rate of decay for the fluctuations (4.62). As the reference solutions u_s and q_s, we used FE solutions computed using the mesh width $h=2^{-10}$ for experiment (a) and $h=2^{-7}$ for experiment (b). The L^2 errors were computed by interpolating the coarser FE solutions onto the grid corresponding to the reference solution. The numerical results are displayed in Figure 7.3. In the case of a single fixed vector $\boldsymbol{y}\in[-1/2,1/2]^{100}$, we obtain the rates $\mathcal{O}(h^{2.01688})$ and $\mathcal{O}(h^{2.00542})$ for the state and adjoint solutions, respectively. The corresponding rates averaged over $n=2^{15}$ lattice quadrature nodes are $\mathcal{O}(h^{2.04011})$ for the state PDE and $\mathcal{O}(h^{2.01617})$ for the adjoint PDE. In both cases, the observed rates adhere nicely with the theoretical rates given in Theorem 7.1.1.

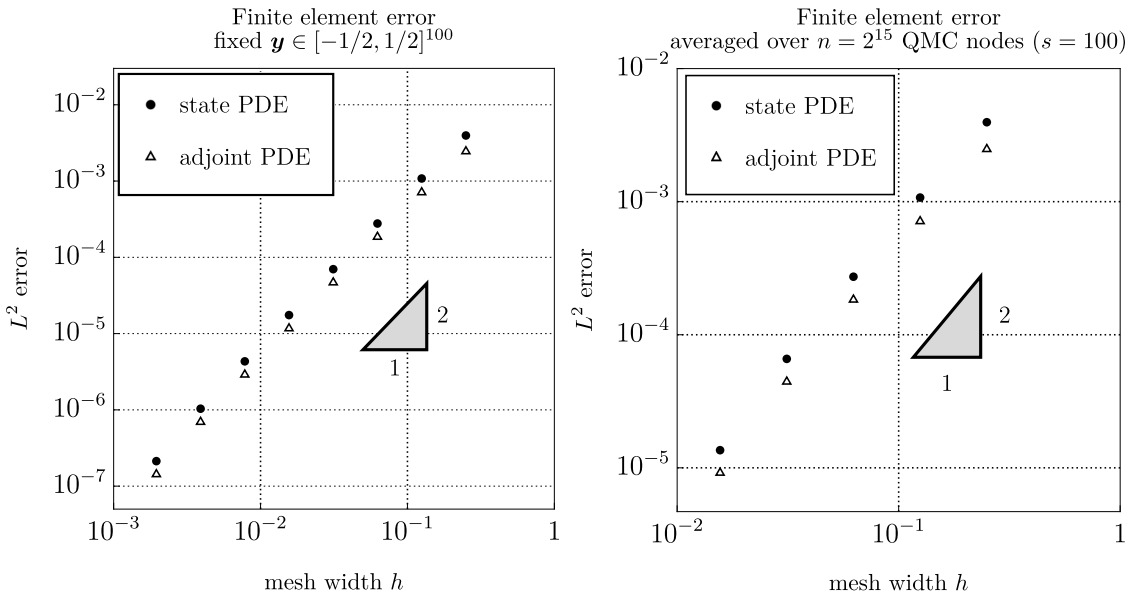


Figure 7.3: The computed finite element errors displayed against the theoretical rates.

7.2 Multilevel quasi-Monte Carlo for optimal control problems

In this chapter, we apply a multilevel method to approximate the gradient of an optimal control problem in order to reduce the computational cost of finding the optimal control using a gradient based method. In particular, we consider the elliptic model problem (see (4.15)) with a lognormally distributed diffusion coefficient and use a multilevel quasi-Monte Carlo (MLQMC) estimator to approximate the expected value appearing in the gradient.
Similar to Section 4.1 we consider the optimal control problem

$$\min_{z \in L^2(D)} J(z), \quad J(z) := \frac{1}{2} \int_\Omega \|u(z) - \widehat{u}\|^2_{L^2(D)} \, \mathrm{d}\mathbb{P} + \frac{\alpha}{2} \|z\|^2_{L^2(D)},$$

where $\alpha > 0$ is a regularization parameter and u as a function of the control z solves the elliptic equation

$$\int_D a(\boldsymbol{x}, \omega) \nabla u(\boldsymbol{x}, \omega) \cdot \nabla v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int_D z(\boldsymbol{x}) v(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \quad \forall v \in H^1_0(D). \tag{7.8}$$

for a.e. $\omega \in \Omega$. The spatial domain $D \subset \mathbb{R}^d$ with $d = 1, 2$ or 3 is a bounded Lipschitz domain and we consider Dirichlet boundary conditions, i.e., $u(\cdot, \omega) \in H^1_0(D) =: V$ has zero trace for a.e. $\omega \in \Omega$, as in Section 4.1.
In this chapter, the random input is assumed to be lognormally distributed as opposed to the uniformly distributed parameters in Section 4.1. We use the notation $a(\boldsymbol{x}, \omega)$ to indicate that the diffusion coefficient is stochastic, i.e., dependends on some random influence $\omega \in \Omega$, where in general ω is an element of the set of events Ω in a suitable probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Provided that the PDE is uniquely solvable, any deterministic control $z \in L^2(D)$ then leads to a solution u that also depends on ω. As we do not assume

additional constraints on the control, the optimality conditions are

$$\int_D a(\boldsymbol{x},\omega)\nabla u(\boldsymbol{x},\omega)\cdot\nabla v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_D z(\boldsymbol{x})v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \quad \forall v\in V, \tag{7.9}$$

$$\int_D a(\boldsymbol{x},\omega)\nabla q(\boldsymbol{x},\omega)\cdot\nabla v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_D (u(\boldsymbol{x},\omega)-\widehat{u}(\boldsymbol{x}))\,v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \quad \forall v\in V, \tag{7.10}$$

$$\nabla J(z) = \mathbb{E}[q] + \alpha z = 0. \tag{7.11}$$

In this section, we address the problem of obtaining an estimate for $\mathbb{E}[q]$ and therefore $\nabla J(z)$ using a multilevel quasi-Monte Carlo method. The resulting gradient could be used in a gradient based optimization problem to find a solution, see Section 4.4.
This section is based on the joint work with Andreas Van Barel [75] in which we draw upon ideas from several previous works. First, the single level quasi-Monte Carlo method was investigated and analyzed for this problem in [76]. Secondly, [116, 119] discusses the application of the MLQMC method to the forward problem (7.8). Both [76] and [116, 119] build on previous papers applying the QMC method to the forward PDE problem; see, e.g., [69, 113]. Next, the application of multilevel Monte Carlo for the optimization problem at hand can be found in [157]. It is itself based on [62, 29] where the MLMC method is applied to the forward problem. In this section we attempt to combine these ideas by employing a MLQMC for the estimation of $\mathbb{E}[q]$ in (7.11). In [76, 119, 157], the uncertain coefficient a is sampled using the Karhunen–Loève (KL) expansion. However, in this manuscript we follow [72], which uses the circulant embedding (CE) method with QMC. Using the CE method, we obtain exact realizations of the random field on a finite set of points and hence there is no truncation error. However, since the FE quadrature points typically do not match the CE grid, we need to interpolate the realizations of the random field. The use of a MLQMC estimator in conjunction with the CE method is new for the optimal control problem as well as for the forward PDE problem.
In this section, we show that the use of QMC points leads to a faster rate of convergence than the ordinary Monte Carlo points. Using the multilevel strategy can further reduce the computational cost. The theoretical convergence rate, as derived in the analysis below, is easily observed in practice. Moreover, the method has little storage costs and is easily parallelizable.

7.2.1 Sampling and discretization

The random field is assumed to be lognormal, i.e., of the form

$$a(\boldsymbol{x},\omega) = \exp(Z(\boldsymbol{x},\omega))\,,$$

where ω is an element of the set of events Ω in the probability space $(\Omega,\mathcal{A},\mathbb{P})$ and $Z(\boldsymbol{x},\omega)$ is a Gaussian random field with prescribed mean $\bar{Z} = \mathbb{E}[Z(\boldsymbol{x},\cdot)]$ and covariance $r_{\mathrm{cov}}(\boldsymbol{x},\boldsymbol{x}') := \mathrm{Cov}\,[Z(\boldsymbol{x},\cdot),Z(\boldsymbol{x}',\cdot)] = \mathbb{E}\big[(Z(\boldsymbol{x},\cdot)-\overline{Z})(Z(\boldsymbol{x}',\cdot)-\overline{Z})\big], \forall \boldsymbol{x},\boldsymbol{x}'\in D$.
One could sample the underlying Gaussian stochastic field using the KL expansion [100, 121] of Z:

$$Z(\boldsymbol{x},\omega) = \mathbb{E}[Z(\boldsymbol{x},\cdot)] + \sum_{n=1}^{\infty}\sqrt{\theta_n}\xi_n(\omega)f_n(\boldsymbol{x}), \quad \boldsymbol{x}\in D,\ \omega\in\Omega. \tag{7.12}$$

The KL expansion is the unique expansion of the above form (with $\|\xi_n\|_{L^2(\Omega)} = \|f_n\|_{L^2(D)} = 1$) that minimizes the total mean square error if the expansion is truncated to a finite

number of terms [58]. This sampling method is widely used, see e.g., [15, 16, 25, 29, 69, 76, 119, 157]. The advantage is that the expansion represents the field Z and therefore $a := \exp(Z)$ at all points in the domain D. In practice, one must however truncate the expansion at some point, introducing a truncation error, see Section 5.3.
Alternatively, one can generate exact realizations of the field in a finite set of m discretization points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ which we collect in the vector

$$\boldsymbol{Z}(\omega) := [Z(\boldsymbol{x}_1, \omega), \ldots, Z(\boldsymbol{x}_m, \omega)]^\top.$$

To that end, consider the resulting covariance matrix $\Sigma = (r_{\text{cov}}(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j=1}^m$ and a factorization of the form $\Sigma = BB^\top$, where $B \in \mathbb{R}^{m\times s}$ with $s \geqslant m$. Defining $\overline{\boldsymbol{Z}} := [\mathbb{E}[Z(\boldsymbol{x}_1, \cdot)], \ldots, \mathbb{E}[Z(\boldsymbol{x}_m, \cdot)]]^\top$,

$$\boldsymbol{Z}(\omega) = B\boldsymbol{Y}(\omega) + \overline{\boldsymbol{Z}},\ \boldsymbol{Y} \sim \mathcal{N}(0, I_{s\times s}) \tag{7.13}$$

then has the desired mean $\mathbb{E}[\boldsymbol{Z}(\omega)] = \overline{\boldsymbol{Z}}$ and covariance

$$\mathbb{E}\big[(\boldsymbol{Z} - \overline{\boldsymbol{Z}})(\boldsymbol{Z} - \overline{\boldsymbol{Z}})^\top\big] = \mathbb{E}\big[B\boldsymbol{Y}\boldsymbol{Y}^\top B^\top\big] = B\mathbb{E}\big[\boldsymbol{Y}\boldsymbol{Y}^\top\big]B^\top = BB^\top = \Sigma.$$

Generating a factorization $\Sigma = BB^\top$ costs in general $\mathcal{O}(m^3)$ operations. However, in what follows we suppose the grids and stochastic fields satisfy the following conditions:

- The set of points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ forms a regular rectangular (also referred to as a uniform rectilinear) grid of points in $\mathbb{R}^d$, with d the dimension.
- The covariance function $r_{\text{cov}}(\boldsymbol{x}, \boldsymbol{x}')$ of the stochastic field is homogeneous, meaning that it is a function of $x - x'$ only. The resulting stochastic field is said to be stationary [1].

In this case, the CE method [21, 43, 69, 161] can be used to very efficiently sample the stochastic field in the given regular rectangular grid of points. In the case $d = 2$, Σ is then block-Toeplitz with Toeplitz blocks and can be embedded in a block-circulant matrix C with circulant blocks, which is the reason for the name of the method. This generalizes to more than two dimensions. The required circulant structure, and the amount of additional padding that may be necessary to ensure positive definiteness determine the size s of $C \in \mathbb{R}^{s\times s}$. Usually, s is of the same order of magnitude as m. A real eigenvalue factorization $C = G\Lambda G^\top$ of this symmetric nested circulant matrix can be obtained using the multidimensional fast Fourier transform, see, e.g., [69]. Since Σ is embedded in a positive definite C, this leads to the desired factorization $\Sigma = BB^\top$ with $B \in \mathbb{R}^{m\times s}$ the first m rows of $G\sqrt{\Lambda}$. For some given realization $\boldsymbol{Y}(\omega)$ of $\boldsymbol{Y}$, a realization

$$\boldsymbol{Z}(\omega) = B\boldsymbol{Y}(\omega) + \overline{\boldsymbol{Z}} \tag{7.14}$$

can then be obtained in $\mathcal{O}(s \log s)$ operations. Some additional details about employing quasi-Monte Carlo values to sample $\boldsymbol{Y}$ follow in Section 7.2.4. The CE method is used in the remainder of this section and allows us to avoid an analysis of the truncation error for the MLQMC estimator. However, the numerical results and the associated analysis of the MLQMC method are not fundamentally dependent on the use of the CE method.
We denote realizations $\boldsymbol{Y}(\omega)$ of the random vector $\boldsymbol{Y}$ by $\boldsymbol{y} = (y_1, \ldots, y_s)$. Since samples of a depend on ω through $\boldsymbol{Y}(\omega)$, we employ the notational convention

$$a(\boldsymbol{x}, \omega) = a_s(\boldsymbol{x}, \boldsymbol{y}) = a_s^{\boldsymbol{y}}(\boldsymbol{x}), \quad \boldsymbol{x} \in \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}.$$

So far, the sample $a_s^{\boldsymbol{y}}(\boldsymbol{x})$ of the lognormal random field is only defined (and exact) at any of the uniform CE grid points $\{\boldsymbol{x}_i\}_{i=1}^m$. For the i-th point $\boldsymbol{x}_i$, this definition is

$$a_s^{\boldsymbol{y}}(\boldsymbol{x}_i) := \exp\big(\sum_{j=1}^{s} B_{i,j} y_j + \overline{Z}_i\big). \tag{7.15}$$

In general these points do not match the quadrature points of the finite element triangulation. Hence an interpolation operator $\mathcal{I}$ is needed. Values of the random field at arbitrary points $\boldsymbol{x} \in D$ are obtained by a multilinear interpolation, i.e., a convex combination of the vertex values $\{\boldsymbol{x}_{k,\boldsymbol{x}}\}_{k=1}^{2^d}$ surrounding $\boldsymbol{x} \in D \subset \mathbb{R}^d$. The resulting approximated sample of $a(\boldsymbol{x}, \omega)$ is denoted by $a_s^{\boldsymbol{y}}(\boldsymbol{x})$ or $a_s(\boldsymbol{x}, \boldsymbol{y})$, and is then defined for all $\boldsymbol{x} \in D$ and $\boldsymbol{y} \in \mathbb{R}^s$ as

$$a_s^{\boldsymbol{y}}(\boldsymbol{x}) := \mathcal{I}(a_s^{\boldsymbol{y}}; \{\boldsymbol{x}_i\}_{i=1}^m)(\boldsymbol{x}) := \sum_{k=1}^{2^d} w_{k,\boldsymbol{x}} a_s^{\boldsymbol{y}}(\boldsymbol{x}_{k,\boldsymbol{x}}), \tag{7.16}$$

with $\sum_{k=1}^{2^d} w_{k,\boldsymbol{x}} = 1$ and $0 \leqslant w_{j,\boldsymbol{x}} \leqslant 1$ for all $k = 1, \ldots, 2^d$. The subscript s indicates the dimension of the random vector $\boldsymbol{y} \in \mathbb{R}^s$ that is used to generate an approximate sample of a. Using this definition the interpolated field matches the exact field at the points $\{\boldsymbol{x}_i\}_{i=1}^m$. Moreover, we observe that the following important properties of the exact sample hold for the interpolated field as well:

- If $a_s^{\boldsymbol{y}}$ is Lipschitz in $\{\boldsymbol{x}_i\}_{i=1}^m$, then $a_s^{\boldsymbol{y}}$ is Lipschitz in all $\boldsymbol{x} \in D$ with the same constant.
- If $a_{\min}(\omega) \leqslant a(\boldsymbol{x}, \omega) \leqslant a_{\max}(\omega)$ holds for all $\boldsymbol{x} \in \{\boldsymbol{x}_i\}_{i=1}^m$, then the same bounds also hold for all $\boldsymbol{x} \in D$.

In Section 7.2.5 (below) the stochastic field properties are discussed in more detail.
Additionally, since we will be employing a multilevel method, it is convenient to be able to generate a sample of a on two different grids starting from a single realization $\boldsymbol{y}$. Consider a first uniform rectilinear grid $\{x_1^1, \ldots, x_{m_1}^1\}$ with m_1 points and a second one $\{x_1^0, \ldots, x_{m_0}^0\}$ consisting of m_0 points. Let us assume the second grid to be coarser, i.e., $m_0 < m_1$. Assume that the CE method requires the vector $\boldsymbol{y}$ to be of dimension s_1 for the fine grid and s_0 for the coarse grid. In the previous paragraph we defined $a_{s_1}^{\boldsymbol{y}}$ for $\boldsymbol{y} \in \mathbb{R}^{s_1}$ and $a_{s_0}^{\boldsymbol{y}}$ for $\boldsymbol{y} \in \mathbb{R}^{s_0}$. We now overload this notation to define $a_{s_0}^{\boldsymbol{y}}$ for $\boldsymbol{y} \in \mathbb{R}^{s_1}$ by

$$a_{s_0}^{\boldsymbol{y}}(\boldsymbol{x}) := \mathcal{I}(a_{s_1}^{\boldsymbol{y}}; \{x_i^0\}_{i=1}^{m_0})(\boldsymbol{x}). \tag{7.17}$$

This means that for a given $\boldsymbol{y} \in \mathbb{R}^{s_1}$, first the stochastic field sample $a_{s_1}^{\boldsymbol{y}}$ is found following (7.16), which is then evaluated in the coarse grid points $\{x_i^0\}_{i=1}^{m_0}$ and used to generate $a_{s_0}^{\boldsymbol{y}}(\boldsymbol{x})$ by linear interpolation between those coarse grid points.
We have two very important properties:

- For a given $\boldsymbol{y} \in \mathbb{R}^{s_1}$, the field samples $a_{s_1}^{\boldsymbol{y}}$ and $a_{s_0}^{\boldsymbol{y}}$ are highly correlated.
- if the coarser grid is nested, i.e., if $\{x_i^0\}_{i=1}^{m_0} \subseteq \{x_i^1\}_{i=1}^{m_1}$, then for either $\boldsymbol{y} \in \mathbb{R}^{s_1}$ or $\boldsymbol{y} \in \mathbb{R}^{s_0}$, a sample $a_{s_0}^{\boldsymbol{y}}$ is exact in the coarse grid points and interpolated in between. This implies that the distribution of $a_{s_0}^{\boldsymbol{Y}}$ with $\boldsymbol{Y} \sim \mathcal{N}(0, I_{s_1 \times s_1})$ is identical to the distribution of $a_{s_0}^{\boldsymbol{Y}}$ with $\boldsymbol{Y} \sim \mathcal{N}(0, I_{s_0 \times s_0})$. If only nested grids are considered, an expression such as $\mathbb{E}\big[a_{s_0}^{\boldsymbol{Y}}\big]$ is then unambiguous, even if the size of $\boldsymbol{Y}$ is not explicitly stated.

Other random variables and random fields in this text depend on ω through their dependence on the stochastic field a. Therefore, we analogously define $u_s(\cdot, \boldsymbol{y}) = u_s^{\boldsymbol{y}}$ and $q_s(\cdot, \boldsymbol{y}) = q_s^{\boldsymbol{y}}$ as realizations of the state u and adjoint q obtained by the interpolated stochastic field $a_s^{\boldsymbol{y}}$, i.e.,

$$\int_D a_s^{\boldsymbol{y}} \nabla u_s^{\boldsymbol{y}} \cdot \nabla v \, \mathrm{d}x = \int_D z \, v \, \mathrm{d}x, \quad \forall v \in V \tag{7.18}$$

$$\int_D a_s^{\boldsymbol{y}} \nabla q_s^{\boldsymbol{y}} \cdot \nabla v \, \mathrm{d}x = \int_D (u_s^{\boldsymbol{y}} - \widehat{u}) \, v \, \mathrm{d}x, \quad \forall v \in V \, . \tag{7.19}$$

Finally the PDEs (7.18) – (7.19) are assumed to be solved using a FE method. Thereby let h be the maximum mesh diameter of the FE grid. The FE solutions of the state and adjoint are denoted as $u_{h,s}^{\boldsymbol{y}}$ and $q_{h,s}^{\boldsymbol{y}}$ respectively and defined as

$$\int_D a_s^{\boldsymbol{y}} \nabla u_{h,s}^{\boldsymbol{y}} \cdot \nabla v_h \, \mathrm{d}x = \int_D z v_h \, \mathrm{d}x, \quad \forall v \in V_h \subset V \tag{7.20}$$

$$\int_D a_s^{\boldsymbol{y}} \nabla q_{h,s}^{\boldsymbol{y}} \cdot \nabla v_h \, \mathrm{d}x = \int_D (u_{h,s}^{\boldsymbol{y}} - \widehat{u}) \, v_h \, \mathrm{d}x, \quad \forall v_h \in V_h \subset V \, , \tag{7.21}$$

where $V_h \subset V$ is the FE space of continuous piecewise linear functions that vanish on the boundary ∂D, see Section 7.1.

7.2.2 Multilevel quasi-Monte Carlo quadrature

As we have seen in the chapter Chapter 6, quasi-Monte methods are equal weight quadrature rules integrating over the s-dimensional unit cube $[0,1]^s$. However, in this section we are interested in finding an approximation for $\mathbb{E}[q_{h,s}(\boldsymbol{x}, \boldsymbol{y})]$ where $\boldsymbol{y}$ follows a normal distribution. Thus it is necessary to perform the change of variables $\boldsymbol{y} = \boldsymbol{\Phi}^{-1}(\boldsymbol{\xi})$, with $\boldsymbol{\Phi}(\cdot)$ the element-wise cumulative normal distribution to obtain

$$\mathbb{E}[q_{h,s}(\boldsymbol{x}, \boldsymbol{y})] = \int_{\mathbb{R}^s} q_{h,s}(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{\Phi}(\boldsymbol{y}) = \int_{[0,1]^s} q_{h,s}(\boldsymbol{x}, \boldsymbol{\Phi}^{-1}(\boldsymbol{\xi})) \, \mathrm{d}\boldsymbol{\xi}. \tag{7.22}$$

To approximate $\mathbb{E}[q_{h,s}(\boldsymbol{x}, \boldsymbol{y})]$, we employ the n-point shifted rank-1 lattice rule $\mathcal{Q}_N$ defined as

$$\mathcal{Q}_N(q_{h,s}(\boldsymbol{x}, \cdot); \boldsymbol{\Delta}) := \frac{1}{n} \sum_{i=1}^{n} q_{h,s} \left(\boldsymbol{x}, \boldsymbol{\Phi}^{-1} \left(\mathrm{frac} \left(\frac{i\boldsymbol{z}}{n} + \boldsymbol{\Delta} \right) \right) \right), \tag{7.23}$$

where $\boldsymbol{z} \in \mathbb{N}^s$ denotes the generating vector and $\boldsymbol{\Delta} \in [0,1]^s$ denotes the shift.
For any a priori choice of the shift $\boldsymbol{\Delta}$, the rule (7.23) is a biased estimator for $\mathbb{E}[q_{h,s}(\boldsymbol{x}, \boldsymbol{y})]$. This bias can be removed by instead considering shifts that are uniformly distributed over $[0,1]^s$. The resulting QMC points $\boldsymbol{\xi}_i = \mathrm{frac}\left(\frac{i\boldsymbol{z}}{n} + \boldsymbol{\Delta}\right)$, $i = 1, \ldots, N$ are then also uniformly distributed over the unit cube. The rule is then an unbiased estimator for $\mathbb{E}[q_{h,s}(\boldsymbol{x}, \boldsymbol{y})]$ since

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\Delta}}[\mathcal{Q}_n(q_{h,s}(\boldsymbol{x}, \cdot); \boldsymbol{\Delta})] &= \int_{[0,1]^s} \frac{1}{n} \sum_{i=1}^{n} q_{h,s} \left(\boldsymbol{x}, \boldsymbol{\Phi}^{-1} \left(\mathrm{frac} \left(\frac{i\boldsymbol{z}}{n} + \boldsymbol{\Delta} \right) \right) \right) \mathrm{d}\boldsymbol{\Delta} \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_{[0,1]^s} q_{h,s} \left(\boldsymbol{x}, \boldsymbol{\Phi}^{-1} \left(\boldsymbol{\xi}_i \right) \right) \mathrm{d}\boldsymbol{\xi}_i = \mathbb{E}[q_{h,s}(\boldsymbol{x}, \boldsymbol{y})].
\end{aligned}$$

The notation $\mathbb{E}_{\boldsymbol{\Delta}}[\cdot]$ emphasizes that the expected value is taken w.r.t. the random shifts. By taking the sample average over R samples of the random shift $\boldsymbol{\Delta}$, and therefore of $\mathcal{Q}_n(q_{h,s}(\boldsymbol{x},\cdot);\boldsymbol{\Delta})$, one obtains the randomly shifted lattice rule

$$\mathcal{Q}_{n,R}(q_{h,s}(\boldsymbol{x},\cdot)) := \frac{1}{R}\sum_{r=1}^{R}\mathcal{Q}_n(q_{h,s}(\boldsymbol{x},\cdot);\boldsymbol{\Delta}_r). \tag{7.24}$$

Another purpose of the random shifts is to facilitate the error estimation. The randomly shifted lattice rule is stochastic, so its root mean square error (RMSE) can be defined as

$$\varepsilon(\mathcal{Q}_{n,R}(q_{h,s})) := \sqrt{\mathbb{E}_{\boldsymbol{\Delta}}[\|\mathcal{Q}_{n,R}(q_{h,s}) - \mathbb{E}[q]\|^2_{L^2(D)}]}. \tag{7.25}$$

Since the means $\mathbb{E}[q_{h,s}]$ and $\mathbb{E}[q]$ are deterministic, it is easily verified that the MSE ε^2 can be expressed as

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Delta}}[\|\mathcal{Q}_{n,R}(q_{h,s}) - \mathbb{E}[q]\|^2_{L^2(D)}] &= \mathbb{E}_{\boldsymbol{\Delta}}[\|\mathcal{Q}_{n,R}(q_{h,s}) - \mathbb{E}[q_{h,s}] + \mathbb{E}[q_{h,s}] - \mathbb{E}[q]\|^2_{L^2(D)}] \\ &= \underbrace{\mathbb{E}_{\boldsymbol{\Delta}}[\|\mathcal{Q}_{n,R}(q_{h,s}) - \mathbb{E}[q_{h,s}]\|^2_{L^2(D)}]}_{\text{QMC quadrature error}} + \underbrace{\|\mathbb{E}[q_{h,s} - q]\|^2_{L^2(D)}}_{\text{Bias}}.\end{aligned} \tag{7.26}$$

The first term is due to the error incurred by the QMC quadrature. It is related to the variance of the randomly shifted lattice rule since

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Delta}}[\|\mathcal{Q}_{n,R}(q_{h,s}) - \mathbb{E}[q_{h,s}]\|^2_{L^2(D)}] &= \int_D \mathbb{E}_{\boldsymbol{\Delta}}[(\mathcal{Q}_{n,R}(q_{h,s}) - \mathbb{E}[\mathcal{Q}_{n,R}(q_{h,s})])^2]\mathrm{d}x \\ &= \int_D \mathbb{V}_{\boldsymbol{\Delta}}[\mathcal{Q}_{n,R}(q_{h,s})]\mathrm{d}x = \int_D \frac{1}{R}\mathbb{V}_{\boldsymbol{\Delta}}[\mathcal{Q}_n(q_{h,s};\boldsymbol{\Delta})]\mathrm{d}x,\end{aligned} \tag{7.27}$$

where we introduced the notation $\mathbb{V}_{\boldsymbol{\Delta}}[\cdot]$ for the variance w.r.t. the random shifts. The R samples of the shift in (7.24) allow the easy estimation

$$\mathbb{V}_{\boldsymbol{\Delta}}[\mathcal{Q}_{n,R}(q_{h,s})] = \frac{1}{R}\mathbb{V}_{\boldsymbol{\Delta}}[\mathcal{Q}_n(q_{h,s};\boldsymbol{\Delta})] \approx \frac{1}{R(R-1)}\sum_{r=1}^{R}(\mathcal{Q}_n(q_{h,s};\boldsymbol{\Delta}_r) - \mathcal{Q}_{n,R}(q_{h,s}))^2. \tag{7.28}$$

This QMC quadrature error depends on the number of QMC points n and the generating vector $\boldsymbol{z}$ in (7.23). The second term in (7.26) is the bias w.r.t. $\mathbb{E}[q]$, due to the discretization error incurred by numerically solving the PDEs. It can be decreased by considering a finer discretization mesh width h.

The multilevel quasi-Monte Carlo (MLQMC) estimator for $\mathbb{E}[q]$ combines estimators of the form (7.24) on a hierarchy of levels $\ell \in \{0, 1, \ldots, L\}$, with level 0 being the coarsest level and L the finest. For each level, we consider a discretization mesh width h_ℓ, with $h_\ell < h_{\ell-1}$, and corresponding spaces $V_{h_0} \subset V_{h_1} \subset \ldots \subset V_{h_L} \subset V = V$ in which approximations u_h for the state and q_h for the adjoint exist.

We define $q_\ell := q_{h_\ell,s_\ell}$, $\ell = 0, \ldots, L$. Using a telescopic sum and the linearity of the expected value operator, we observe that the expected value on the finest discretization level is equal to the expected value on the coarsest level plus a series of corrections, i.e.,

$$\mathbb{E}[q_L] = \mathbb{E}[q_0] + \sum_{\ell=1}^{L}\mathbb{E}[q_\ell - q_{\ell-1}] = \sum_{\ell=0}^{L}\mathbb{E}[q_\ell - q_{\ell-1}], \tag{7.29}$$

where we follow the convention $q_{-1} := 0$. The multilevel quasi-Monte Carlo estimator for $\mathbb{E}[q]$ is obtained by estimating each of the terms in the right-hand side with a randomly shifted lattice rule (7.24), yielding

$$\mathcal{Q}^{\mathrm{ML}}_{\boldsymbol{n},\boldsymbol{R}}(q) := \sum_{\ell=0}^{L} \mathcal{Q}_{n_\ell,R_\ell}(q_\ell - q_{\ell-1}) = \sum_{\ell=0}^{L} \frac{1}{R_\ell} \sum_{r=1}^{R_\ell} \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \left(q_\ell(\cdot, \boldsymbol{y}_\ell^{(i,r)}) - q_{\ell-1}(\cdot, \boldsymbol{y}_\ell^{(i,r)}) \right),$$

where $\boldsymbol{y}_\ell^{(i,r)} := \boldsymbol{\Phi}^{-1}(\mathrm{frac}(i\boldsymbol{z}_\ell n_\ell^{-1} + \boldsymbol{\Delta}_{\ell,r})) \in \mathbb{R}^{s_\ell}$, with $\boldsymbol{z}_\ell \in \mathbb{N}^{s_\ell}$ the generating vector on level ℓ and s_ℓ the stochastic dimension on level ℓ. All random shifts $\boldsymbol{\Delta}_{\ell,r}$ are independent. Both s_ℓ and $\boldsymbol{z}_\ell$ are in general different from level to level.
It is important that both terms $q_\ell(\cdot, \boldsymbol{y}_\ell^{(i,r)})$ and $q_{\ell-1}(\cdot, \boldsymbol{y}_\ell^{(i,r)})$ are evaluated for the same approximate realization $a_{s_\ell}(\cdot, \boldsymbol{y}_\ell^{(i,r)})$ of the stochastic field. Note that if $s_{\ell-1} < s_\ell$, then $q_{\ell-1}(\cdot, \boldsymbol{y}_\ell^{(i,r)}) = q_{h_{\ell-1},s_{\ell-1}}(\cdot, \boldsymbol{y}_\ell^{(i,r)})$ is evaluated as stated by (7.17): first $a_{s_\ell}(\cdot, \boldsymbol{y}_\ell^{(i,r)})$ is evaluated in the CE grid points corresponding to level $\ell-1$ and then $a_{s_{\ell-1}}(\cdot, \boldsymbol{y}_\ell^{(i,r)})$ is formed by linear interpolation between those grid points. The quantity $q_{\ell-1}(\cdot, \boldsymbol{y}_\ell^{(i,r)})$ is the adjoint solution corresponding to that interpolated diffusion coefficient $a_{s_{\ell-1}}(\cdot, \boldsymbol{y}_\ell^{(i,r)})$. Now, in order to ensure $\mathbb{E}[\mathcal{Q}^{\mathrm{ML}}_{\boldsymbol{n},\boldsymbol{R}}(q)] = \mathbb{E}[q_L]$ through the telescopic sum (7.29), the distribution of $q_{\ell-1}(\cdot, \boldsymbol{y}_\ell^{(i,r)})$ must equal the distribution of $q_{\ell-1}(\cdot, \boldsymbol{y}_{\ell-1}^{(i,r)})$, and therefore the distribution of $a_\ell(\cdot, \boldsymbol{y}_{\ell-1}^{(i,r)})$ equals the distribution of $a_{\ell-1}(\cdot, \boldsymbol{y}_{\ell-1}^{(i,r)})$. As discussed in Section 7.2.1, this necessitates that the uniform rectilinear grids involved in the CE sampling of the diffusion coefficient are nested. If we denote the m_ℓ point CE grid at level ℓ by $\{x_i^\ell\}_{i=1}^{m_\ell}$, we therefore must choose grids such that $\{x_i^0\}_{i=1}^{m_0} \subseteq \{x_i^1\}_{i=1}^{m_1} \subseteq \ldots \subseteq \{x_i^L\}_{i=1}^{m_L}$ and therefore we also have $s_0 \leqslant s_1 \leqslant \ldots \leqslant s_L$.

7.2.3 Error and cost

Analogous to (7.25), and due to the independence of the random shifts used for each level, the RMSE of the MLQMC estimator can be shown to equal

$$\varepsilon(\mathcal{Q}^{\mathrm{ML}}_{\boldsymbol{n},\boldsymbol{R}}(q))^2 := \sum_{\ell=0}^{L} \mathcal{V}_\ell + \|\mathbb{E}[q_L - q]\|^2_{L^2(D)}, \tag{7.30}$$

with

$$\mathcal{V}_\ell := \int_D \mathbb{V}_{\boldsymbol{\Delta}}[\mathcal{Q}_{n_\ell,R_\ell}(q_\ell - q_{\ell-1})]\mathrm{d}x. \tag{7.31}$$

As in (7.26), the first term quantifies the quadrature errors of the QMC methods on all levels. They can be estimated using the sample variance of the R_ℓ samples as demonstrated in (7.28). The second term is the bias, which coincides with the single-level bias term in (7.26) for $h = h_L$.
The basic cost and convergence theorems are presented following [119], but applied to our specific case where the circulant embedding method is used as opposed to the KL expansion. To that end, we first formulate a set of general assumptions about the convergence rate of the PDE discretization, the RMSE of the QMC estimator and the computational cost of the sample generation. We introduce the notation $a \lesssim b$ implies that $a \leqslant cb$ with $c > 0$ some constant independent of a and b, and $a \eqsim b$ as $a \lesssim b$ and $b \lesssim a$.
Let $M_\ell := \dim(V_{h_\ell})$ denote the number of degrees of freedom associated with the FE approximation of the PDE at level ℓ. We assume that

Assumption 7.2.1. *$M_\ell \eqsim h_\ell^{-d}$ and $s_\ell \lesssim M_\ell \log M_\ell$.*

The first part of the assumption holds for a variety of mesh families, including locally or anisotropically refined meshes [72]. The second part here details the refinement of the CE grid in relation to the refinement of the FE grid. The assumption allows the CE grid to contain all the quadrature points in the FE triangulation. Even if the FE grid is not a subgrid of the CE grid, it allows the mesh width of the CE grid to be proportional to the FE mesh width, which is a straightforward choice in practice. In those cases, due to the padding requirements in general being dependent on the grid refinement, this leads to a stochastic dimension s_ℓ at level ℓ proportional to $M_\ell \log M_\ell$; see [71] for a detailed analysis. If no padding is required in the CE method, then $s_\ell \simeq M_\ell$. The assumption then trivially also covers the case where the CE grid is refined more slowly than the FE grid. Finally, also covered is the case where the CE grid is refined until some predetermined maximum refinement level $L_{\max}$ is reached. In that case, $s_\ell = s_{L_{\max}}$ for $\ell \geqslant L_{\max}$.
We assume that the hierarchy of discretization levels for the PDE (7.8) has a weak order of convergence ρ, i.e.,

Assumption 7.2.2. *$\|\mathbb{E}[q_\ell - q]\|_{L^2(D)} \lesssim h_\ell^\rho$ for some constant $\rho > 0$.*

This assumption and the next two are stated in terms of h_ℓ. Due to Assumption Assumption 7.2.1, any possible dependence on s_ℓ is incorporated into a dependence on h_ℓ. For elliptic problems such as the Laplace problem described in this section, one expects $\rho = 2$, at least for diffusion coefficients that are smooth enough. However, the simultaneous refining of the random field itself may lead to an order $\rho = 1$.
Next we make an assumption on the variance of the QMC estimator, the justification of which is the subject of the analysis later in this section.

Assumption 7.2.3. *$\mathcal{V}_\ell \lesssim R_\ell^{-1} n_\ell^{-1/\lambda} h_\ell^\varphi$ for some constants $\lambda, \varphi > 0$, with $\mathcal{V}_\ell$ as defined in* (7.31).

Usually one expects $\varphi = 2\rho$. For a standard Monte Carlo method, one would have $\lambda = 1$, i.e., the variance would be inversely proportional to the number of Monte Carlo samples. We will see that the QMC method yields a better rate of convergence. The theoretical results in Section 7.2.5 show that $\lambda \in (1/2, 1]$ can be attained.
Finally, let the cost to compute a sample $q_\ell(\cdot, \boldsymbol{y}_\ell)$ with $\boldsymbol{y}_\ell \in \mathbb{R}^{s_\ell}$ on level ℓ be denoted as $\mathcal{C}_\ell$. We assume

Assumption 7.2.4. *The computational cost for a single sample, denoted $\mathcal{C}_\ell$, satisfies $\mathcal{C}_\ell \lesssim h_\ell^{-\kappa}$ for some constant κ.*

The cost $\mathcal{C}_\ell$ consists of two parts. First, there is the cost $\mathcal{C}_\ell^{\mathrm{FE}}$ of the FE solver. If a multigrid solver is used, this cost is typically at most of the order $\mathcal{O}(M_\ell \log M_\ell)$ and typically of the order $\mathcal{O}(M_\ell)$. Next, there is a cost $\mathcal{C}_\ell^{\mathrm{CE}}$ of $\mathcal{O}(s_\ell \log s_\ell)$ operations for generating the diffusion coefficient sample through the CE method. Due to Assumption 7.2.1, $\mathcal{C}_\ell^{\mathrm{CE}} = \mathcal{O}(M_\ell (\log M_\ell)^2)$. Assumption 7.2.4 then holds with $\kappa = d + \delta$ for an arbitrary small $\delta > 0$.
Supposing that constants $\lambda, \rho, \varphi, \kappa > 0$ exist such that Assumption 7.2.1 – Assumption 7.2.4 hold for $\ell = 0, \dots, L$, it follows immediately from the (7.30) and the discussion of the cost above that

$$\varepsilon(\mathcal{Q}^{\mathrm{ML}}_{\boldsymbol{n},\boldsymbol{R}}(q))^2 \lesssim h_L^{2\rho} + \sum_{\ell=0}^{L} R_\ell^{-1} n_\ell^{-1/\lambda} h_\ell^\varphi \quad \text{and} \quad \mathcal{C}(\mathcal{Q}^{\mathrm{ML}}_{\boldsymbol{n},\boldsymbol{R}}(q)) \lesssim \sum_{\ell=0}^{L} R_\ell n_\ell h_\ell^{-\kappa}. \tag{7.32}$$

Theorem 7.2.5. *Suppose that constants $\lambda, \rho, \varphi, \kappa > 0$ exist such that Assumption 7.2.1 – Assumption 7.2.4 hold for $\ell = 0, \dots, L$. If the meshes have mesh widths $h_\ell \simeq q^{-\ell}$ for some $q > 1$ and the choice $R_\ell = R$ is made for some $R \in \mathbb{R}$, then for any $\epsilon > 0$, there exists a choice of L and of $N_0, \dots, N_L$ such that*

$$\varepsilon(\mathcal{Q}^{ML}_{\boldsymbol{n},\boldsymbol{R}}(q))^2 \lesssim \epsilon^2 \text{ and } \mathcal{C}(\mathcal{Q}^{ML}_{\boldsymbol{n},\boldsymbol{R}}(q)) \lesssim \begin{cases} \epsilon^{-2\lambda} & \text{if } \varphi\lambda > \kappa, \\ \epsilon^{-2\lambda}(\log_2 \epsilon^{-1})^{\lambda+1} & \text{if } \varphi\lambda = \kappa, \\ \epsilon^{-2\lambda-(\kappa-\varphi\lambda)/\rho} & \text{if } \varphi\lambda < \kappa. \end{cases} \tag{7.33}$$

The proof is analogous to the one presented in [119, Corollary 2]. In fact, Theorem 7.2.5 can be understood as equivalent to [119, Theorem 1 and Corollary 2] with the constants α' and β' defined there equal to $-\infty$ and the dimension d there, due to the assumptions in this section being slightly different, replaced by our κ.

7.2.4 Numerical experiments

In this section we present numerical evidence that the MLQMC method outperforms the MLMC method and the single level QMC and MC methods for gradient calculations involving the elliptic model problem. Assumption 7.2.3 is verified numerically to hold for λ smaller than 1, thus outperforming standard Monte Carlo methods. Practical aspects and implementational details are also briefly discussed.

Problem specification

We consider a spatial domain $D = (0,1)^2$. The gradient is calculated for the target function

$$\widehat{u}(\boldsymbol{x}) = \begin{cases} 1 & \boldsymbol{x} \in [0.25, 0.75] \times [0.25, 0.75], \\ 0 & \text{otherwise}, \end{cases}$$

in the control point $z(\boldsymbol{x}) = 5(1-\cos(2\pi\boldsymbol{x}_1))(1-\cos(2\pi\boldsymbol{x}_2))$, see Figure 7.4. The stochastic diffusion coefficient has a Matérn covariance

$$r_{\text{cov}}(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \Big(\sqrt{2\nu}\frac{\|x - x'\|_2}{\lambda_c}\Big)^\nu K_\nu\Big(\sqrt{2\nu}\frac{\|x - x'\|_2}{\lambda_c}\Big), \tag{7.34}$$

where Γ is the gamma function and K_ν is the modified Bessel function of the second kind. Here, σ^2 is the variance, λ_c the correlation length and ν a parameter determining the smoothness of the resulting field samples. We choose $\sigma^2 = 0.1$, $\lambda_c = 1$ and consider two values for ν. Problem 1 has $\nu = 0.5$, which yields an exponential covariance, and Problem 2 has $\nu = 2.5$. These particular parameters were also investigated in a MLQMC context in [119].

Level definitions, CE and FE details

We consider 7 levels for which the FE grids are regular rectangular grids having size $(2^{2+\ell}+1) \times (2^{2+\ell}+1), \ell = 0, \dots, 6$, including the boundary points. For the CE, we consider coarser regular rectangular grids of size $(2^\ell + 1) \times (2^\ell + 1), \ell = 0, \dots, 6$. The resulting FE and stochastic CE dimensions are shown in Figure 7.5a. The stochastic dimension is different for the two model problems since the different stochastic field parameters necessitate a different amount of padding in the CE method. The resulting computational

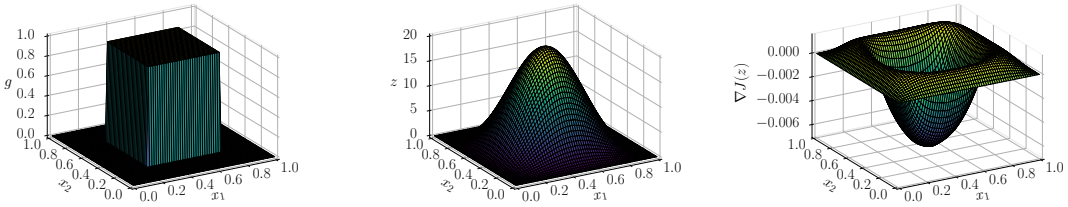
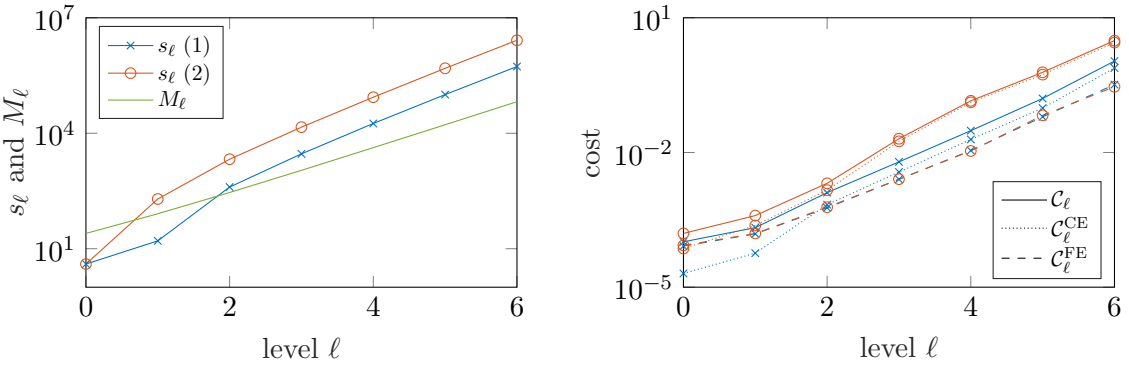


Figure 7.4: Target g, control z and gradient $\nabla J(z)$.



(a) Stochastic dimension s_ℓ and FE dimension M_ℓ as a function of level ℓ. M_ℓ is the same for both problems.

(b) At each level, the CE sampling cost $\mathcal{C}_\ell^{\mathrm{CE}}$, the FE cost $\mathcal{C}_\ell^{\mathrm{FE}}$ and their sum $\mathcal{C}_\ell$, measured in run time.

Figure 7.5: CE and FE details. Problem 1 is marked by blue ×, Problem 2 by red ○.

single threaded performance on an Intel® Core i5–4690K CPU @ 3.50GHz is shown in Figure 7.5b. These costs are only important relative to one another; the scaling of the figure has no further consequence. The CE and FE costs are comparable, which is the reason for choosing the CE grid slightly coarser than the FE grid.

As indicated in (7.30), the RMSE is composed of a variance term, due to the QMC quadrature error, and a bias term due to the FE discretization. The maximum level L determines the bias. For the numerical experiments in this section however, we make abstraction of the FE error and study only the QMC quadrature error. The levels we use and thus L are fixed. This does not fundamentally alter the computational cost for a multilevel methods (MLQMC or MLMC), since the number of samples is small on any additional fine levels. Furthermore, in a context of optimization, fixing the levels is a natural thing to do since it allows an optimization algorithm access to gradients at a known and consistent discretization level, independent of the requested tolerance ϵ, which, for performance reasons, may differ from optimization step to optimization step [157].

QMC details

We use $R = R_\ell = 10$ random shifts for the single level QMC estimator, as well as for each level in the MLQMC estimator. We use an embedded lattice rule with a generating vector that can be found online at [112, lattice-32001-1024-1048576.3600.txt]. This rule works optimally for a number of QMC points $n_\ell \in [2^{10}, 2^{20}] = [1024, 1048576]$. Note that this lattice rule is not specifically tuned to the problem at hand, as one could do by incorporating information about certain constants in Section 7.2.5. Even though there is

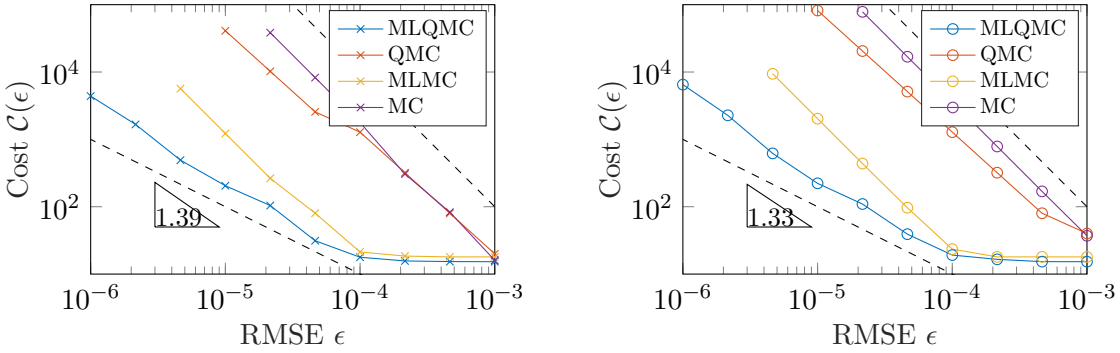
thus no theoretical justification to use this particular lattice rule, numerical experiments in [70] and [119] show that such generic lattice rules have comparable performance. An issue is that the generating vector provided here has length 3600, making it only usable for integrals of dimension up to 3600. Due to the circulant embedding method, the stochastic dimension s_ℓ grows with ℓ, see Assumption 7.2.1. In the experiments that follow, a stochastic dimension in the millions is not uncommon, see Figure 7.5a. The construction of a custom lattice rule tuned to our problem with POD weights (see Section 7.2.5) for all stochastic dimensions is not feasible as the cost of constructing the generating vector using a CBC algorithm scales as $\mathcal{O}(s_\ell^2 n_\ell + s_\ell n_\ell \log n_\ell)$, with s_ℓ the stochastic dimension on level ℓ, see e.g., [72]. Therefore, the generating vector [112] is appended with as many as necessary independent uniformly distributed random integers between 1 and $2^{20} - 1$. Before applying the QMC method, the stochastic dimensions are sorted from most important to least important. The most important dimensions are then handled by the first, high quality elements of the random vector. The importance of a stochastic dimension is taken to be proportional to the corresponding eigenvalue of the circulant matrix C, see Section 7.2.1. As suggested in, e.g., [119], the optimal number of samples to take at each of the L levels, given a tolerance on the QMC quadrature error ϵ, is attained dynamically by Algorithm 6. It ensures that $\mathcal{V}_\ell \simeq n_\ell \mathcal{C}_\ell$, i.e., it ensures that the computational effort required to further reduce the variance contribution $\mathcal{V}_\ell$ at any level is comparable.

Algorithm 6 Determining $\boldsymbol{N} = (N_0, \ldots, N_L)$

1: Set $N_0 = N_1 = \ldots = N_L = 1$.
2: Estimate $\mathcal{V}_0, \ldots, \mathcal{V}_L$ using (7.28)
3: **if** $\sum_{\ell=1}^{L} \mathcal{V}_\ell > \epsilon^2$ **then**
4: Double n_ℓ at ℓ where $\mathcal{V}_\ell/(n_\ell \mathcal{C}_\ell)$ is largest.
5: **end if**
6: (An algorithm with adaptive L could estimate and check the bias here.)

Results

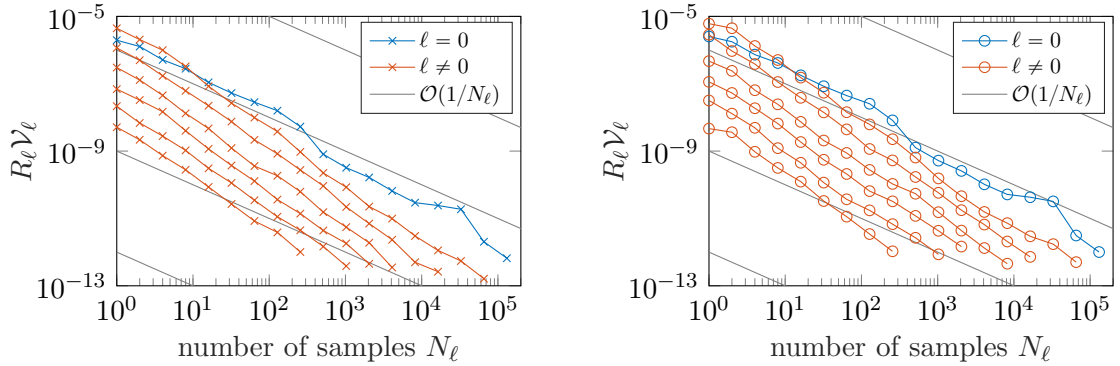
The performance for both problems is shown in Figure 7.6. Clearly, the MLQMC method outperforms the other methods. Note that due to the fixed number of levels L, the MC and MLMC methods follow the typical convergence rate of $\mathcal{O}(\epsilon^{-2})$. If L were not fixed, then smaller and smaller tolerances on ϵ would eventually prompt a refinement of the single grid at which all samples are taken, resulting in a sudden massive increase in computational cost. The rate at which the single level methods become more expensive with decreasing ϵ is thus underestimated in the results shown. This in contrast to the multilevel methods, for which an increase in L would at most incur a moderate cost increase. The flat costs for the multilevel methods for large ϵ are due to warm-up samples.
Section 7.2.4 illustrates Assumption 7.2.4. Shown is $R_\ell \mathcal{V}_\ell$ since that quantity does not depend on the chosen number of shifts. Remark that of course the precision of the numerical estimation (7.28) of $\mathcal{V}_\ell$ does depend on R_ℓ. Clearly, the variance contributions for each of the levels go down faster than the MC rate of n_ℓ^{-1}. Furthermore, the variances decay with ℓ as some power of h_ℓ. Curiously, for $\ell = 0$, the variances take a large N_0 before their faster decay starts. Should this be a problem in practice, a method different from the QMC method could be used to estimate at the coarsest level, especially considering that the stochastic dimension there is very small (4 in this case), see Figure 7.5a.



(a) Problem 1: $\nu = 0.5, \sigma^2 = 0.1, \lambda_c = 1$.

(b) Problem 2: $\nu = 2.5, \sigma^2 = 0.1, \lambda_c = 1$.

Figure 7.6: Performance of the MLQMC method compared with the MLMC method and their single level counterparts. The cost is expressed in equivalent finest level PDE solves.



(a) Problem 1: $\nu = 0.5, \sigma^2 = 0.1, \lambda_c = 1$.

(b) Problem 2: $\nu = 2.5, \sigma^2 = 0.1, \lambda_c = 1$.

Figure 7.7: MSE contribution $\mathcal{V}_\ell$ as a function of the number of QMC samples n_ℓ used for each of the $R_\ell = R = 10$ shifts. Shown is $R_\ell \mathcal{V}_\ell$, since this quantity does not depend on R_ℓ. Lower lines correspond to finer levels, except in the case $\ell = 0$ for low N_0.

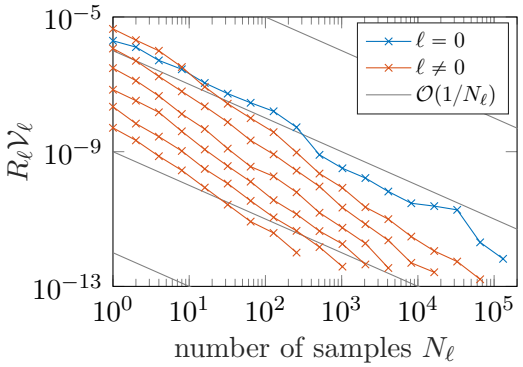


Figure 7.8: MSE contribution $\mathcal{V}_\ell$ as a function of the total number of QMC samples Rn_ℓ (full line). The analogue for the MC method is also provided (dashed line). Shown for problem 1 ($\nu = 0.5, \sigma^2 = 0.1, \lambda_c = 1$).

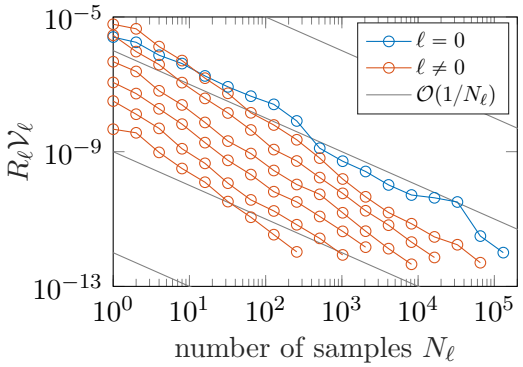


Figure 7.9: MSE contribution $\mathcal{V}_\ell$ as a function of the total number of QMC samples Rn_ℓ (full line). The analogue for the MC method is also provided (dashed line). Shown for problem 1 ($\nu = 0.5, \sigma^2 = 0.1, \lambda_c = 1$).

7.2.5 Convergence analysis

This section provides a theoretical justification for Assumption 7.2.3. In this analysis, we confine ourselves to the following assumption:

Assumption 7.2.6. *There exists some $L_{\max} \in \mathbb{N}$ such that $s_{\ell+1} = s_\ell$ for all $\ell \geqslant L_{\max}$, i.e., there is a finest CE grid with $s_{L_{\max}}$ points.*

This is a stronger assumption on s_ℓ than Assumption 7.2.1. The FE grid can still be refined for $\ell \geqslant L_{\max}$. A similar assumptions is made in [119], where the authors analyze a MLQMC method to approximate expected values of elliptic PDEs with lognormal random inputs parameterized by a Karhunen–Loève expansion with a fixed number of terms. Our restriction is less strict in the sense that our analysis allows simultaneous refinement of the CE grid up to a fixed arbitrary fine level $L_{\max}$.

The novelties in the regularity analysis are the following. Firstly, we analyze the adjoint equation, which has a right-hand side that depends on the uncertain variables through the solution of the state equation. Moreover, our integration error is stated in terms of L^2 errors over the spatial domain D, we do not apply a bounded linear functional to the PDE solution. Both aspects occur in [76], where the regularity analysis for the solution of the adjoint equation is provided with a complete error analysis for the single level method with uniformly distributed parameters. In this manuscript we study lognormally distributed parameters using a multilevel estimator. While multilevel methods are well studied for problems with deterministic right-hand sides, the regularity anaylsis for a multilevel method has not been studied for the problem class considered in this manuscript. Secondly, we sample the random field using the circulant embedding method instead of a series expansion. We therefore first show that the linearly interpolated random field inherits important properties from the true random field.

Properties of the random field

For $\beta \in (0,1]$, we denote by $C^\beta(\overline{D})$ the space of Hölder continuous functions on $\overline{D}$ with exponent β and norm $\|v\|_{C^\beta(\overline{D})} := \sup_{\boldsymbol{x}\in\overline{D}} |v(\boldsymbol{x})| + |v|_{C^\beta(\overline{D})}$ with seminorm $|v|_{C^\beta(\overline{D})} := \sup_{\boldsymbol{x}_1,\boldsymbol{x}_2\in\overline{D},\boldsymbol{x}_1\neq\boldsymbol{x}_2} |v(\boldsymbol{x}_1)-v(\boldsymbol{x}_2)|/\|\boldsymbol{x}_1-\boldsymbol{x}_2\|^\beta < \infty$. The space $L^p(\Omega,X)$ denotes the Bochner space of all random fields in a separable Banach space X with bounded p-th moments over Ω, i.e., $L^p(\Omega,X)$ contains strongly measurable functions that have finite norm given by

$$\|v\|_{L^p(\Omega,X)} := \begin{cases} \left(\int_\Omega \|v\|_X^p \mathrm{d}\mathbb{P}\right)^{1/p}, & \text{for } p<\infty\,, \\ \operatorname{ess\,sup}_{\omega\in\Omega}\|v\|_X, & \text{for } p=\infty\,. \end{cases}$$

The variational form (7.8) is based on the Sobolev space V with norm

$$\|v\|_V := \|\nabla v\|_{L^2(D)}$$

and dual space $H^{-1}(D) := V'$. By $|\cdot|$ we denote the Euclidean norm in $\mathbb{R}^n$.
We suppose that the stochastic field has the property

$$Z(\cdot,\omega) \in C^\beta(\overline{D}), \text{ for some } \beta\in(0,1] \ \mathbb{P}\text{-a.s.} \tag{7.35}$$

Then, using Fernique's Theorem, one can show (see [23]), that $a \in L^p(\Omega, C^\beta(\overline{D}))$ for all $p\in[1,\infty)$ and furthermore that

$$a_{\max}(\omega) := \max_{\boldsymbol{x}\in\overline{D}} a(\boldsymbol{x},\omega) \in L^p(\Omega) \qquad \text{and} \qquad \frac{1}{a_{\min}(\omega)} := \frac{1}{\min_{\boldsymbol{x}\in\overline{D}} a(\boldsymbol{x},\omega)} \in L^p(\Omega)\,,$$

for all $p\in[1,\infty)$, i.e., $0 < a_{\min}(\omega) \leqslant a_{\max}(\omega) < \infty$ $\mathbb{P}$-a.s. Clearly, for $\boldsymbol{x}$ in any set of points $\{\boldsymbol{x}_i\}_{i=1}^m \subset D$, we have

$$0 < a_{\min}(\omega) \leqslant \min_{\boldsymbol{x}\in\{\boldsymbol{x}_i\}_{i=1}^m} a(\boldsymbol{x},\omega) \leqslant \max_{\boldsymbol{x}\in\{\boldsymbol{x}_i\}_{i=1}^m} a(\boldsymbol{x},\omega) \leqslant a_{\max}(\omega) < \infty \quad \mathbb{P}\text{-}a.s.$$

Hence for any realization of the linearly interpolated field $a_s^{\boldsymbol{y}}(\boldsymbol{x})$ (see (7.16)), which is exact on $\{\boldsymbol{x}_i\}_{i=1}^m$, the bounds can only be tighter

$$0 < a_{\min}^{\boldsymbol{y}} \leqslant \min_{\boldsymbol{x}\in\{\boldsymbol{x}_i\}_{i=1}^m} a_s^{\boldsymbol{y}}(\boldsymbol{x}) \leqslant \min_{\boldsymbol{x}\in\overline{D}} a_s^{\boldsymbol{y}}(\boldsymbol{x}) \leqslant \max_{\boldsymbol{x}\in\overline{D}} a_s^{\boldsymbol{y}}(\boldsymbol{x}) \leqslant \max_{\boldsymbol{x}\in\{\boldsymbol{x}_i\}_{i=1}^m} a_s^{\boldsymbol{y}}(\boldsymbol{x}) \leqslant a_{\max}^{\boldsymbol{y}} < \infty\,,$$

where we use the convention $a_{\min}^{\boldsymbol{y}} := a_{\min}(\omega)$ and $a_{\max}^{\boldsymbol{y}} := a_{\max}(\omega)$.
The piecewise linear interpolant $a_s^{\boldsymbol{y}}(\boldsymbol{x})$ is clearly Lipschitz, i.e., $a_s^{\boldsymbol{y}}(\boldsymbol{x}) \in C^\beta(\overline{D})$ for $\beta = 1$ (and thus also for all $\beta < 1$). In fact, since $w_{k,\boldsymbol{x}}$ in (7.16) are first-order polynomials in $\boldsymbol{x}$,

$$\begin{aligned} |a_s^{\boldsymbol{y}}|_{C^1(\overline{D})} \leqslant \|a_s^{\boldsymbol{y}}\|_{W^{1,\infty}(D)} &= \max\{\|a_s^{\boldsymbol{y}}\|_{L^\infty(D)}, \|\nabla a_s^{\boldsymbol{y}}\|_{L^\infty(D)}\} \\ &\leqslant \max\left\{a_{\max}^{\boldsymbol{y}}, \sum_{k=1}^{2^d} \|\nabla w_{k,\boldsymbol{x}}\|_{L^\infty(D)} a_{\max}^{\boldsymbol{y}}\right\} \\ &= C_d a_{\max}^{\boldsymbol{y}}, \end{aligned} \tag{7.36}$$

where $C_d := \max\{1, \sum_{k=1}^{2^d} \|\nabla w_{k,\boldsymbol{x}}\|_{L^\infty(D)}\}$.
In order to analyze the regularity w.r.t. the uncertain variables, we will denote

$$\boldsymbol{b} := (b_1,\ldots,b_s) \text{ with } b_j := \|B_{\cdot,j}\|_{\max}, \tag{7.37}$$

i.e, the maximum of the j-th column of the matrix B in (7.13).
Since $a_s^{\boldsymbol{y}}(\boldsymbol{x}_i) = \exp\left(\sum_{j=1}^s B_{i,j} y_j + \overline{Z}_i\right) \geqslant 0$ for any of the uniform CE grid points $\boldsymbol{x}_i \in \{\boldsymbol{x}_i\}_{i=1}^m$, see (7.15), the chain rule results in $|\partial^{\boldsymbol{\nu}} a_s^{\boldsymbol{y}}(\boldsymbol{x}_i)| = a_s^{\boldsymbol{y}}(\boldsymbol{x}_i) \prod_{j=1}^s |B_{i,j}^{\nu_j}| \leqslant a_s^{\boldsymbol{y}}(\boldsymbol{x}_i)\, \boldsymbol{b}^{\boldsymbol{\nu}}$.
With the intermediate points included, the random field is specified by the interpolation (7.16). Since $w_{k,\boldsymbol{x}} \geqslant 0$ for all $k = 1, \dots 2^d$ and $\boldsymbol{x} \in D$, this result generalizes to all $\boldsymbol{x} \in D$:

$$|\partial^{\boldsymbol{\nu}} a_s^{\boldsymbol{y}}(\boldsymbol{x})| = \sum_{k=1}^{2^d} w_{k,\boldsymbol{x}} |\partial^{\boldsymbol{\nu}} a_s^{\boldsymbol{y}}(\boldsymbol{x}_{k,\boldsymbol{x}})| \leqslant \sum_{k=1}^{2^d} w_{k,\boldsymbol{x}} a_s^{\boldsymbol{y}}(\boldsymbol{x}_{k,\boldsymbol{x}}) \boldsymbol{b}^{\boldsymbol{\nu}} = a_s^{\boldsymbol{y}}(\boldsymbol{x}) \boldsymbol{b}^{\boldsymbol{\nu}}. \tag{7.38}$$

It then follows immediately that

$$\left\| \frac{\partial^{\boldsymbol{\nu}} a_s^{\boldsymbol{y}}(\boldsymbol{x})}{a_s^{\boldsymbol{y}}(\boldsymbol{x})} \right\|_{L^\infty(D)} \leqslant \boldsymbol{b}^{\boldsymbol{\nu}}\,. \tag{7.39}$$

Furthermore,

$$\begin{aligned}
\left\| \nabla \left(\frac{\partial^{\boldsymbol{\nu}} a_s^{\boldsymbol{y}}(\boldsymbol{x})}{a_s^{\boldsymbol{y}}(\boldsymbol{x})} \right) \right\|_{L^\infty(D)} &= \left\| \frac{a_s^{\boldsymbol{y}}(\boldsymbol{x}) \nabla(\partial^{\boldsymbol{\nu}} a_s^{\boldsymbol{y}}(\boldsymbol{x})) - \nabla a_s^{\boldsymbol{y}}(\boldsymbol{x}) \partial^{\boldsymbol{\nu}} a_s^{\boldsymbol{y}}(\boldsymbol{x})}{(a_s^{\boldsymbol{y}}(\boldsymbol{x}))^2} \right\|_{L^\infty(D)} \\
&\leqslant \left\| \frac{a_s^{\boldsymbol{y}}(\boldsymbol{x}) \nabla\left(a_s^{\boldsymbol{y}}(\boldsymbol{x}) \boldsymbol{b}^{\boldsymbol{\nu}}\right)}{(a_s^{\boldsymbol{y}}(\boldsymbol{x}))^2} \right\|_{L^\infty(D)} + \left\| \frac{\nabla a_s^{\boldsymbol{y}}(\boldsymbol{x}) \left(a_s^{\boldsymbol{y}}(\boldsymbol{x}) \boldsymbol{b}^{\boldsymbol{\nu}}\right)}{(a_s^{\boldsymbol{y}}(\boldsymbol{x}))^2} \right\|_{L^\infty(D)} \\
&= \left\| \frac{\nabla\left(a_s^{\boldsymbol{y}}(\boldsymbol{x}) \boldsymbol{b}^{\boldsymbol{\nu}}\right)}{a_s^{\boldsymbol{y}}(\boldsymbol{x})} \right\|_{L^\infty(D)} + \left\| \frac{\nabla a_s^{\boldsymbol{y}}(\boldsymbol{x}) \boldsymbol{b}^{\boldsymbol{\nu}}}{a_s^{\boldsymbol{y}}(\boldsymbol{x})} \right\|_{L^\infty(D)} \\
&= 2 \left\| \frac{\nabla a_s^{\boldsymbol{y}}(\boldsymbol{x}) \boldsymbol{b}^{\boldsymbol{\nu}}}{a_s^{\boldsymbol{y}}(\boldsymbol{x})} \right\|_{L^\infty(D)} = 2 \left\| \frac{\nabla a_s^{\boldsymbol{y}}(\boldsymbol{x})}{a_s^{\boldsymbol{y}}(\boldsymbol{x})} \right\|_{L^\infty(D)} \boldsymbol{b}^{\boldsymbol{\nu}} \leqslant 2 \boldsymbol{b}^{\boldsymbol{\nu}} \frac{C_d a_{\max}^{\boldsymbol{y}}}{a_{\min}^{\boldsymbol{y}}}\,,
\end{aligned} \tag{7.40}$$

where the last inequality follows from (7.36).
The following lemma is based on [72, Lemma 1] and bounds the interpolation error for functions in $C^\beta(\overline{D})$ for some $\beta \in (0, 1]$.

Lemma 7.2.7. *Let $a \in C^\beta(\overline{D})$ for some $\beta \in (0, 1]$. Let b be the linear interpolant of a in interpolation points $\{\boldsymbol{x}_i\}_{i=1}^m$ forming some uniform mesh with mesh width $\hat{h}$, i.e., $b(\boldsymbol{x}) = \mathcal{I}(a; \{\boldsymbol{x}_i\}_{i=1}^m)(\boldsymbol{x})$. Then we have for any $\boldsymbol{x} \in D$ that*

$$|a(\boldsymbol{x}) - b(\boldsymbol{x})| \leqslant (\sqrt{d}\hat{h})^\beta |a|_{C^\beta(\overline{D})}.$$

Proof. The statement follows from

$$\begin{aligned}
|a(\boldsymbol{x}) - b(\boldsymbol{x})| = |a(\boldsymbol{x}) - \sum_{k=1}^{2^d} w_{k,\boldsymbol{x}} a(\boldsymbol{x}_{k,\boldsymbol{x}})| &= |\sum_{k=1}^{2^d} w_{k,\boldsymbol{x}} (a(\boldsymbol{x}) - a(\boldsymbol{x}_{k,\boldsymbol{x}}))| \\
&\leqslant \sum_{k=1}^{2^d} w_{k,\boldsymbol{x}} |a(\boldsymbol{x}) - a(\boldsymbol{x}_{k,\boldsymbol{x}})| \leqslant \sum_{k=1}^{2^d} w_{k,\boldsymbol{x}} |a|_{C^\beta(\overline{D})} |x - \boldsymbol{x}_{k,\boldsymbol{x}}|^\beta \\
&\leqslant \sum_{k=1}^{2^d} w_{k,\boldsymbol{x}} |a|_{C^\beta(\overline{D})} (\sqrt{d}\hat{h})^\beta\,,
\end{aligned}$$

since $\sum_{k=1}^{2^d} w_{k,\boldsymbol{x}} = 1$. □

The above lemma can be applied to the diffusion coefficient and its interpolation. Taking a above to be the exact diffusion coefficient $a(\cdot,\omega)$ for some ω and b its interpolation $a_s^{\boldsymbol{y}}$, as defined in (7.16), we find

$$|a(\boldsymbol{x},\omega) - a_s^{\boldsymbol{y}}(\boldsymbol{x})| \leqslant (\sqrt{d}\hat{h})^\beta |a(\cdot,\omega)|_{C^\beta(\overline{D})}.$$

The quantity $\hat{h}$ is then the mesh width of the uniform CE mesh on which the diffusion coefficient is sampled exactly. Furthermore, since we use nested but not necessarily equal CE grids, the mesh width depends on ℓ. Denoting the CE mesh width at level ℓ by $\hat{h}_\ell$, we have by the above lemma that

$$|a(\boldsymbol{x},\omega) - a_{s_\ell}^{\boldsymbol{y}}(\boldsymbol{x})| \lesssim (\sqrt{d}\hat{h}_\ell)^\beta |a(\cdot,\omega)|_{C^\beta(\overline{D})}. \tag{7.41}$$

Lemma 7.2.8. *Let $a_{s_\ell}^{\boldsymbol{y}}$ be generated with the CE method from $\boldsymbol{y}$ and let $a_{s_{\ell-1}}^{\boldsymbol{y}}$ be its interpolation in the points $\{\boldsymbol{x}_i\}_{i=1}^{m_{\ell-1}}$ forming some uniform mesh with mesh width $\hat{h}_\ell$, i.e., $a_{s_{\ell-1}}^{\boldsymbol{y}}(\boldsymbol{x}) = \mathcal{I}(a_{s_\ell}^{\boldsymbol{y}}; \{\boldsymbol{x}_i\}_{i=1}^{m_{\ell-1}})(\boldsymbol{x})$. Then we have for any $\boldsymbol{x} \in D$ that*

$$|\partial^{\boldsymbol{\nu}}(a_{s_\ell}^{\boldsymbol{y}}(\boldsymbol{x}) - a_{s_{\ell-1}}^{\boldsymbol{y}}(\boldsymbol{x}))| \leqslant \hat{h}_\ell \sqrt{d} C_d a_{\max}^{\boldsymbol{y}} \boldsymbol{b}^{\boldsymbol{\nu}}.$$

Proof. By linearity we obtain

$$\begin{aligned}\partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}(\boldsymbol{x}) = \partial^{\boldsymbol{\nu}} \mathcal{I}(a_{s_\ell}^{\boldsymbol{y}}; \{\boldsymbol{x}_i\}_{i=1}^{m_\ell})(\boldsymbol{x}) = \partial^{\boldsymbol{\nu}} \sum_{k=1}^{2^d} w_{k,\boldsymbol{x}} a_{s_\ell}^{\boldsymbol{y}}(\boldsymbol{x}_{k,\boldsymbol{x}}) &= \sum_{k=1}^{2^d} w_{k,\boldsymbol{x}} \partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}(\boldsymbol{x}_{k,\boldsymbol{x}}) \\ &= \mathcal{I}(\partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}; \{\boldsymbol{x}_i\}_{i=1}^{m_\ell})(\boldsymbol{x}),\end{aligned}$$

where points $\boldsymbol{x}_{k,\boldsymbol{x}}$ denote the vertex values surrounding $\boldsymbol{x} \in D$. In particular, the $\boldsymbol{\nu}$-th partial derivative of the piecewise linear interpolation of the field remains piecewise linear and is hence Lipschitz continuous. Moreover, we have

$$\partial^{\boldsymbol{\nu}} a_{s_{\ell-1}}^{\boldsymbol{y}}(\boldsymbol{x}) := \mathcal{I}(\partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}; \{\boldsymbol{x}_i\}_{i=1}^{m_{\ell-1}})(\boldsymbol{x}).$$

We conclude that

$$\begin{aligned}|\partial^{\boldsymbol{\nu}}(a_{s_\ell}^{\boldsymbol{y}}(\boldsymbol{x}) - a_{s_{\ell-1}}^{\boldsymbol{y}}(\boldsymbol{x}))| &= |\partial^{\boldsymbol{\nu}}(a_{s_\ell}^{\boldsymbol{y}}(\boldsymbol{x}) - \mathcal{I}(a_{s_\ell}^{\boldsymbol{y}}; \{\boldsymbol{x}_i\}_{i=1}^{m_{\ell-1}})(\boldsymbol{x}))| \\ &= |\partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}(\boldsymbol{x}) - \partial^{\boldsymbol{\nu}} \mathcal{I}(a_{s_\ell}^{\boldsymbol{y}}; \{\boldsymbol{x}_i\}_{i=1}^{m_{\ell-1}})(\boldsymbol{x})| \\ &= |\partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}(\boldsymbol{x}) - \mathcal{I}(\partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}; \{\boldsymbol{x}_i\}_{i=1}^{m_{\ell-1}})(\boldsymbol{x})|.\end{aligned}$$

Since $\partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}(\boldsymbol{x}) \in C^\beta$, by Lemma 7.2.7, we have

$$|\partial^{\boldsymbol{\nu}}(a_{s_\ell}^{\boldsymbol{y}}(\boldsymbol{x}) - a_{s_{\ell-1}}^{\boldsymbol{y}}(\boldsymbol{x}))| \leqslant (\sqrt{d}\hat{h}_\ell)^\beta |\partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}|_{C^\beta(\overline{D})} < \infty,$$

which is particularly true for $\beta = 1$. It remains to find a bound for $|\partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}|_{C^\beta(\overline{D})}$ in terms of $a_{\max}^{\boldsymbol{y}}$. To this end, we note

$$|\partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}|_{C^\beta(\overline{D})} \leqslant \|\partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}\|_{W^{1,\infty}(D)} = \max\{\|\partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}\|_{L^\infty(D)}, \|\nabla \partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}\|_{L^\infty(D)}\}.$$

We have

$$\|\partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}\|_{L^\infty(D)} \leqslant \left\| \sum_{k=1}^{2^d} w_{k,\boldsymbol{x}} \partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}(\boldsymbol{x}_{k,\boldsymbol{x}}) \right\|_{L^\infty(D)}$$

$$= \left\| \sum_{k=1}^{2^d} w_{k,\boldsymbol{x}} a_{s_\ell}^{\boldsymbol{y}}(\boldsymbol{x}_{k,\boldsymbol{x}}) \prod_{j=1}^{s} B_{(k,\boldsymbol{x}),j}^{\nu_j} \right\|_{L^\infty(D)}$$
$$\leqslant a_{\max}^{\boldsymbol{y}} \boldsymbol{b}^{\boldsymbol{\nu}},$$

and

$$\left\| \nabla \partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}} \right\|_{L^\infty(D)} \leqslant \left\| \nabla \sum_{k=1}^{2^d} w_{k,\boldsymbol{x}} \partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}(\boldsymbol{x}_{k,\boldsymbol{x}}) \right\|_{L^\infty(D)}$$
$$= \left\| \sum_{k=1}^{2^d} \nabla w_{k,\boldsymbol{x}} a_{s_\ell}^{\boldsymbol{y}}(\boldsymbol{x}_{k,\boldsymbol{x}}) \prod_{j=1}^{s} B_{(k,\boldsymbol{x}),j}^{\nu_j} \right\|_{L^\infty(D)}$$
$$\leqslant \sum_{k=1}^{2^d} \|\nabla w_{k,\boldsymbol{x}}\|_{L^\infty(D)} a_{\max}^{\boldsymbol{y}} \boldsymbol{b}^{\boldsymbol{\nu}}.$$

Combining the two estimates, we get

$$|\partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}|_{C^\beta(\overline{D})} \leqslant \|\partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}\|_{W^{1,\infty}(D)} = \max\left\{\|\partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}\|_{L^\infty(D)}, \|\nabla \partial^{\boldsymbol{\nu}} a_{s_\ell}^{\boldsymbol{y}}\|_{L^\infty(D)}\right\}$$
$$\leqslant \left(1 + \sum_{k=1}^{2^d} \|\nabla w_{k,\boldsymbol{x}}\|_{L^\infty(D)}\right) a_{\max}^{\boldsymbol{y}} \boldsymbol{b}^{\boldsymbol{\nu}}.$$

as required. □

Remark 7.2.9. *The constant* $C_d = \max\{1, \sum_{k=1}^{2^d} \|\nabla w_{k,x}\|_{L^\infty(D)}\}$ *might depend inversely proportional on* $\hat{h}_\ell$ *through the term* $\nabla w_{k,x}$*. Due to Assumption 7.2.6,* C_d *can be chosen as the minimum of* $\max\{1, \sum_{k=1}^{2^d} \|\nabla w_{k,x}\|_{L^\infty(D)}\}$ *over all levels* $\ell = 1, \ldots, L_{\max}$*.*

Bounds on partial derivatives of $u_s^{\boldsymbol{y}}$ and $q_s^{\boldsymbol{y}}$

The error estimates for the QMC method require bounds on the partial derivatives of the integrands in (7.29), as we will see in Section 7.2.5 below. We introduce the frequently used notation

$$C_q^{\boldsymbol{y}} := \max\left(1, \frac{c_1 c_2}{a_{\min}^{\boldsymbol{y}}}\right) \qquad \text{and} \qquad C_{zg} := \left(\|z\|_{V'} + \|\widehat{u}\|_{V'}\right),$$

where $c_1, c_2 > 0$ are the embedding constants from (4.6) – (4.7). Note that $C_q^{\boldsymbol{y}} \leqslant 1 + \frac{c_1 c_2}{a_{\min}^{\boldsymbol{y}}} \in L^p(\Omega)$ because $1/a_{\min}^{\boldsymbol{y}} \in L^p(\Omega)$ for all $p \in [1, \infty)$.

Lemma 7.2.10. *Let* $u_s^{\boldsymbol{y}}$ *and* $q_s^{\boldsymbol{y}}$ *be as defined previously in* (7.18) – (7.19)*. Then*

$$\|\partial^{\boldsymbol{\nu}} u_s^{\boldsymbol{y}}\|_V \leqslant |\boldsymbol{\nu}|! \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} \frac{\|z\|_{V'}}{a_{\min}^{\boldsymbol{y}}}, \tag{7.42}$$
$$\|\partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}\|_V \leqslant (|\boldsymbol{\nu}| + 1)! \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} \frac{C_q^{\boldsymbol{y}}}{a_{\min}^{\boldsymbol{y}}} \left(\|z\|_{V'} + \|\widehat{u}\|_{V'}\right),$$

with $\boldsymbol{b}$ *as defined in* (7.37)*.*

Proof. Let $f^{\boldsymbol{y}} := u_s^{\boldsymbol{y}} - \widehat{u}$, then taking the $\boldsymbol{\nu}$-th derivative of (7.19) yields by Leibniz product rule

$$\sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \int_D \partial^{\boldsymbol{\nu}-\boldsymbol{m}} a_s^{\boldsymbol{y}}(\boldsymbol{x}) \nabla \partial^{\boldsymbol{m}} q_s^{\boldsymbol{y}}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \,\mathrm{d}x = \int_D \partial^{\boldsymbol{\nu}} f^{\boldsymbol{y}}(\boldsymbol{x})\, v(\boldsymbol{x}) \,\mathrm{d}x$$

for all $v \in V$. Setting $v = \partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}$ and separating out the $\boldsymbol{\nu} = \boldsymbol{m}$ term gives

$$\begin{aligned}
&\int_D a_s^{\boldsymbol{y}}(\boldsymbol{x}) |\nabla \partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}(\boldsymbol{x})|^2 \,\mathrm{d}x \qquad (7.43)\\
&\quad = -\sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}, \boldsymbol{m} \neq \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \int_D \partial^{\boldsymbol{\nu}-\boldsymbol{m}} a_s^{\boldsymbol{y}}(\boldsymbol{x}) \nabla \partial^{\boldsymbol{m}} q_s^{\boldsymbol{y}}(\boldsymbol{x}) \cdot \nabla \partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}(\boldsymbol{x}) \,\mathrm{d}x\\
&\qquad + \int_D \partial^{\boldsymbol{\nu}} f^{\boldsymbol{y}}(\boldsymbol{x}) \partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}(\boldsymbol{x}) \,\mathrm{d}x\\
&\quad = -\sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}, \boldsymbol{m} \neq \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \int_D \left(\frac{\partial^{\boldsymbol{\nu}-\boldsymbol{m}} a_s^{\boldsymbol{y}}}{a_s^{\boldsymbol{y}}}\right) a_s^{\boldsymbol{y}}(\boldsymbol{x}) \nabla \partial^{\boldsymbol{m}} q_s^{\boldsymbol{y}}(\boldsymbol{x}) \cdot \nabla \partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}(\boldsymbol{x}) \,\mathrm{d}x\\
&\qquad + \int_D \partial^{\boldsymbol{\nu}} f^{\boldsymbol{y}}(\boldsymbol{x}) \partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}(\boldsymbol{x}) \,\mathrm{d}x\\
&\quad \leqslant \sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}, \boldsymbol{m} \neq \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \left\| \frac{\partial^{\boldsymbol{\nu}-\boldsymbol{m}} a_s^{\boldsymbol{y}}}{a_s^{\boldsymbol{y}}} \right\|_{L^\infty(D)} \left| \int_D a_s^{\boldsymbol{y}}(\boldsymbol{x}) \nabla \partial^{\boldsymbol{m}} q_s^{\boldsymbol{y}}(\boldsymbol{x}) \cdot \nabla \partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}(\boldsymbol{x}) \,\mathrm{d}x \right|\\
&\qquad + \left| \int_D \partial^{\boldsymbol{\nu}} f^{\boldsymbol{y}}(\boldsymbol{x}) \partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}(\boldsymbol{x}) \,\mathrm{d}x \right|. \qquad (7.44)
\end{aligned}$$

We can now use the Cauchy–Schwarz inequality on both integrals above. For the right-hand side in particular we get, $|\int_D \partial^{\boldsymbol{\nu}} f^{\boldsymbol{y}}(\boldsymbol{x}) \partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}(\boldsymbol{x}) \,\mathrm{d}x| \leqslant \|\partial^{\boldsymbol{\nu}} f^{\boldsymbol{\nu}}\|_{V'} \|\partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}\|_V$ and furthermore

$$\|\partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}\|_V = \left(\int_D |\nabla \partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}(\boldsymbol{x})|^2 \mathrm{d}x \right)^{1/2} \leqslant \frac{1}{(a_{\min}^{\boldsymbol{y}})^{1/2}} \left(\int_D a_s^{\boldsymbol{y}}(\boldsymbol{x}) |\nabla \partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}(\boldsymbol{x})|^2 \mathrm{d}x \right)^{1/2} \qquad (7.45)$$

such that (7.44) can be bounded using (7.39) by

$$\begin{aligned}
&\int_D a_s^{\boldsymbol{y}}(\boldsymbol{x}) |\nabla(\partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}(\boldsymbol{x}))|^2 \,\mathrm{d}x\\
&\leqslant \sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}, \boldsymbol{m} \neq \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{m}} \left(\int_D a_s^{\boldsymbol{y}}(\boldsymbol{x}) |\nabla(\partial^{\boldsymbol{m}} q_s^{\boldsymbol{y}}(\boldsymbol{x}))|^2 \mathrm{d}x \right)^{1/2} \left(\int_D a_s^{\boldsymbol{y}}(\boldsymbol{x}) |\nabla(\partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}(\boldsymbol{x}))|^2 \mathrm{d}x \right)^{1/2}\\
&\quad + \|\partial^{\boldsymbol{\nu}} f^{\boldsymbol{\nu}}\|_{V'} \frac{1}{(a_{\min}^{\boldsymbol{y}})^{1/2}} \left(\int_D a_s^{\boldsymbol{y}}(\boldsymbol{x}) |\nabla(\partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}(\boldsymbol{x}))|^2 \mathrm{d}x \right)^{1/2}.
\end{aligned}$$

Noting that $\int_D a_s^{\boldsymbol{y}}(\boldsymbol{x}) |\nabla(\partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}(\boldsymbol{x}))|^2 \,\mathrm{d}x = \|(a_s^{\boldsymbol{y}})^{1/2} \nabla(\partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}})\|_{L^2(D)}^2$ and cancelling out a common factor, we obtain

$$\begin{aligned}
\underbrace{\|(a_s^{\boldsymbol{y}})^{1/2} \nabla(\partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}})\|_{L^2(D)}}_{\mathbb{A}_{\boldsymbol{\nu}}} &\leqslant \sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}, \boldsymbol{m} \neq \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{m}} \underbrace{\|(a_s^{\boldsymbol{y}})^{1/2} \nabla(\partial^{\boldsymbol{m}} q_s^{\boldsymbol{y}})\|_{L^2(D)}}_{\mathbb{A}_{\boldsymbol{m}}}\\
&\quad + \underbrace{(a_{\min}^{\boldsymbol{y}})^{-1/2} \left(\|\partial^{\boldsymbol{\nu}} f^{\boldsymbol{y}}\|_{V'}\right)}_{\mathbb{B}_{\boldsymbol{\nu}}}. \qquad (7.46)
\end{aligned}$$

We may apply Lemma 4.6.1 to get

$$\begin{aligned}
&\|(a_s^{\boldsymbol{y}})^{1/2}\nabla(\partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}})\|_{L^2(D)} \\
&\quad\leqslant \sum_{\boldsymbol{k}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{k}} \frac{|\boldsymbol{k}|!}{(\ln 2)^{|\boldsymbol{k}|}} \boldsymbol{b}^{\boldsymbol{k}} \frac{1}{(a_{\min}^{\boldsymbol{y}})^{1/2}} \left(\|\partial^{\boldsymbol{\nu}-\boldsymbol{k}} f^{\boldsymbol{y}}\|_{V'}\right) \\
&\quad\leqslant \sum_{\boldsymbol{k}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{k}} \frac{|\boldsymbol{k}|!}{(\ln 2)^{|\boldsymbol{k}|}} \boldsymbol{b}^{\boldsymbol{k}} \frac{1}{(a_{\min}^{\boldsymbol{y}})^{1/2}} \left(\|\partial^{\boldsymbol{\nu}-\boldsymbol{k}} u_s^{\boldsymbol{y}}\|_{V'} + \|\partial^{\boldsymbol{\nu}-\boldsymbol{k}}\widehat{u}\|_{V'}\right) \\
&\quad\leqslant \sum_{\boldsymbol{k}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{k}} \frac{|\boldsymbol{k}|!}{(\ln 2)^{|\boldsymbol{k}|}} \boldsymbol{b}^{\boldsymbol{k}} \frac{1}{(a_{\min}^{\boldsymbol{y}})^{1/2}} \left(c_1 c_2 \|\partial^{\boldsymbol{\nu}-\boldsymbol{k}} u_s^{\boldsymbol{y}}\|_{V} + \|\partial^{\boldsymbol{\nu}-\boldsymbol{k}}\widehat{u}\|_{V'}\right) \qquad (7.47)
\end{aligned}$$

for all multi-indices $\boldsymbol{\nu} \in \mathbb{N}_0^s$. In order to further estimate (7.47), we need an estimate for the partial derivatives of the state PDE solution $u_s^{\boldsymbol{y}}$. This can be obtained as follows: beginning this proof with the $\boldsymbol{\nu}$-th partial derivatives of the weak formulation of (7.18) (instead of (7.19)), one gets an analogous recursion to (7.46) with $q_s^{\boldsymbol{y}}$ replaced by $u_s^{\boldsymbol{y}}$ and $f^{\boldsymbol{y}}$ replaced by the control z:

$$\underbrace{\|(a_s^{\boldsymbol{y}})^{1/2}\nabla(\partial^{\boldsymbol{\nu}} u_s^{\boldsymbol{y}})\|_{L^2(D)}}_{\mathbb{A}_{\boldsymbol{\nu}}} \leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},\boldsymbol{m}\neq\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{m}} \underbrace{\|(a_s^{\boldsymbol{y}})^{1/2}\nabla(\partial^{\boldsymbol{m}} u_s^{\boldsymbol{y}})\|_{L^2(D)}}_{\mathbb{A}_{\boldsymbol{m}}} + \underbrace{\frac{\|\partial^{\boldsymbol{\nu}} z\|_{V'}}{(a_{\min}^{\boldsymbol{y}})^{1/2}}}_{\mathbb{B}_{\boldsymbol{\nu}}}.$$

In this case, the application of Lemma 4.6.1 gives

$$\|(a_s^{\boldsymbol{y}})^{1/2}\nabla(\partial^{\boldsymbol{\nu}} u_s^{\boldsymbol{y}})\|_{L^2(D)} \leqslant \sum_{\boldsymbol{k}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{k}} |\boldsymbol{k}|! \frac{\boldsymbol{b}^{\boldsymbol{k}}}{(\ln 2)^{|\boldsymbol{k}|}} \frac{\|\partial^{\boldsymbol{\nu}-\boldsymbol{k}} z\|_{V'}}{(a_{\min}^{\boldsymbol{y}})^{1/2}} = |\boldsymbol{\nu}|! \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} \frac{\|z\|_{V'}}{(a_{\min}^{\boldsymbol{y}})^{1/2}}.$$

Then, (7.42) follows directly from $(a_{\min}^{\boldsymbol{y}})^{1/2}\|\partial^{\boldsymbol{\nu}} u_s^{\boldsymbol{y}}\|_V \leqslant \|(a_s^{\boldsymbol{y}})^{1/2}\nabla(\partial^{\boldsymbol{\nu}} u_s^{\boldsymbol{y}})\|_{L^2(D)}$. Using (7.42) we can now further estimate (7.47) to get

$$\begin{aligned}
&\|(a_s^{\boldsymbol{y}})^{1/2}\nabla(\partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}})\|_{L^2(D)} \\
&\quad\leqslant \sum_{\boldsymbol{k}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{k}} \frac{|\boldsymbol{k}|!}{(\ln 2)^{|\boldsymbol{k}|}} \boldsymbol{b}^{\boldsymbol{k}} \frac{1}{(a_{\min}^{\boldsymbol{y}})^{1/2}} \left(c_1 c_2 |\boldsymbol{\nu}-\boldsymbol{k}|! \frac{\boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{k}}}{(\ln 2)^{|\boldsymbol{\nu}-\boldsymbol{k}|}} \frac{\|z\|_{V'}}{a_{\min}} + \|\partial^{\boldsymbol{\nu}-\boldsymbol{k}}\widehat{u}\|_{V'}\right).
\end{aligned}$$

Note that g is independent of $\boldsymbol{y}$, i.e., we have for $\boldsymbol{\nu} \leqslant \boldsymbol{k}$

$$\begin{aligned}
\|\partial^{\boldsymbol{\nu}-\boldsymbol{k}}\widehat{u}\|_{V'} &= \begin{cases} \|\widehat{u}\|_{V'} & \boldsymbol{\nu} = \boldsymbol{k} \\ 0 & \text{else} \end{cases} \\
&\leqslant |\boldsymbol{\nu}-\boldsymbol{k}|! \frac{\boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{k}}}{(\ln 2)^{|\boldsymbol{\nu}-\boldsymbol{k}|}} \|\widehat{u}\|_{V'}.
\end{aligned}$$

This and setting $C_q^{\boldsymbol{y}} := \max\left(\frac{c_1 c_2}{a_{\min}^{\boldsymbol{y}}}, 1\right)$ and $C_{zg} := \|z\|_{V'} + \|\widehat{u}\|_{V'}$ gives

$$\begin{aligned}
\|(a_s^{\boldsymbol{y}})^{1/2}\nabla(\partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}})\|_{L^2(D)} &\leqslant \sum_{\boldsymbol{k}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{k}} |\boldsymbol{k}|! \frac{\boldsymbol{b}^{\boldsymbol{k}}}{(\ln 2)^{|\boldsymbol{k}|}} |\boldsymbol{\nu}-\boldsymbol{k}|! \frac{\boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{k}}}{(\ln 2)^{|\boldsymbol{\nu}-\boldsymbol{k}|}} \frac{C_q^{\boldsymbol{y}}}{(a_{\min}^{\boldsymbol{y}})^{1/2}} C_{zg} \\
&= (|\boldsymbol{\nu}|+1)! \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} \frac{C_q^{\boldsymbol{y}}}{(a_{\min}^{\boldsymbol{y}})^{1/2}} C_{zg},
\end{aligned}$$

where the last equality follows from (4.77). The assertion then follows from (7.45). □

Lemma 7.2.11. *Let Δ be the Laplace operator. Under the assumptions of the previous lemma, it holds that*

$$\|\Delta(\partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}})\|_{L^2(D)} \leqslant \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} \frac{(|\boldsymbol{\nu}|+4)!}{(|\boldsymbol{\nu}|+2)(|\boldsymbol{\nu}|+3)} \frac{\widetilde{C}^{\boldsymbol{y}} C_q^{\boldsymbol{y}}}{a_{\min}^{\boldsymbol{y}}} (\|z\|_{V'} + \|\widehat{u}\|_{V'}),$$

where $\widetilde{C}^{\boldsymbol{y}} = \max\left(1, 2\,\frac{C_d a_{\max}^{\boldsymbol{y}}}{a_{\min}^{\boldsymbol{y}}}\right)$.

Proof. We have

$$\begin{aligned}\partial^{\boldsymbol{\nu}} f^{\boldsymbol{y}}(\boldsymbol{x}) &= \partial^{\boldsymbol{\nu}}\big(-\nabla\cdot(a_s^{\boldsymbol{y}}(\boldsymbol{x})\nabla q_s^{\boldsymbol{y}}(\boldsymbol{x}))\big)\\ &= -\nabla\cdot\partial^{\boldsymbol{\nu}}(a_s^{\boldsymbol{y}}(\boldsymbol{x})\nabla q_s^{\boldsymbol{y}}(\boldsymbol{x}))\,.\end{aligned}$$

Thus we get by Leibniz product rule that

$$-\nabla\cdot\partial^{\boldsymbol{\nu}}(a_s^{\boldsymbol{y}}\nabla q_s^{\boldsymbol{y}}) = -\nabla\cdot\left(\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}(\partial^{\boldsymbol{\nu}-\boldsymbol{m}}a_s^{\boldsymbol{y}})\nabla(\partial^{\boldsymbol{m}}q_s^{\boldsymbol{y}})\right) = \partial^{\boldsymbol{\nu}} f^{\boldsymbol{y}}.$$

Separating out the $\boldsymbol{m}=\boldsymbol{\nu}$ term yields

$$\begin{aligned}k_{\boldsymbol{\nu}} &:= \nabla\cdot(a_s^{\boldsymbol{y}}\nabla(\partial^{\boldsymbol{\nu}}q_s^{\boldsymbol{y}}))\\ &= -\nabla\cdot\left(\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},\boldsymbol{m}\neq\boldsymbol{\nu}}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}(\partial^{\boldsymbol{\nu}-\boldsymbol{m}}a_s^{\boldsymbol{y}})\nabla(\partial^{\boldsymbol{m}}q_s^{\boldsymbol{y}})\right) - \partial^{\boldsymbol{\nu}} f^{\boldsymbol{y}}\\ &= -\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},\boldsymbol{m}\neq\boldsymbol{\nu}}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}\nabla\cdot\left(\frac{\partial^{\boldsymbol{\nu}-\boldsymbol{m}}a_s^{\boldsymbol{y}}}{a_s^{\boldsymbol{y}}}(a_s^{\boldsymbol{y}}\nabla(\partial^{\boldsymbol{m}}q_s^{\boldsymbol{y}}))\right) - \partial^{\boldsymbol{\nu}} f^{\boldsymbol{y}}\\ &= -\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},\boldsymbol{m}\neq\boldsymbol{\nu}}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}\left(\frac{\partial^{\boldsymbol{\nu}-\boldsymbol{m}}a_s^{\boldsymbol{y}}}{a_s^{\boldsymbol{y}}}k_{\boldsymbol{m}} + \nabla\left(\frac{\partial^{\boldsymbol{\nu}-\boldsymbol{m}}a_s^{\boldsymbol{y}}}{a_s^{\boldsymbol{y}}}\right)\cdot(a_s^{\boldsymbol{y}}\nabla(\partial^{\boldsymbol{m}}q_s^{\boldsymbol{y}}))\right) - \partial^{\boldsymbol{\nu}} f^{\boldsymbol{y}},\end{aligned}$$

where we used $\nabla\cdot(AB) = A\nabla\cdot B + \nabla A\cdot B$ in the last equality. We can multiply $k_{\boldsymbol{\nu}}$ by $(a_s^{\boldsymbol{y}})^{-1/2}$ and obtain the bound

$$\begin{aligned}\|(a_s^{\boldsymbol{y}})^{-1/2}k_{\boldsymbol{\nu}}\|_{L^2(D)} \leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},\boldsymbol{m}\neq\boldsymbol{\nu}}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}\Bigg(&\left\|\frac{\partial^{\boldsymbol{\nu}-\boldsymbol{m}}a_s^{\boldsymbol{y}}}{a_s^{\boldsymbol{y}}}\right\|_{L^\infty(D)}\|(a_s^{\boldsymbol{y}})^{-1/2}k_{\boldsymbol{m}}\|_{L^2(D)}\\ &+\left\|\nabla\left(\frac{\partial^{\boldsymbol{\nu}-\boldsymbol{m}}a_s^{\boldsymbol{y}}}{a_s^{\boldsymbol{y}}}\right)\right\|_{L^\infty(D)}\|(a_s^{\boldsymbol{y}})^{1/2}\nabla(\partial^{\boldsymbol{m}}q_s^{\boldsymbol{y}})\|_{L^2(D)}\Bigg) + \|(a_s^{\boldsymbol{y}})^{-1/2}\partial^{\boldsymbol{\nu}} f^{\boldsymbol{y}}\|_{L^2(D)}.\end{aligned}$$

From the assumption that $g\in L^2(D)$ and $z\in V'$ implies $k_{\boldsymbol{0}}\in L^2(D)$. From the inequality above we then deduce by induction w.r.t. $|\boldsymbol{\nu}|$ that $(a_s^{\boldsymbol{y}})^{-1/2}k_{\boldsymbol{\nu}}$ and thus by (7.35) also $k_{\boldsymbol{\nu}}\in L^2(D)$ for all multi-indices $\boldsymbol{\nu}\in\mathbb{N}_0^s$. Using the properties (7.39) and (7.40) of $a_s^{\boldsymbol{y}}$, allows to reformulate the previous inequality as

$$\underbrace{\|(a_s^{\boldsymbol{y}})^{-1/2}k_{\boldsymbol{\nu}}\|_{L^2(D)}}_{\mathbb{A}_{\boldsymbol{\nu}}} \leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},\boldsymbol{m}\neq\boldsymbol{\nu}}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}\boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{m}}\underbrace{\|(a_s^{\boldsymbol{y}})^{-1/2}k_{\boldsymbol{m}}\|_{L^2(D)}}_{\mathbb{A}_{\boldsymbol{m}}} + \mathbb{B}'_{\boldsymbol{\nu}},$$

with

$$\mathbb{B}'_{\boldsymbol{\nu}} := \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},\boldsymbol{m}\neq\boldsymbol{\nu}}\binom{\boldsymbol{\nu}}{\boldsymbol{m}}\left(2\boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{m}}\frac{C_d a_{\max}^{\boldsymbol{y}}}{a_{\min}^{\boldsymbol{y}}}\|(a_s^{\boldsymbol{y}})^{1/2}\nabla(\partial^{\boldsymbol{m}}q_s^{\boldsymbol{y}})\|_{L^2(D)}\right) + \|(a_s^{\boldsymbol{y}})^{-1/2}\partial^{\boldsymbol{\nu}} f^{\boldsymbol{y}}\|_{L^2(D)}.$$

In the next section of the proof we first find a simple expression $\mathbb{B}_{\boldsymbol{\nu}}$ such that $\mathbb{B}'_{\boldsymbol{\nu}} \leqslant \mathbb{B}_{\boldsymbol{\nu}}$ and then apply Lemma 4.6.1 to obtain

$$\mathbb{A}_{\boldsymbol{\nu}} \leqslant \sum_{\boldsymbol{k} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{k}} \frac{|\boldsymbol{k}|!}{(\ln 2)^{|\boldsymbol{k}|}} \boldsymbol{b}^{\boldsymbol{k}} \mathbb{B}_{\boldsymbol{\nu}-\boldsymbol{k}}. \tag{7.48}$$

Introducing $C^{\boldsymbol{y}} := 2\,\frac{C_d a^{\boldsymbol{y}}_{\max}}{a^{\boldsymbol{y}}_{\min}}$ to ease readability, we find, using Lemma 7.2.10,

$$\begin{aligned}
\mathbb{B}'_{\boldsymbol{\nu}} &\leqslant \sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}, \boldsymbol{m} \neq \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \left(C^{\boldsymbol{y}} \boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{m}} \boldsymbol{b}^{\boldsymbol{m}} \frac{(|\boldsymbol{m}|+1)!}{(\ln 2)^{|\boldsymbol{m}|}} \frac{C_q^{\boldsymbol{y}}}{(a^{\boldsymbol{y}}_{\min})^{1/2}} C_{zg} \right) + \|(a_s^{\boldsymbol{y}})^{-1/2} \partial^{\boldsymbol{\nu}} f^{\boldsymbol{y}}\|_{L^2(D)} \\
&= \frac{C^{\boldsymbol{y}} C_q^{\boldsymbol{y}} C_{zg}}{(a^{\boldsymbol{y}}_{\min})^{1/2}} \boldsymbol{b}^{\boldsymbol{\nu}} \sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}, \boldsymbol{m} \neq \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \frac{(|\boldsymbol{m}|+1)!}{(\ln 2)^{|\boldsymbol{m}|}} + \|(a_s^{\boldsymbol{y}})^{-1/2} \partial^{\boldsymbol{\nu}} f^{\boldsymbol{y}}\|_{L^2(D)}\,.
\end{aligned}$$

Using (4.79) finally leads to

$$\mathbb{B}'_{\boldsymbol{\nu}} \leqslant \mathbb{B}_{\boldsymbol{\nu}} := \frac{C^{\boldsymbol{y}} C_q^{\boldsymbol{y}} C_{zg}}{(a^{\boldsymbol{y}}_{\min})^{1/2}} \boldsymbol{b}^{\boldsymbol{\nu}} \frac{(|\boldsymbol{\nu}|+1)!}{(\ln 2)^{|\boldsymbol{\nu}|}} + \|(a_s^{\boldsymbol{y}})^{-1/2} \partial^{\boldsymbol{\nu}} (u_s^{\boldsymbol{y}} - g)\|_{L^2(D)}.$$

Now we apply Lemma 4.6.1, yielding

$$\begin{aligned}
\mathbb{A}_{\boldsymbol{\nu}} &\leqslant \sum_{\boldsymbol{k} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{k}} \frac{|\boldsymbol{k}|!}{(\ln 2)^{|\boldsymbol{k}|}} \boldsymbol{b}^{\boldsymbol{k}} \left(\frac{C^{\boldsymbol{y}} C_q^{\boldsymbol{y}} C_{zg}}{(a^{\boldsymbol{y}}_{\min})^{1/2}} \boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{k}} \frac{(|\boldsymbol{\nu}-\boldsymbol{k}|+1)!}{(\ln 2)^{|\boldsymbol{\nu}-\boldsymbol{k}|}} + \|(a_s^{\boldsymbol{y}})^{-1/2} \partial^{\boldsymbol{\nu}-\boldsymbol{k}} (u_s^{\boldsymbol{y}} - g)\|_{L^2(D)} \right) \\
&\leqslant \sum_{\boldsymbol{k} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{k}} \frac{|\boldsymbol{k}|!}{(\ln 2)^{|\boldsymbol{k}|}} \boldsymbol{b}^{\boldsymbol{k}} \left(\frac{C^{\boldsymbol{y}} C_q^{\boldsymbol{y}} C_{zg}}{(a^{\boldsymbol{y}}_{\min})^{1/2}} \boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{k}} \frac{(|\boldsymbol{\nu}-\boldsymbol{k}|+1)!}{(\ln 2)^{|\boldsymbol{\nu}-\boldsymbol{k}|}} + |\boldsymbol{\nu}-\boldsymbol{k}|! \frac{\boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{k}}}{(\ln 2)^{|\boldsymbol{\nu}-\boldsymbol{k}|}} \frac{C_q^{\boldsymbol{y}} C_{zg}}{(a^{\boldsymbol{y}}_{\min})^{1/2}} \right) \\
&\leqslant \frac{C_q^{\boldsymbol{y}} C_{zg}}{(a^{\boldsymbol{y}}_{\min})^{1/2}} \max(C^{\boldsymbol{y}}, 1)\, \boldsymbol{b}^{\boldsymbol{\nu}} \sum_{\boldsymbol{k} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{k}} \frac{|\boldsymbol{k}|!}{(\ln 2)^{|\boldsymbol{k}|}} \left(\frac{(|\boldsymbol{\nu}-\boldsymbol{k}|+1)!}{(\ln 2)^{|\boldsymbol{\nu}-\boldsymbol{k}|}} + \frac{|\boldsymbol{\nu}-\boldsymbol{k}|!}{(\ln 2)^{|\boldsymbol{\nu}-\boldsymbol{k}|}} \right) \\
&\leqslant \frac{C_q^{\boldsymbol{y}} C_{zg}}{(a^{\boldsymbol{y}}_{\min})^{1/2}} \max(C^{\boldsymbol{y}}, 1) \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} \Big((|\boldsymbol{\nu}|+2)! + (|\boldsymbol{\nu}|+1)! \Big) \\
&= \frac{C_q^{\boldsymbol{y}} C_{zg}}{(a^{\boldsymbol{y}}_{\min})^{1/2}} \max(C^{\boldsymbol{y}}, 1) \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} \frac{(|\boldsymbol{\nu}|+3)!}{|\boldsymbol{\nu}|+2}\,.
\end{aligned}$$

Since $(a_s^{\boldsymbol{y}})^{-1/2} k_{\boldsymbol{\nu}} = (a_s^{\boldsymbol{y}})^{-1/2} \nabla \cdot (a_s^{\boldsymbol{y}} \nabla(\partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}})) = (a_s^{\boldsymbol{y}})^{1/2} \Delta(\partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}) + (a_s^{\boldsymbol{y}})^{-1/2} \nabla a_s^{\boldsymbol{y}} \cdot \nabla(\partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}})$, we have

$$\begin{aligned}
\|(a_s^{\boldsymbol{y}})^{1/2} \Delta(\partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}})\|_{L^2(D)} &\leqslant \|(a_s^{\boldsymbol{y}})^{-1/2} k_{\boldsymbol{\nu}}\|_{L^2(D)} + \|(a_s^{\boldsymbol{y}})^{-1/2} (\nabla a_s^{\boldsymbol{y}} \cdot \nabla(\partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}}))\|_{L^2(D)} \\
&\leqslant \frac{C_q^{\boldsymbol{y}} C_{zg}}{(a^{\boldsymbol{y}}_{\min})^{1/2}} \max(C^{\boldsymbol{y}}, 1) \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} \frac{(|\boldsymbol{\nu}|+3)!}{|\boldsymbol{\nu}|+2} + \frac{C_d a^{\boldsymbol{y}}_{\max}}{a^{\boldsymbol{y}}_{\min}} \|(a_s^{\boldsymbol{y}})^{1/2} \nabla(\partial^{\boldsymbol{\nu}} q_s^{\boldsymbol{y}})\|_{L^2(D)} \\
&\leqslant \frac{C_q^{\boldsymbol{y}} C_{zg}}{(a^{\boldsymbol{y}}_{\min})^{1/2}} \max(C^{\boldsymbol{y}}, 1) \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} \frac{(|\boldsymbol{\nu}|+3)!}{|\boldsymbol{\nu}|+2} + \frac{C_d a^{\boldsymbol{y}}_{\max}}{a^{\boldsymbol{y}}_{\min}} (|\boldsymbol{\nu}|+1)! \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} \frac{C_q^{\boldsymbol{y}} C_{zg}}{(a^{\boldsymbol{y}}_{\min})^{1/2}} \\
&\leqslant \max\big(1, C^{\boldsymbol{y}}, \tfrac{C_d a^{\boldsymbol{y}}_{\max}}{a^{\boldsymbol{y}}_{\min}}\big) \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} \frac{(|\boldsymbol{\nu}|+4)!}{(|\boldsymbol{\nu}|+2)(|\boldsymbol{\nu}|+3)} \frac{C_q^{\boldsymbol{y}} C_{zg}}{(a^{\boldsymbol{y}}_{\min})^{1/2}} \\
&= \widetilde{C}^{\boldsymbol{y}} \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} \frac{(|\boldsymbol{\nu}|+4)!}{(|\boldsymbol{\nu}|+2)(|\boldsymbol{\nu}|+3)} \frac{C_q^{\boldsymbol{y}} C_{zg}}{(a^{\boldsymbol{y}}_{\min})^{1/2}}\,,
\end{aligned}$$

with $\widetilde{C}^{\boldsymbol{y}} = \max(1, C^{\boldsymbol{y}}, \frac{C^{\boldsymbol{y}}}{2}) = \max(1, C^{\boldsymbol{y}})$. The third inequality above follows from Lemma 7.2.10. □

Note that $\widetilde{C}^{\boldsymbol{y}} = \max\left(1, 2\,\frac{C_d a^{\boldsymbol{y}}_{\max}}{a^{\boldsymbol{y}}_{\min}}\right) \leqslant 1 + 2\frac{C_d a^{\boldsymbol{y}}_{\max}}{a^{\boldsymbol{y}}_{\min}} \in L^p(\Omega)$ because $\frac{1}{a^{\boldsymbol{y}}_{\min}}$ and $a^{\boldsymbol{y}}_{\max}$ are both in $L^p(\Omega)$ for all $p \in [1,\infty)$.

Lemma 7.2.12. *Let $q^{\boldsymbol{y}}_{s,h}$ be the unique solution of (7.21). Then, under the assumptions of the previous lemma, it holds that*

$$\begin{aligned}(a^{\boldsymbol{y}}_{\min})^{1/2}\|\partial^{\boldsymbol{k}}(q^{\boldsymbol{y}}_s - q^{\boldsymbol{y}}_{s,h})\|_V &\leqslant \|(a^{\boldsymbol{y}}_s)^{1/2}\nabla\partial^{\boldsymbol{k}}(q^{\boldsymbol{y}}_s - q^{\boldsymbol{y}}_{s,h})\|_{L^2(D)}\\ &\lesssim h\,\frac{\boldsymbol{b}^{\boldsymbol{k}}}{(\ln 2)^{|\boldsymbol{k}|}}\,\frac{(|\boldsymbol{k}|+2)!(|\boldsymbol{k}|+6)}{3}\,\frac{C_q^{\boldsymbol{y}}\widetilde{C}^{\boldsymbol{y}}(a^{\boldsymbol{y}}_{\max})^{1/2}}{a^{\boldsymbol{y}}_{\min}}C_{zg}\,.\end{aligned}$$

Proof. Let $P_h = P_h(\boldsymbol{y}) : V \to V_h : w \mapsto w_h$ denote the parametric FE projection onto V_h which is defined, for arbitrary $w \in V$, by

$$\int_D a^{\boldsymbol{y}}_s\,\nabla(P_h(\boldsymbol{y})w - w)\cdot\nabla v_h \mathrm{d}x = 0, \quad \forall v_h \in V_h. \tag{7.49}$$

In particular, we have $P_h(\boldsymbol{y})w = w_h$ in V_h and $P_h^2(\boldsymbol{y}) = P_h(\boldsymbol{y})$. We conclude, using $\partial^{\boldsymbol{\nu}} w_h \in V_h$ for every $\boldsymbol{\nu} \in \mathbb{N}_0^s$, that $(Id - P_h(\boldsymbol{y}))(\partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}}_h) = 0$. We stress here that, since the parametric FE projection $P_h(\boldsymbol{y})$ depends on $\boldsymbol{y}$, in general

$$\partial^{\boldsymbol{\nu}}(w^{\boldsymbol{y}} - w^{\boldsymbol{y}}_h) \neq (Id - P_h(\boldsymbol{y}))(\partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}}).$$

Thus

$$\begin{aligned}&\|(a^{\boldsymbol{y}}_s)^{1/2}\nabla\partial^{\boldsymbol{k}}(q^{\boldsymbol{y}}_s - q^{\boldsymbol{y}}_{s,h})\|_{L^2(D)}\\ &\quad= \|(a^{\boldsymbol{y}}_s)^{1/2}\nabla P_h(\boldsymbol{y})\partial^{\boldsymbol{k}}(q^{\boldsymbol{y}}_s - q^{\boldsymbol{y}}_{s,h}) + (a^{\boldsymbol{y}}_s)^{1/2}\nabla(Id - P_h(\boldsymbol{y}))\partial^{\boldsymbol{k}}(q^{\boldsymbol{y}}_s - q^{\boldsymbol{y}}_{s,h})\|_{L^2(D)}\\ &\quad\leqslant \|(a^{\boldsymbol{y}}_s)^{1/2}\nabla P_h(\boldsymbol{y})\partial^{\boldsymbol{k}}(q^{\boldsymbol{y}}_s - q^{\boldsymbol{y}}_{s,h})\|_{L^2(D)} + \|(a^{\boldsymbol{y}}_s)^{1/2}\nabla(Id - P_h(\boldsymbol{y}))\partial^{\boldsymbol{k}} q^{\boldsymbol{y}}_s\|_{L^2(D)}.\end{aligned} \tag{7.50}$$

Now applying $\partial^{\boldsymbol{k}}$ to

$$\int_D a^{\boldsymbol{y}}_s\nabla(q^{\boldsymbol{y}}_s - q^{\boldsymbol{y}}_{s,h})\cdot\nabla v_h\,\mathrm{d}x = 0 \quad \forall\, v \in V_h\,,$$

and separating out the $\boldsymbol{m} = \boldsymbol{k}$ term, we get for all $v_h \in V_h$

$$\int_D a^{\boldsymbol{y}}_s\nabla\partial^{\boldsymbol{k}}(q^{\boldsymbol{y}}_s - q^{\boldsymbol{y}}_{s,h})\cdot\nabla v_h\mathrm{d}x = -\sum_{\boldsymbol{m}\leqslant\boldsymbol{k},\boldsymbol{m}\neq\boldsymbol{k}}\binom{\boldsymbol{k}}{\boldsymbol{m}}\int_D(\partial^{\boldsymbol{k}-\boldsymbol{m}}a^{\boldsymbol{y}}_s)\nabla\partial^{\boldsymbol{m}}(q^{\boldsymbol{y}}_s - q^{\boldsymbol{y}}_{s,h})\cdot\nabla v_h\,\mathrm{d}x.$$

Choosing $v_h = P_h\partial^{\boldsymbol{k}}(q^{\boldsymbol{y}}_s - q^{\boldsymbol{y}}_{s,h})$, the left-hand side becomes

$$\int_D a^{\boldsymbol{y}}_s|\nabla P_h\partial^{\boldsymbol{k}}(q^{\boldsymbol{y}}_s - q^{\boldsymbol{y}}_{s,h})|^2\mathrm{d}x + \int_D a^{\boldsymbol{y}}_s\nabla(Id - P_h)\partial^{\boldsymbol{k}}(q^{\boldsymbol{y}}_s - q^{\boldsymbol{y}}_{s,h})\cdot\nabla P_h\partial^{\boldsymbol{k}}(q^{\boldsymbol{y}}_s - q^{\boldsymbol{y}}_{s,h})\mathrm{d}x,$$

where the second term cancels due to the projection definition (7.49). Dividing and multiplying the right-hand side by $a^{\boldsymbol{y}}_s$ and using the Cauchy–Schwarz inequality, one obtains

$$\begin{aligned}\int_D a^{\boldsymbol{y}}_s|\nabla P_h\partial^{\boldsymbol{k}}(q^{\boldsymbol{y}}_s - q^{\boldsymbol{y}}_{s,h})|^2\mathrm{d}x &\leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{k},\boldsymbol{m}\neq\boldsymbol{k}}\binom{\boldsymbol{k}}{\boldsymbol{m}}\left\|\frac{\partial^{\boldsymbol{k}-\boldsymbol{m}}a^{\boldsymbol{y}}_s}{a^{\boldsymbol{y}}_s}\right\|_{L^\infty(D)}\\ &\quad\times\left(\int_D a^{\boldsymbol{y}}_s|\nabla\partial^{\boldsymbol{m}}(q^{\boldsymbol{y}}_s - q^{\boldsymbol{y}}_{s,h})|^2\mathrm{d}x\right)^{1/2}\left(\int_D a^{\boldsymbol{y}}_s|\nabla P_h\partial^{\boldsymbol{k}}(q^{\boldsymbol{y}}_s - q^{\boldsymbol{y}}_{s,h})|^2\mathrm{d}x\right)^{1/2}.\end{aligned}$$

Cancelling the common factor in both sides and using (7.39) we arrive at

$$\|(a_s^{\boldsymbol{y}})^{1/2}\nabla P_h\partial^{\boldsymbol{k}}(q_s^{\boldsymbol{y}}-q_{s,h}^{\boldsymbol{y}})\|_{L^2(D)} \leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{k},\boldsymbol{m}\neq\boldsymbol{k}}\binom{\boldsymbol{k}}{\boldsymbol{m}}\boldsymbol{b}^{\boldsymbol{k}-\boldsymbol{m}}\|(a_s^{\boldsymbol{y}})^{1/2}\nabla\partial^{\boldsymbol{m}}(q_s^{\boldsymbol{y}}-q_{s,h}^{\boldsymbol{y}})\|_{L^2(D)}.$$

Substituting this into (7.50) we obtain

$$\begin{aligned}\underbrace{\|(a_s^{\boldsymbol{y}})^{1/2}\nabla\partial^{\boldsymbol{k}}(q_s^{\boldsymbol{y}}-q_{s,h}^{\boldsymbol{y}})\|_{L^2(D)}}_{\mathbb{A}_{\boldsymbol{k}}} \leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{k},\boldsymbol{m}\neq\boldsymbol{k}}\binom{\boldsymbol{k}}{\boldsymbol{m}}\boldsymbol{b}^{\boldsymbol{k}-\boldsymbol{m}}\underbrace{\|(a_s^{\boldsymbol{y}})^{1/2}\nabla\partial^{\boldsymbol{m}}(q_s^{\boldsymbol{y}}-q_{s,h}^{\boldsymbol{y}})\|_{L^2(D)}}_{\mathbb{A}_{\boldsymbol{m}}}\\ + \underbrace{\|(a_s^{\boldsymbol{y}})^{1/2}\nabla(Id-P_h)\partial^{\boldsymbol{k}}q_s^{\boldsymbol{y}}\|_{L^2(D)}}_{\mathbb{B}_{\boldsymbol{k}}}\end{aligned}$$

leading by Lemma 4.6.1 to

$$\begin{aligned}&\|(a_s^{\boldsymbol{y}})^{1/2}\nabla\partial^{\boldsymbol{k}}(q_s^{\boldsymbol{y}}-q_{s,h}^{\boldsymbol{y}})\|_{L^2(D)}\\ &\leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{k}}\binom{\boldsymbol{k}}{\boldsymbol{m}}\frac{|\boldsymbol{m}|!\,\boldsymbol{b}^{\boldsymbol{m}}}{(\ln 2)^{|\boldsymbol{m}|}}\|(a_s^{\boldsymbol{y}})^{1/2}\nabla(Id-P_h)\partial^{\boldsymbol{k}-\boldsymbol{m}}q_s^{\boldsymbol{y}}\|_{L^2(D)}\\ &\lesssim h\,(a_{\max}^{\boldsymbol{y}})^{1/2}\sum_{\boldsymbol{m}\leqslant\boldsymbol{k}}\binom{\boldsymbol{k}}{\boldsymbol{m}}\frac{|\boldsymbol{m}|!\,\boldsymbol{b}^{\boldsymbol{m}}}{(\ln 2)^{|\boldsymbol{m}|}}\|\Delta(\partial^{\boldsymbol{k}-\boldsymbol{m}}q_s^{\boldsymbol{y}})\|_{L^2(D)}\\ &\leqslant h\,(a_{\max}^{\boldsymbol{y}})^{1/2}\sum_{\boldsymbol{m}\leqslant\boldsymbol{k}}\binom{\boldsymbol{k}}{\boldsymbol{m}}\frac{|\boldsymbol{m}|!\,\boldsymbol{b}^{\boldsymbol{m}}}{(\ln 2)^{|\boldsymbol{m}|}}\widetilde{C}^{\boldsymbol{y}}\,\frac{\boldsymbol{b}^{\boldsymbol{k}-\boldsymbol{m}}}{(\ln 2)^{|\boldsymbol{k}-\boldsymbol{m}|}}\,\frac{(|\boldsymbol{k}-\boldsymbol{m}|+4)!}{(|\boldsymbol{k}-\boldsymbol{m}|+2)(|\boldsymbol{k}-\boldsymbol{m}|+3)}\,\frac{C_q^{\boldsymbol{y}}}{a_{\min}^{\boldsymbol{y}}}C_{zg}\\ &= h\,\frac{\boldsymbol{b}^{\boldsymbol{k}}}{(\ln 2)^{|\boldsymbol{k}|}}\sum_{\boldsymbol{m}\leqslant\boldsymbol{k}}\binom{\boldsymbol{k}}{\boldsymbol{m}}|\boldsymbol{m}|!\frac{(|\boldsymbol{k}-\boldsymbol{m}|+4)!}{(|\boldsymbol{k}-\boldsymbol{m}|+2)(|\boldsymbol{k}-\boldsymbol{m}|+3)}\,\frac{\widetilde{C}^{\boldsymbol{y}}C_q^{\boldsymbol{y}}(a_{\max}^{\boldsymbol{y}})^{1/2}}{a_{\min}^{\boldsymbol{y}}}C_{zg}\\ &= h\,\frac{\boldsymbol{b}^{\boldsymbol{k}}}{(\ln 2)^{|\boldsymbol{k}|}}\,\frac{(|\boldsymbol{k}|+2)!(|\boldsymbol{k}|+6)}{3}\,\frac{\widetilde{C}^{\boldsymbol{y}}C_q^{\boldsymbol{y}}(a_{\max}^{\boldsymbol{y}})^{1/2}}{a_{\min}^{\boldsymbol{y}}}C_{zg}\,.\end{aligned}$$

In order to justify the second inequality, note that by the product rule $q_s^{\boldsymbol{y}}$ satisfies the following PDE

$$-\Delta q_s^{\boldsymbol{y}} = \frac{1}{a_s^{\boldsymbol{y}}}(u_s^{\boldsymbol{y}}-g+\nabla a_s^{\boldsymbol{y}}\cdot\nabla q_s^{\boldsymbol{y}})\,,$$

allowing us to derive $H^2(D)$-regularity

$$\begin{aligned}\|q_s^{\boldsymbol{y}}\|_{H^2(D)} := \|\Delta q_s^{\boldsymbol{y}}\|_{L^2(D)} &\leqslant \frac{1}{a_{\min}^{\boldsymbol{y}}}\left(1+\frac{C_d a_{\max}^{\boldsymbol{y}}}{a_{\min}^{\boldsymbol{y}}}\right)\|u_s^{\boldsymbol{y}}-\widehat{u}\|_{L^2(D)}\\ &\leqslant \frac{1}{a_{\min}^{\boldsymbol{y}}}\left(1+\frac{C_d a_{\max}^{\boldsymbol{y}}}{a_{\min}^{\boldsymbol{y}}}\right)C_q^{\boldsymbol{y}}\big(\|z\|_{L^2(D)}+\|\widehat{u}\|_{L^2(D)}\big)\,.\end{aligned}$$

Classical results from FE theory for $H^2(D)$-regular functions on a convex domain D (see, e.g., [59]) lead, as $h\to 0$, to

$$\inf_{v\in V_{h_\ell}}\|q_s^{\boldsymbol{y}}-v\|_V \lesssim h_\ell\|\Delta q_s^{\boldsymbol{y}}\|_{L^2(D)}\,.$$

This result together with Céa's lemma and the definition of $a_{\max}^{\boldsymbol{y}}$ then proves

$$\|(a_s^{\boldsymbol{y}})^{1/2}\nabla(q_s^{\boldsymbol{y}}-q_{s,h_\ell}^{\boldsymbol{y}})\|_{L^2(D)} \lesssim h_\ell(a_{\max}^{\boldsymbol{y}})^{1/2}\|\Delta q_s^{\boldsymbol{y}}\|_{L^2(D)}\,.$$

as required. □

Note that one can apply a standard Aubin–Nitsche duality argument to obtain quadratic convergence in the meshwidth h measured in the $L^2(D)$-norm.
Let $u_{s_\ell}^{\boldsymbol{y}}$ be the solution of

$$\int_D a_{s_\ell}^{\boldsymbol{y}} \nabla u_{s_\ell}^{\boldsymbol{y}} \cdot \nabla v \,\mathrm{d}x = \int_D zv \,\mathrm{d}x, \quad \forall v \in V \tag{7.51}$$

and $u_{s_{\ell-1}}^{\boldsymbol{y}}$ be the solution of

$$\int_D a_{s_{\ell-1}}^{\boldsymbol{y}} \nabla u_{s_{\ell-1}}^{\boldsymbol{y}} \cdot \nabla v \,\mathrm{d}x = \int_D zv \,\mathrm{d}x, \quad \forall v \in V\,. \tag{7.52}$$

Subtracting (7.52) from (7.51) we get

$$0 = \int_D a_{s_\ell}^{\boldsymbol{y}} (\nabla u_{s_\ell}^{\boldsymbol{y}} - \nabla u_{s_{\ell-1}}^{\boldsymbol{y}}) \cdot \nabla v \,\mathrm{d}x + \int_D (a_{s_\ell}^{\boldsymbol{y}} - a_{s_{\ell-1}}^{\boldsymbol{y}}) \nabla u_{s_{\ell-1}}^{\boldsymbol{y}} \cdot \nabla v \,\mathrm{d}x\,. \tag{7.53}$$

This is used in [31] to show, that

$$\|u_{s_\ell}^{\boldsymbol{y}} - u_{s_{\ell-1}}^{\boldsymbol{y}}\|_V \leqslant \|a_{s_\ell}^{\boldsymbol{y}} - a_{s_{\ell-1}}^{\boldsymbol{y}}\|_{L^\infty(D)} \frac{\|z\|_{V'}}{(a_{\min}^{\boldsymbol{y}})^2}.$$

We are next going to show an analogous result for the $\boldsymbol{\nu}$-th partial derivatives with respect to the uncertain variable.

Lemma 7.2.13. *Let $u_{s_\ell}^{\boldsymbol{y}}$ be the unique solution of (7.51) and $u_{s_{\ell-1}}^{\boldsymbol{y}}$ the unique solution of (7.52). Then, under the assumptions of the previous lemma, it holds that*

$$\|\partial^{\boldsymbol{\nu}}(u_{s_\ell}^{\boldsymbol{y}} - u_{s_{\ell-1}}^{\boldsymbol{y}})\|_V \leqslant \hat{h}_\ell 2\sqrt{d} C_d \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} (|\boldsymbol{\nu}| + 1)! \frac{a_{\max}^{\boldsymbol{y}}}{(a_{\min}^{\boldsymbol{y}})^{3/2}} \|z\|_{V'}\,.$$

Proof. Taking the $\boldsymbol{\nu}$-th partial derivative on both sides of (7.53), we get with Leibniz product rule

$$\begin{aligned}
&\sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \int_D \partial^{\boldsymbol{\nu}-\boldsymbol{m}} a_{s_\ell}^{\boldsymbol{y}} \nabla(\partial^{\boldsymbol{m}}(u_{s_\ell}^{\boldsymbol{y}} - u_{s_{\ell-1}}^{\boldsymbol{y}})) \cdot \nabla v \,\mathrm{d}x \\
&= -\sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \int_D \partial^{\boldsymbol{\nu}-\boldsymbol{m}} (a_{s_\ell}^{\boldsymbol{y}} - a_{s_{\ell-1}}^{\boldsymbol{y}}) \nabla(\partial^{\boldsymbol{m}} u_{s_{\ell-1}}^{\boldsymbol{y}}) \cdot \nabla v \,\mathrm{d}x\,.
\end{aligned}$$

Introducing the notation $w^{\boldsymbol{y}} := u_{s_\ell}^{\boldsymbol{y}} - u_{s_{\ell-1}}^{\boldsymbol{y}}$, separating out the $\boldsymbol{m} = \boldsymbol{\nu}$ term on the left-hand side and setting $v = \partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}}$ gives

$$\begin{aligned}
&\int_D a_{s_\ell}^{\boldsymbol{y}} |\nabla(\partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}})|^2 \mathrm{d}x \\
&\quad = -\sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}, \boldsymbol{m} \neq \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \int_D \partial^{\boldsymbol{\nu}-\boldsymbol{m}} a_{s_\ell}^{\boldsymbol{y}} \nabla(\partial^{\boldsymbol{m}} w^{\boldsymbol{y}}) \cdot \nabla(\partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}}) \,\mathrm{d}x \\
&\qquad - \sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \int_D \partial^{\boldsymbol{\nu}-\boldsymbol{m}} (a_{s_\ell}^{\boldsymbol{y}} - a_{s_{\ell-1}}^{\boldsymbol{y}}) \nabla(\partial^{\boldsymbol{m}} u_{s_{\ell-1}}^{\boldsymbol{y}}) \cdot \nabla(\partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}}) \,\mathrm{d}x \\
&\quad = -\sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}, \boldsymbol{m} \neq \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \int_D \frac{\partial^{\boldsymbol{\nu}-\boldsymbol{m}} a_{s_\ell}^{\boldsymbol{y}}}{a_{s_\ell}^{\boldsymbol{y}}} a_{s_\ell}^{\boldsymbol{y}} \nabla(\partial^{\boldsymbol{m}} w^{\boldsymbol{y}}) \cdot \nabla(\partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}}) \,\mathrm{d}x
\end{aligned}$$

$$
\begin{aligned}
&- \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \int_D \partial^{\boldsymbol{\nu}-\boldsymbol{m}}(a_{s_\ell}^{\boldsymbol{y}} - a_{s_{\ell-1}}^{\boldsymbol{y}})\nabla(\partial^{\boldsymbol{m}} u_{s_{\ell-1}}^{\boldsymbol{y}}) \cdot \nabla(\partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}})\,\mathrm{d}x \\
\leqslant &\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},\boldsymbol{m}\neq\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{m}} \|(a_{s_\ell}^{\boldsymbol{y}})^{1/2}\nabla(\partial^{\boldsymbol{m}} w^{\boldsymbol{y}})\|_{L^2(D)} \|(a_{s_\ell}^{\boldsymbol{y}})^{1/2}\nabla(\partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}})\|_{L^2(D)} \\
&+ \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \|\partial^{\boldsymbol{\nu}-\boldsymbol{m}}(a_{s_\ell}^{\boldsymbol{y}} - a_{s_{\ell-1}}^{\boldsymbol{y}})\|_{L^\infty(D)} \|\nabla(\partial^{\boldsymbol{m}} u_{s_{\ell-1}}^{\boldsymbol{y}})\|_{L^2(D)} \|(a_{s_\ell}^{\boldsymbol{y}})^{1/2}\nabla(\partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}})\|_{L^2(D)}\,.
\end{aligned}
$$

Cancelling one common factor on both sides we obtain

$$
\begin{aligned}
\underbrace{\|(a_{s_\ell}^{\boldsymbol{y}})^{1/2}\nabla(\partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}})\|_{L^2(D)}}_{\mathbb{A}_{\boldsymbol{\nu}}} \leqslant &\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu},\boldsymbol{m}\neq\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{m}} \underbrace{\|(a_{s_\ell}^{\boldsymbol{y}})^{1/2}\nabla(\partial^{\boldsymbol{m}} w^{\boldsymbol{y}})\|_{L^2(D)}}_{\mathbb{A}_{\boldsymbol{m}}} \\
&+ \underbrace{\sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \|\partial^{\boldsymbol{\nu}-\boldsymbol{m}}(a_{s_\ell}^{\boldsymbol{y}} - a_{s_{\ell-1}}^{\boldsymbol{y}})\|_{L^\infty(D)} \|\nabla(\partial^{\boldsymbol{m}} u_{s_{\ell-1}}^{\boldsymbol{y}})\|_{L^2(D)}}_{\mathbb{B}_{\boldsymbol{\nu}}}\,.
\end{aligned}
$$

We know that

$$
\|\nabla(\partial^{\boldsymbol{m}} u_{s_{\ell-1}}^{\boldsymbol{y}})\|_{L^2(D)} = \|\partial^{\boldsymbol{m}} u_{s_{\ell-1}}^{\boldsymbol{y}}\|_V \leqslant |\boldsymbol{m}|! \frac{\boldsymbol{b}^{\boldsymbol{m}}}{(\ln 2)^{|\boldsymbol{m}|}} \frac{\|z\|_{V'}}{a_{\min}^{\boldsymbol{y}}}\,,
$$

and using Lemma 7.2.8 we get

$$
\begin{aligned}
\mathbb{B}_{\boldsymbol{\nu}} &\leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \hat{h}_\ell \sqrt{d} C_d a_{\max}^{\boldsymbol{y}} \boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{m}} |\boldsymbol{m}|! \frac{\boldsymbol{b}^{\boldsymbol{m}}}{(\ln 2)^{|\boldsymbol{m}|}} \frac{\|z\|_{V'}}{a_{\min}^{\boldsymbol{y}}} \\
&\leqslant \boldsymbol{b}^{\boldsymbol{\nu}} \hat{h}_\ell \sqrt{d} C_d a_{\max}^{\boldsymbol{y}} \frac{\|z\|_{V'}}{a_{\min}^{\boldsymbol{y}}} \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \frac{|\boldsymbol{m}|!}{(\ln 2)^{|\boldsymbol{m}|}} \\
&\leqslant \frac{|\boldsymbol{\nu}|!}{(\ln 2)^{|\boldsymbol{\nu}|}} \boldsymbol{b}^{\boldsymbol{\nu}} 2\hat{h}_\ell \sqrt{d} C_d a_{\max}^{\boldsymbol{y}} \frac{\|z\|_{V'}}{a_{\min}^{\boldsymbol{y}}}\,,
\end{aligned}
$$

where we used (4.78). We can now apply Lemma 4.6.1 to get

$$
\begin{aligned}
\|(a_{s_\ell}^{\boldsymbol{y}})^{1/2}\nabla(\partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}})\|_{L^2(D)} &\leqslant \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \frac{|\boldsymbol{m}|!}{(\ln 2)^{|\boldsymbol{m}|}} \boldsymbol{b}^{\boldsymbol{m}} \frac{|\boldsymbol{\nu}-\boldsymbol{m}|!}{(\ln 2)^{|\boldsymbol{\nu}-\boldsymbol{m}|}} \boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{m}} 2\hat{h}_\ell \sqrt{d} C_d a_{\max}^{\boldsymbol{y}} \frac{\|z\|_{V'}}{a_{\min}^{\boldsymbol{y}}} \\
&= 2\hat{h}_\ell \sqrt{d} C_d a_{\max}^{\boldsymbol{y}} \frac{\|z\|_{V'}}{a_{\min}^{\boldsymbol{y}}} \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} \sum_{\boldsymbol{m}\leqslant\boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} |\boldsymbol{m}|!|\boldsymbol{\nu}-\boldsymbol{m}|! \\
&= 2\hat{h}_\ell \sqrt{d} C_d a_{\max}^{\boldsymbol{y}} \frac{\|z\|_{V'}}{a_{\min}^{\boldsymbol{y}}} \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} (|\boldsymbol{\nu}|+1)!\,,
\end{aligned}
$$

as required. □

A similar result holds for the adjoint variable. Therefore let $q_{s_\ell}^{\boldsymbol{y}}$ be the solution of

$$
\int_D a_{s_\ell}^{\boldsymbol{y}} \nabla q_{s_\ell}^{\boldsymbol{y}} \cdot \nabla v\,\mathrm{d}x = \int_D (u_{s_\ell}^{\boldsymbol{y}} - g) v\,\mathrm{d}x, \quad \forall v \in V \tag{7.54}
$$

and $q_{s_{\ell-1}}^{\boldsymbol{y}}$ be the solution of

$$
\int_D a_{s_{\ell-1}}^{\boldsymbol{y}} \nabla q_{s_{\ell-1}}^{\boldsymbol{y}} \cdot \nabla v\,\mathrm{d}x = \int_D (u_{s_{\ell-1}}^{\boldsymbol{y}} - g) v\,\mathrm{d}x, \quad \forall v \in V\,. \tag{7.55}
$$

Subtracting (7.55) from (7.54) we get

$$0 = \int_D a_{s_\ell}^{\boldsymbol{y}}(\nabla q_{s_\ell}^{\boldsymbol{y}} - \nabla q_{s_{\ell-1}}^{\boldsymbol{y}}) \cdot \nabla v \,\mathrm{d}x + \int_D (a_{s_\ell}^{\boldsymbol{y}} - a_{s_{\ell-1}}^{\boldsymbol{y}}) \nabla q_{s_{\ell-1}}^{\boldsymbol{y}} \cdot \nabla v \,\mathrm{d}x - \int_D (u_{s_\ell}^{\boldsymbol{y}} - u_{s_{\ell-1}}^{\boldsymbol{y}}) v \,\mathrm{d}x\,. \tag{7.56}$$

Lemma 7.2.14. *Let $q_{s_\ell}^{\boldsymbol{y}}$ be the unique solution of* (7.54) *and $q_{s_{\ell-1}}^{\boldsymbol{y}}$ the unique solution of* (7.55)*. Then, under the assumptions of the previous lemma, it holds that*

$$\|\partial^{\boldsymbol{\nu}}(q_{s_\ell}^{\boldsymbol{y}} - q_{s_{\ell-1}}^{\boldsymbol{y}})\|_V \leqslant \hat{h}_\ell (|\boldsymbol{\nu}|+2)! \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} 2\sqrt{d} C_d \frac{a_{\max}^{\boldsymbol{y}} C_q^{\boldsymbol{y}}}{(a_{\min}^{\boldsymbol{y}})^{3/2}} C_{zg}\,.$$

Proof. Taking the $\boldsymbol{\nu}$-th partial derivative on both sides of (7.56), we get by Leibniz product rule

$$\begin{aligned}
&\sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \int_D \partial^{\boldsymbol{\nu}-\boldsymbol{m}} a_{s_\ell}^{\boldsymbol{y}} \nabla(\partial^{\boldsymbol{m}}(q_{s_\ell}^{\boldsymbol{y}} - q_{s_{\ell-1}}^{\boldsymbol{y}})) \cdot \nabla v \,\mathrm{d}x \\
&= -\sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \int_D \partial^{\boldsymbol{\nu}-\boldsymbol{m}}(a_{s_\ell}^{\boldsymbol{y}} - a_{s_{\ell-1}}^{\boldsymbol{y}}) \nabla(\partial^{\boldsymbol{m}} q_{s_{\ell-1}}^{\boldsymbol{y}}) \cdot \nabla v \,\mathrm{d}x + \int_D \partial^{\boldsymbol{\nu}}(u_{s_\ell}^{\boldsymbol{y}} - u_{s_{\ell-1}}^{\boldsymbol{y}}) v \,\mathrm{d}x\,.
\end{aligned}$$

Introducing the notation $w^{\boldsymbol{y}} := q_{s_\ell}^{\boldsymbol{y}} - q_{s_{\ell-1}}^{\boldsymbol{y}}$, separating out the $\boldsymbol{m} = \boldsymbol{\nu}$ term on the left-hand side, setting $v = \partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}}$ and cancelling the common factor $\|(a_{s_\ell}^{\boldsymbol{y}})^{-1/2} \partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}}\|_V$, gives

$$\begin{aligned}
\underbrace{\|(a_{s_\ell}^{\boldsymbol{y}})^{1/2} \nabla(\partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}})\|_{L^2(D)}}_{\mathbb{A}_{\boldsymbol{\nu}}} &\leqslant \sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}, \boldsymbol{m} \neq \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{m}} \underbrace{\|(a_{s_\ell}^{\boldsymbol{y}})^{1/2} \nabla(\partial^{\boldsymbol{m}} w^{\boldsymbol{y}})\|_{L^2(D)}}_{\mathbb{A}_{\boldsymbol{m}}} \\
&+ \underbrace{\sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \|\partial^{\boldsymbol{\nu}-\boldsymbol{m}}(a_{s_\ell}^{\boldsymbol{y}} - a_{s_{\ell-1}}^{\boldsymbol{y}})\|_{L^\infty(D)} \|\nabla(\partial^{\boldsymbol{m}} q_{s_{\ell-1}}^{\boldsymbol{y}})\|_{L^2(D)} + \frac{\|\partial^{\boldsymbol{\nu}}(u_{s_\ell}^{\boldsymbol{y}} - u_{s_{\ell-1}}^{\boldsymbol{y}})\|_{V'}}{(a_{\min}^{\boldsymbol{y}})^{1/2}}}_{\mathbb{B}_{\boldsymbol{\nu}}},
\end{aligned}$$

where we used $\int_D \partial^{\boldsymbol{\nu}}(u_{s_\ell}^{\boldsymbol{y}} - u_{s_{\ell-1}}^{\boldsymbol{y}}) \partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}} \,\mathrm{d}x = \int_D (a_{s_\ell}^{\boldsymbol{y}})^{1/2} \partial^{\boldsymbol{\nu}}(u_{s_\ell}^{\boldsymbol{y}} - u_{s_{\ell-1}}^{\boldsymbol{y}}) (a_{s_\ell}^{\boldsymbol{y}})^{-1/2} \partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}} \,\mathrm{d}x \leqslant \|(a_{s_\ell}^{\boldsymbol{y}})^{1/2} \partial^{\boldsymbol{\nu}}(u_{s_\ell}^{\boldsymbol{y}} - u_{s_{\ell-1}}^{\boldsymbol{y}})\|_{V'} \|(a_{s_\ell}^{\boldsymbol{y}})^{-1/2} \partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}}\|_V$ in order to cancel the common factors.
We know from Lemma 7.2.10 that

$$\|\nabla(\partial^{\boldsymbol{m}} q_{s_{\ell-1}}^{\boldsymbol{y}})\|_{L^2(D)} = \|\partial^{\boldsymbol{m}} q_{s_{\ell-1}}^{\boldsymbol{y}}\|_V \leqslant (|\boldsymbol{m}|+1)! \frac{\boldsymbol{b}^{\boldsymbol{m}}}{(\ln 2)^{|\boldsymbol{m}|}} \frac{C_q^{\boldsymbol{y}}}{a_{\min}^{\boldsymbol{y}}} C_{zg}\,.$$

This bound holds because $q_{s_{\ell-1}}^{\boldsymbol{y}}$ is the adjoint state corresponding to the stochastic field $a_{s_{\ell-1}}^{\boldsymbol{y}}$, which in turn is obtained by interpolating the field $a_s^{\boldsymbol{y}}$ in the nodes of a coarser CE method; see Section 7.2.1. Note that in both cases $\boldsymbol{y} \in \mathbb{R}^s$. Importantly, the stochastic field $a_{s_{\ell-1}}^{\boldsymbol{y}}$ thus originates from the CE method of dimension s. Since the $\boldsymbol{b}$ are characterized by the CE method, the $\boldsymbol{b}$ in the bound of Lemma 7.2.10 is the same for $q_{s_\ell}^{\boldsymbol{y}}$ and $q_{s_{\ell-1}}^{\boldsymbol{y}}$. Furthermore, from Lemma 7.2.13 we know that

$$\begin{aligned}
\|\partial^{\boldsymbol{\nu}}(u_{s_\ell}^{\boldsymbol{y}} - u_{s_{\ell-1}}^{\boldsymbol{y}})\|_{V'} &\leqslant c_1 c_2 \|\partial^{\boldsymbol{\nu}}(u_{s_\ell}^{\boldsymbol{y}} - u_{s_{\ell-1}}^{\boldsymbol{y}})\|_V \\
&\leqslant c_1 c_2 2\hat{h}_\ell \sqrt{d} C_d \frac{a_{\max}^{\boldsymbol{y}}}{(a_{\min}^{\boldsymbol{y}})^{3/2}} \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} (|\boldsymbol{\nu}|+1)! \|z\|_{V'}\,.
\end{aligned}$$

This and Lemma 7.2.8 gives

$$\mathbb{B}_{\boldsymbol{\nu}} \leqslant \sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \hat{h}_\ell \sqrt{d} C_d a_{\max}^{\boldsymbol{y}} \boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{m}} (|\boldsymbol{m}|+1)! \frac{\boldsymbol{b}^{\boldsymbol{m}}}{(\ln 2)^{|\boldsymbol{m}|}} \frac{C_q^{\boldsymbol{y}}}{a_{\min}^{\boldsymbol{y}}} C_{zg}$$

$$
\begin{aligned}
&+ \frac{1}{(a^{\boldsymbol{y}}_{\min})^{1/2}} c_1 c_2 2\hat{h}_\ell \sqrt{d} C_d a^{\boldsymbol{y}}_{\max} \frac{\|z\|_{V'}}{(a^{\boldsymbol{y}}_{\min})^{3/2}} \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} (|\boldsymbol{\nu}|+1)! \\
&\leqslant \boldsymbol{b}^{\boldsymbol{\nu}} \hat{h}_\ell \sqrt{d} C_d a^{\boldsymbol{y}}_{\max} \frac{C^{\boldsymbol{y}}_q}{a^{\boldsymbol{y}}_{\min}} (\|z\|_{V'} + \|\widehat{u}\|_{V'}) \sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \frac{(|\boldsymbol{m}|+1)!}{(\ln 2)^{|\boldsymbol{m}|}} \\
&\quad + \frac{1}{(a^{\boldsymbol{y}}_{\min})^{1/2}} c_1 c_2 2\hat{h}_\ell \sqrt{d} C_d a^{\boldsymbol{y}}_{\max} \frac{\|z\|_{V'}}{(a^{\boldsymbol{y}}_{\min})^{3/2}} \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} (|\boldsymbol{\nu}|+1)! \\
&\leqslant \boldsymbol{b}^{\boldsymbol{\nu}} \hat{h}_\ell \sqrt{d} C_d a^{\boldsymbol{y}}_{\max} \frac{C^{\boldsymbol{y}}_q}{a^{\boldsymbol{y}}_{\min}} (\|z\|_{V'} + \|\widehat{u}\|_{V'}) 2 \frac{(|\boldsymbol{\nu}|+1)!}{(\ln 2)^{|\boldsymbol{\nu}|}} \\
&\quad + \frac{1}{(a^{\boldsymbol{y}}_{\min})^{1/2}} c_1 c_2 2\hat{h}_\ell \sqrt{d} C_d a^{\boldsymbol{y}}_{\max} \frac{\|z\|_{V'}}{(a^{\boldsymbol{y}}_{\min})^{3/2}} \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} (|\boldsymbol{\nu}|+1)! \\
&= \hat{h}_\ell (|\boldsymbol{\nu}|+1)! \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} \sqrt{d} C_d \frac{a^{\boldsymbol{y}}_{\max}}{a^{\boldsymbol{y}}_{\min}} \big(2 C^{\boldsymbol{y}}_q C_{zg} + \frac{2 c_1 c_2}{a^{\boldsymbol{y}}_{\min}} \|z\|_{V'}\big) \\
&\leqslant \hat{h}_\ell (|\boldsymbol{\nu}|+1)! \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} 4\sqrt{d} C_d \frac{a^{\boldsymbol{y}}_{\max}}{a^{\boldsymbol{y}}_{\min}} C^{\boldsymbol{y}}_q C_{zg} \,,
\end{aligned}
$$

where we used (7.36) in the second inequality and (4.80) in the third inequality. We can now apply Lemma 4.6.1 to get

$$
\begin{aligned}
\|(a^{\boldsymbol{y}}_{s_\ell})^{1/2} \nabla(\partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}})\|_{L^2(D)} &\leqslant \sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} \frac{|\boldsymbol{m}|!}{(\ln 2)^{|\boldsymbol{m}|}} \boldsymbol{b}^{\boldsymbol{m}} \frac{|(\boldsymbol{\nu}-\boldsymbol{m}|+1)!}{(\ln 2)^{|\boldsymbol{\nu}-\boldsymbol{m}|}} \boldsymbol{b}^{\boldsymbol{\nu}-\boldsymbol{m}} \\
&\quad \times \hat{h}_\ell 4\sqrt{d} C_d \frac{a^{\boldsymbol{y}}_{\max}}{a^{\boldsymbol{y}}_{\min}} C^{\boldsymbol{y}}_q (\|z\|_{V'} + \|\widehat{u}\|_{V'}) \\
&= \hat{h}_\ell \frac{(|\boldsymbol{\nu}|+2)!}{2} \frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}} 4\sqrt{d} C_d \frac{a^{\boldsymbol{y}}_{\max}}{a^{\boldsymbol{y}}_{\min}} C^{\boldsymbol{y}}_q (\|z\|_{V'} + \|\widehat{u}\|_{V'}) \,,
\end{aligned}
$$

where we used the equality $\sum_{\boldsymbol{m} \leqslant \boldsymbol{\nu}} \binom{\boldsymbol{\nu}}{\boldsymbol{m}} |\boldsymbol{m}|! (|\boldsymbol{\nu}-\boldsymbol{m}|+1)! = \frac{(|\boldsymbol{\nu}|+2)!}{2}$, which is stated, e.g., in [113, equation 9.5]. From

$$
(a^{\boldsymbol{y}}_{\min})^{1/2} \|\partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}}\|_V \leqslant \|(a^{\boldsymbol{y}}_{s_\ell})^{1/2} \nabla(\partial^{\boldsymbol{\nu}} w^{\boldsymbol{y}})\|_{L^2(D)},
$$

the claim follows directly. □

Integration error on difference of two levels

In this section we analyze the expected (w.r.t. the random shifts) MSE for approximating the difference of two consecutive levels in the MLQMC estimator. To this end, we introduce the weighted Sobolev space $\mathcal{W}_{s,\boldsymbol{\gamma}}$, with norm given by

$$
\|F\|^2_{\mathcal{W}_{s,\boldsymbol{\gamma}}} := \sum_{\mathfrak{u} \subseteq \{1:s\}} \frac{1}{\gamma_{\mathfrak{u}}} \int_{\mathbb{R}^{|\mathfrak{u}|}} \bigg(\int_{\mathbb{R}^{s-|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|} F}{\partial \boldsymbol{y}_{\mathfrak{u}}} (\boldsymbol{y}_{\mathfrak{u}}, \boldsymbol{y}_{\{1:s\}\setminus\mathfrak{u}}) \prod_{j \in \{1:s\}\setminus\mathfrak{u}} \phi(y_j) \, \mathrm{d}\boldsymbol{y}_{\{1:s\}\setminus\mathfrak{u}} \bigg)^2 \prod_{j \in \mathfrak{u}} \psi_j^2(y_j) \, \mathrm{d}\boldsymbol{y}_{\mathfrak{u}} \,.
$$

Here $\{1:s\}$ is a shorthand notation for the set of indices $\{1, 2, \ldots, s\}$. In the sum, $\boldsymbol{y}_{\mathfrak{u}} = (y_j)_{j \in \mathfrak{u}}$ denotes the active variables, while $\boldsymbol{y}_{\{1:s\}\setminus\mathfrak{u}} = (y_j)_{j \notin \mathfrak{u}}$ denotes the inactive variables. The constants $\gamma_{\mathfrak{u}}$ are weights, collected formally in $\boldsymbol{\gamma}$, and the functions $\psi_j : \mathbb{R} \to \mathbb{R}^+$ determine the behavior of the functions in the space. For the analysis, based on [70, 115, 123] to hold, we consider functions $\psi_j^2(y) = \exp(-9\alpha_j |y|)$ with $\alpha_j > 0$ to be specified below.

In the multilevel estimator for our gradient we want to apply the QMC rule to the difference $q_{s_\ell}^{\boldsymbol{y}} - q_{s_{\ell-1}}^{\boldsymbol{y}}$. On a level $\ell \in \{1, \dots, L\}$ we can use Fubini's theorem and [70, Theorem 15] to get

$$\begin{aligned}
\mathcal{V}_\ell &= \int_D \mathbb{V}_\Delta[\mathcal{Q}_{n_\ell,R_\ell}(q_{s_\ell}^{\boldsymbol{y}} - q_{s_{\ell-1}}^{\boldsymbol{y}})]\mathrm{d}x = \frac{1}{R_\ell}\int_D \mathbb{V}_\Delta[\mathcal{Q}_{n_\ell}(q_{s_\ell}^{\boldsymbol{y}} - q_{s_{\ell-1}}^{\boldsymbol{y}})]\mathrm{d}x \\
&= \mathbb{E}_\Delta\big[\|\mathcal{Q}_n(q_{s_\ell}^{\boldsymbol{y}} - q_{s_{\ell-1}}^{\boldsymbol{y}}) - \mathbb{E}[F]\|_{L^2(D)}^2\big] = \int_D \mathbb{E}_\Delta[(\mathcal{Q}_n(q_{s_\ell}^{\boldsymbol{y}} - q_{s_{\ell-1}}^{\boldsymbol{y}}) - \mathbb{E}[q_{s_\ell}^{\boldsymbol{y}} - q_{s_{\ell-1}}^{\boldsymbol{y}}])^2]\mathrm{d}x \\
&\leqslant \frac{1}{R_\ell}\bigg(\sum_{\varnothing\neq\mathfrak{u}\subseteq\{1:s_\ell\}} \gamma_\mathfrak{u}^\lambda \prod_{j\in\mathfrak{u}} \varrho_j(\lambda)\bigg)^{1/\lambda} (\varphi_{\mathrm{tot}}(N))^{-1/\lambda} \int_D \|q_{s_\ell}^{\boldsymbol{y}} - q_{s_{\ell-1}}^{\boldsymbol{y}}\|_{\mathcal{W}_{s_\ell,\boldsymbol{\gamma}}}^2 \mathrm{d}x \qquad (7.57)
\end{aligned}$$

where

$$\varrho_j(\lambda) := 2\bigg(\frac{\sqrt{2\pi}\exp(\alpha_j^2/\eta^*)}{\pi^{2-2\eta^*}(1-\eta^*)\eta^*}\bigg)^\lambda \zeta\left(\lambda + \frac{1}{2}\right).$$

Here $\eta^* = (2\lambda-1)/(4\lambda)$, $\zeta(\boldsymbol{x})$ denotes the Riemann Zeta function and $\varphi_{\mathrm{tot}}(N) := |\{1 \leqslant z \leqslant N \mid \gcd(z,N) = 1\}|$ denotes the Euler totient function. In particular, if N is a power of a prime, it can be shown that $1/\varphi_{\mathrm{tot}}(N) \leqslant 2/N$. By using the shorthand notation $F(\boldsymbol{y}) := q_{s_\ell}^{\boldsymbol{y}} - q_{s_{\ell-1}}^{\boldsymbol{y}}$, we observe that

$$\begin{aligned}
&\int_D \|F\|_{\mathcal{W}_{s,\boldsymbol{\gamma}}}^2 \mathrm{d}x \\
&= \int_D \sum_{\mathfrak{u}\subseteq\{1:s\}} \frac{1}{\gamma_\mathfrak{u}} \int_{\mathbb{R}^{|\mathfrak{u}|}} \bigg(\int_{\mathbb{R}^{s-|\mathfrak{u}|}} \frac{\partial^{|\mathfrak{u}|}F}{\partial \boldsymbol{y}_\mathfrak{u}} \prod_{j\in\{1:s\}\setminus\mathfrak{u}} \phi(y_j)\mathrm{d}\boldsymbol{y}_{\{1:s\}\setminus\mathfrak{u}}\bigg)^2 \prod_{j\in\mathfrak{u}} \psi_j^2(y_j)\mathrm{d}\boldsymbol{y}_\mathfrak{u}\mathrm{d}x \\
&\leqslant \int_D \sum_{\mathfrak{u}\subseteq\{1:s\}} \frac{1}{\gamma_\mathfrak{u}} \int_{\mathbb{R}^{|\mathfrak{u}|}} \int_{\mathbb{R}^{s-|\mathfrak{u}|}} \bigg(\frac{\partial^{|\mathfrak{u}|}F}{\partial \boldsymbol{y}_\mathfrak{u}}\bigg)^2 \prod_{j\in\{1:s\}\setminus\mathfrak{u}} \phi(y_j)\mathrm{d}\boldsymbol{y}_{\{1:s\}\setminus\mathfrak{u}} \prod_{j\in\mathfrak{u}} \psi_j^2(y_j)\mathrm{d}\boldsymbol{y}_\mathfrak{u}\mathrm{d}x \\
&= \sum_{\mathfrak{u}\subseteq\{1:s\}} \frac{1}{\gamma_\mathfrak{u}} \int_{\mathbb{R}^{|\mathfrak{u}|}} \int_{\mathbb{R}^{s-|\mathfrak{u}|}} \bigg\|\frac{\partial^{|\mathfrak{u}|}F}{\partial \boldsymbol{y}_\mathfrak{u}}\bigg\|_{L^2(D)}^2 \prod_{j\in\{1:s\}\setminus\mathfrak{u}} \phi(y_j)\mathrm{d}\boldsymbol{y}_{\{1:s\}\setminus\mathfrak{u}} \prod_{j\in\mathfrak{u}} \psi_j^2(y_j)\mathrm{d}\boldsymbol{y}_\mathfrak{u} \\
&\leqslant c_2^2 \sum_{\mathfrak{u}\subseteq\{1:s\}} \frac{1}{\gamma_\mathfrak{u}} \int_{\mathbb{R}^{|\mathfrak{u}|}} \int_{\mathbb{R}^{s-|\mathfrak{u}|}} \bigg\|\frac{\partial^{|\mathfrak{u}|}F}{\partial \boldsymbol{y}_\mathfrak{u}}\bigg\|_{V}^2 \prod_{j\in\{1:s\}\setminus\mathfrak{u}} \phi(y_j)\mathrm{d}\boldsymbol{y}_{\{1:s\}\setminus\mathfrak{u}} \prod_{j\in\mathfrak{u}} \psi_j^2(y_j)\mathrm{d}\boldsymbol{y}_\mathfrak{u}\,. \qquad (7.58)
\end{aligned}$$

Thus we take $F = q_{s_\ell}^{\boldsymbol{y}} - q_{s_{\ell-1}}^{\boldsymbol{y}}$ and plug (7.58) into (7.57) to obtain the following result.

Theorem 7.2.15. *Let $\psi_j^2(y) := \exp(-9\alpha_j|y|)$ for $\max(b_j, \alpha_{\min}) < \alpha_j < \alpha_{\max}$ for all $j \in \mathfrak{u} \subseteq \{1:s\}$ and some $0 < \alpha_{\min} < \alpha_{\max} < \infty$. Given $s_\ell, n_\ell \in \mathbb{N}$, and weights $\boldsymbol{\gamma}$, a generating vector $\boldsymbol{z} \in \mathbb{N}^s$ for a randomly shifted lattice rule can be constructed using a component-by-component algorithm such that the variance $\mathcal{V}_\ell$, defined in (7.31), for approximating the difference of two consecutive levels in the MLQMC estimator satisfies, for all $\lambda \in (1/2, 1]$,*

$$\mathcal{V}_\ell \leqslant \frac{1}{R_\ell}\varphi_{tot}(n_\ell)^{-1/\lambda} C_{s_\ell,\boldsymbol{\gamma}} \bigg(h_{\ell-1}c_2 C_{zg} C_P \exp(\|\bar{Z}\|_\infty)\bigg)^2 \exp\left(\frac{81}{2}\|\boldsymbol{b}\|_2^2 + 2\frac{9}{\sqrt{2\pi}}\|\boldsymbol{b}\|_1\right),$$

with C_P some constant depending only on c_1, c_2 and C_d and where

$$C_{s,\boldsymbol{\gamma}} := \bigg(\sum_{\varnothing\neq\mathfrak{u}\subseteq\{1:s_\ell\}} \gamma_\mathfrak{u}^\lambda \prod_{j\in\mathfrak{u}} \varrho_j(\lambda)\bigg)^{1/\lambda} \sum_{\mathfrak{u}\subseteq\{1:s_\ell\}} \frac{1}{\gamma_\mathfrak{u}} \bigg(\frac{(|\mathfrak{u}|+2)!(|\mathfrak{u}|+6)}{3(\ln 2)^{|\mathfrak{u}|}}\bigg)^2 \bigg(\prod_{j\in\mathfrak{u}} \frac{\tilde{b}_j^2}{\alpha_j - b_j}\bigg)$$

and

$$\varrho_j(\lambda) := 2\left(\frac{\sqrt{2\pi}\exp(\alpha_j^2/\eta^*)}{\pi^{2-2\eta^*}(1-\eta^*)\eta^*}\right)^{\lambda}\zeta\left(\lambda+\frac{1}{2}\right). \tag{7.59}$$

Proof. For this proof it is important to recall from Section 7.2.1 that $q^{\boldsymbol{y}}_{s_{\ell-1}}$ is the adjoint state corresponding to the stochastic field $a^{\boldsymbol{y}}_{s_{\ell-1}}$, which in turn is obtained by interpolating the field $a^{\boldsymbol{y}}_{s_\ell}$ in the nodes of a coarser CE method. In both cases $\boldsymbol{y}\in\mathbb{R}^{s_\ell}$. By the triangle inequality we have

$$\begin{aligned}&\|\partial^{\boldsymbol{\nu}}(q^{\boldsymbol{y}}_{s_\ell,h_\ell}-q^{\boldsymbol{y}}_{s_{\ell-1},h_{\ell-1}})\|_V \\ &\quad\leqslant \underbrace{\|\partial^{\boldsymbol{\nu}}(q^{\boldsymbol{y}}_{s_\ell,h_\ell}-q^{\boldsymbol{y}}_{s_\ell})\|_V}_{\text{term}_1}+\underbrace{\|\partial^{\boldsymbol{\nu}}(q^{\boldsymbol{y}}_{s_\ell}-q^{\boldsymbol{y}}_{s_{\ell-1}})\|_V}_{\text{term}_2}+\underbrace{\|\partial^{\boldsymbol{\nu}}(q^{\boldsymbol{y}}_{s_{\ell-1}}-q^{\boldsymbol{y}}_{s_{\ell-1},h_{\ell-1}})\|_V}_{\text{term}_3},\end{aligned} \tag{7.60}$$

which in turn can be estimated using Lemma 7.2.12 (term_1 and term_3) and Lemma 7.2.14 (term_2):

$$\begin{aligned}\text{term}_1 &\lesssim h_\ell\,\frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}}\,\frac{(|\boldsymbol{\nu}|+2)!(|\boldsymbol{\nu}|+6)}{3}\,\frac{(a^{\boldsymbol{y}}_{\max})^{1/2}\widetilde{C}^{\boldsymbol{y}}C^{\boldsymbol{y}}_q}{a^{\boldsymbol{y}}_{\min}}(\|z\|_{V'}+\|\widehat{u}\|_{V'})\\
\text{term}_2 &\leqslant \hat{h}_\ell\,\frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}}(|\boldsymbol{\nu}|+2)!\,\frac{a^{\boldsymbol{y}}_{\max}C^{\boldsymbol{y}}_q}{(a^{\boldsymbol{y}}_{\min})^{3/2}}2\sqrt{d}C_d(\|z\|_{V'}+\|\widehat{u}\|_{V'})\\
\text{term}_3 &\lesssim h_{\ell-1}\,\frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}}\,\frac{(|\boldsymbol{\nu}|+2)!(|\boldsymbol{\nu}|+6)}{3}\,\frac{(a^{\boldsymbol{y}}_{\max})^{1/2}\widetilde{C}^{\boldsymbol{y}}C^{\boldsymbol{y}}_q}{a^{\boldsymbol{y}}_{\min}}(\|z\|_{V'}+\|\widehat{u}\|_{V'}).\end{aligned}$$

For $\ell\leqslant L_{\max}$ we can find a constant such that $\hat{h}_\ell\lesssim h_\ell$. Due to Assumption 7.2.6 we have that $\text{term}_2=0$ for $\ell>L_{\max}$. The precise form of the bound of term_2 is then not important. However, if the constant C_d in Lemma 7.2.8 can be found independent of ℓ (see Remark 7.2.9), then Assumption 7.2.6 can be omitted. In that case, the form of term_2 and therefore Lemma 7.2.14 are relevant for the remaining analysis. Recalling that $\widetilde{C}^{\boldsymbol{y}}=\max(1,C^{\boldsymbol{y}})=\max(1,2\frac{C_d a^{\boldsymbol{y}}_{\max}}{a^{\boldsymbol{y}}_{\min}})$ and $C^{\boldsymbol{y}}_q=\max(1,\frac{c_1c_2}{a^{\boldsymbol{y}}_{\min}})$, we can further estimate

$$\begin{aligned}\text{term}_1+\text{term}_2+\text{term}_3 &\lesssim h_{\ell-1}\frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}}\frac{(|\boldsymbol{\nu}|+2)!(|\boldsymbol{\nu}|+6)}{3}(\|z\|_{V'}+\|\widehat{u}\|_{V'})\\
&\times\left(2\frac{(a^{\boldsymbol{y}}_{\max})^{1/2}}{a^{\boldsymbol{y}}_{\min}}\Big(1+2C_d\frac{a^{\boldsymbol{y}}_{\max}}{a^{\boldsymbol{y}}_{\min}}\Big)\Big(1+\frac{c_1c_2}{a^{\boldsymbol{y}}_{\min}}\Big)+\Big(\frac{a^{\boldsymbol{y}}_{\max}}{(a^{\boldsymbol{y}}_{\min})^{3/2}}\Big)\Big(1+\frac{c_1c_2}{a^{\boldsymbol{y}}_{\min}}\Big)\right),\end{aligned} \tag{7.61}$$

so the bound depends on $\boldsymbol{y}$ only through $a^{\boldsymbol{y}}_{\min}$ and $a^{\boldsymbol{y}}_{\max}$. We use

$$(a^{\boldsymbol{y}}_{\min})^{-1}, a^{\boldsymbol{y}}_{\max}\leqslant\exp(\|\bar{Z}\|_\infty)\exp(\boldsymbol{b}^\top|\boldsymbol{y}|)$$

to derive the bounds

$$\begin{aligned}\frac{(a^{\boldsymbol{y}}_{\max})^{1/2}}{a^{\boldsymbol{y}}_{\min}} &\leqslant \big(\exp(\|\bar{Z}\|_\infty)\exp(\boldsymbol{b}^\top|\boldsymbol{y}|)\big)^{3/2},\\
\frac{a^{\boldsymbol{y}}_{\max}}{(a^{\boldsymbol{y}}_{\min})^{3/2}} &\leqslant \big(\exp(\|\bar{Z}\|_\infty)\exp(\boldsymbol{b}^\top|\boldsymbol{y}|)\big)^{5/2},\\
\frac{c_1c_2}{a^{\boldsymbol{y}}_{\min}} &\leqslant c_1c_2\exp(\|\bar{Z}\|_\infty)\exp(\boldsymbol{b}^\top|\boldsymbol{y}|),\end{aligned}$$

$$2C_d\frac{a^{\boldsymbol{y}}_{\max}}{a^{\boldsymbol{y}}_{\min}} \leqslant 2C_d\big(\exp(\|\bar{Z}\|_\infty)\,\exp(\boldsymbol{b}^\top|\boldsymbol{y}|)\big)^2.$$

Moreover, we have $1 \leqslant \exp(\|\bar{Z}\|_\infty)\,\exp(\boldsymbol{b}^\top|\boldsymbol{y}|) \leqslant \big(\exp(\|\bar{Z}\|_\infty)\,\exp(\boldsymbol{b}^\top|\boldsymbol{y}|)\big)^2$. Using these estimates we conclude that

$$(7.61) \leqslant h_{\ell-1}\frac{\boldsymbol{b}^{\boldsymbol{\nu}}}{(\ln 2)^{|\boldsymbol{\nu}|}}\frac{(|\boldsymbol{\nu}|+2)!(|\boldsymbol{\nu}|+6)}{3}C_{zg}C_P\Big(\exp(\|\bar{Z}\|_\infty)\,\exp(\boldsymbol{b}^\top|\boldsymbol{y}|)\Big)^{9/2}$$

with C_P some constant which depends only on c_1, c_2 and C_d.
Replacing $\partial^{\boldsymbol{\nu}}$ by $\frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{y}_\mathfrak{u}}$ with $\mathfrak{u} \subseteq \{1:s_\ell\}$ in (7.60), i.e., restricting to the case where all $\nu_j \leqslant 1$ as is the case in the definition of the $\mathcal{W}_{s,\boldsymbol{\gamma}}$-norm, we obtain

$$\begin{aligned}&\Big\|\frac{\partial^{|\mathfrak{u}|}}{\partial \boldsymbol{y}_\mathfrak{u}}(q^{\boldsymbol{y}}_{s_\ell,h_\ell} - q^{\boldsymbol{y}}_{s_{\ell-1},h_{\ell-1}})\Big\|_V\\ &\qquad\leqslant h_{\ell-1}\Big(\prod_{j\in\mathfrak{u}} b_j\Big)\frac{(|\mathfrak{u}|+2)!(|\mathfrak{u}|+6)}{3(\ln 2)^{|\mathfrak{u}|}}C_{zg}C_P\Big(\exp(\|\bar{Z}\|_\infty)\exp(\boldsymbol{b}^\top|\boldsymbol{y}|)\Big)^{9/2}.\end{aligned}$$

Moreover, the product form of this bound allows us to group the factors in (7.58), with F taken to be $q^{\boldsymbol{y}}_{s_\ell,h_\ell} - q^{\boldsymbol{y}}_{s_{\ell-1},h_{\ell-1}}$, for $j \in \mathfrak{u}$ and $j \in \{1:s_\ell\}\backslash\mathfrak{u}$ separately, i.e.,

$$\exp\big(\tfrac{9}{2}\boldsymbol{b}^\top|\boldsymbol{y}|\big) = \prod_{j\in\mathfrak{u}}\exp\big(\tfrac{9}{2}b_j|y_j|\big)\prod_{j\in\{1:s\}\backslash\mathfrak{u}}\exp\big(\tfrac{9}{2}b_j|y_j|\big). \tag{7.62}$$

We first estimate the factors $j \in \{1:s_\ell\}\backslash\mathfrak{u}$

$$\begin{aligned}&\int_{\mathbb{R}^{s_\ell-|\mathfrak{u}|}}\Big(\prod_{j\in\{1:s_\ell\}\backslash\mathfrak{u}}\exp\big(\tfrac{9}{2}b_j|y_j|\big)\Big)^2\prod_{j\in\{1:s_\ell\}\backslash\mathfrak{u}}\phi(y_j)\,\mathrm{d}\boldsymbol{y}_{\{1:s_\ell\}\backslash\mathfrak{u}}\\
&\qquad=\int_{\mathbb{R}^{s_\ell-|\mathfrak{u}|}}\prod_{j\in\{1:s_\ell\}\backslash\mathfrak{u}}\exp\big(9b_j|y_j|\big)\prod_{j\in\{1:s_\ell\}\backslash\mathfrak{u}}\phi(y_j)\,\mathrm{d}\boldsymbol{y}_{\{1:s_\ell\}\backslash\mathfrak{u}}\\
&\qquad=\int_{\mathbb{R}^{s_\ell-|\mathfrak{u}|}}\prod_{j\in\{1:s_\ell\}\backslash\mathfrak{u}}\exp\big(9b_j|y_j|\big)\frac{1}{\sqrt{2\pi}}\exp\Big(\frac{-y_j^2}{2}\Big)\,\mathrm{d}\boldsymbol{y}_{\{1:s_\ell\}\backslash\mathfrak{u}}\\
&\qquad=\prod_{j\in\{1:s_\ell\}\backslash\mathfrak{u}}2\int_0^\infty\exp\big(9b_j|y_j|\big)\frac{1}{\sqrt{2\pi}}\exp\Big(\frac{-y_j^2}{2}\Big)\,\mathrm{d}y\\
&\qquad=\prod_{j\in\{1:s_\ell\}\backslash\mathfrak{u}}2\int_0^\infty\exp\Big(\frac{81}{2}b_j^2\Big)\frac{1}{\sqrt{2\pi}}\exp\Big(\frac{-(9b_j-y_j)^2}{2}\Big)\,\mathrm{d}y\\
&\qquad=\prod_{j\in\{1:s_\ell\}\backslash\mathfrak{u}}\exp\Big(\frac{81}{2}b_j^2\Big)2\int_0^\infty\frac{1}{\sqrt{2\pi}}\exp\Big(\frac{-(9b_j-y_j)^2}{2}\Big)\,\mathrm{d}y\\
&\qquad=\prod_{j\in\{1:s_\ell\}\backslash\mathfrak{u}}\exp\Big(\frac{81}{2}b_j^2\Big)2\Phi(9b_j)\end{aligned}$$

where Φ denotes the univariate cumulative standard normal distribution function.
Secondly, we estimate the factors $j \in \mathfrak{u}$

$$\int_{\mathbb{R}^{|\mathfrak{u}|}}\prod_{j\in\mathfrak{u}}\exp(9b_j|y_j|)\psi_j^2(y_j)\mathrm{d}\boldsymbol{y}_\mathfrak{u} = \prod_{j\in\mathfrak{u}}\Big(\int_{-\infty}^\infty\exp(9b_j|y|)\psi_j^2(y)\mathrm{d}y\Big).$$

With $\psi_j^2(y) := \exp(-9\alpha_j|y|)$ for $\max(b_j, \alpha_{\min}) < \alpha_j < \alpha_{\max}$ for all $j \in \mathfrak{u}$ and some $0 < \alpha_{\min} < \alpha_{\max} < \infty$, we get

$$\int_{\mathbb{R}^{|\mathfrak{u}|}} \prod_{j\in\mathfrak{u}} \exp(9b_j|y_j|)\psi_j^2(y_j)\mathrm{d}\boldsymbol{y}_{\mathfrak{u}} = \prod_{j\in\mathfrak{u}} \frac{1}{\alpha_j - b_j}\,.$$

Defining

$$\tilde{b}_j := \frac{b_j}{2\exp(\frac{81}{2}b_j^2)\Phi(9b_j)}$$

we arrive at

$$\begin{aligned}
&\int_{\mathbb{R}^{|\mathfrak{u}|}} \Bigg(\int_{\mathbb{R}^{s_\ell-|\mathfrak{u}|}} \left(\exp\left(\frac{9}{2}\boldsymbol{b}^\top|\boldsymbol{y}|\right)\right)^2 \prod_{j\in\mathfrak{u}} b_j \prod_{j\in\{1:s\}\setminus\mathfrak{u}} \phi(y_j)\,\mathrm{d}\boldsymbol{y}_{\{1:s_\ell\}\setminus\mathfrak{u}}\Bigg) \prod_{j\in\mathfrak{u}} \psi_j^2(y_j)\,\mathrm{d}\boldsymbol{y}_{\mathfrak{u}} \\
&\qquad = \Bigg(\prod_{j\in\{1:s_\ell\}\setminus\mathfrak{u}} 2\exp\left(\frac{81}{2}b_j^2\right)\Phi(9b_j)\Bigg)\Bigg(\prod_{j\in\mathfrak{u}} \frac{b_j^2}{\alpha_j - b_j}\Bigg) \\
&\qquad = \Bigg(\prod_{j\in\{1:s_\ell\}} 2\exp\left(\frac{81}{2}b_j^2\right)\Phi(9b_j)\Bigg)\Bigg(\prod_{j\in\mathfrak{u}} \frac{\tilde{b}_j^2}{\alpha_j - b_j}\Bigg)\,.
\end{aligned}$$

Using $2\Phi(9b_j) = 1 + \mathrm{erf}\big(\frac{9b_j}{\sqrt{2}}\big) \leqslant 1 + 2\frac{9}{\sqrt{2\pi}}b_j \leqslant \exp(2\frac{9}{\sqrt{2\pi}}b_j)$ for all j, where erf denotes the Gauss error function, we have

$$\begin{aligned}
\prod_{j\in\{1:s_\ell\}} 2\exp\left(\frac{81}{2}b_j^2\right)\Phi(9b_j) &\leqslant \prod_{j\in\{1:s\}} \exp\left(\frac{81}{2}b_j^2\right)\exp\left(2\frac{9}{\sqrt{2\pi}}b_j\right) \\
&< \exp\left(\frac{81}{2}\sum_{j\in\{1:s_\ell\}} b_j^2 + 2\frac{9}{\sqrt{\pi}}\sum_{j\in\{1:s\}} b_j\right) \\
&= \exp\left(\frac{81}{2}\|\boldsymbol{b}\|_2^2 + 2\frac{9}{\sqrt{2\pi}}\|\boldsymbol{b}\|_1\right).
\end{aligned}$$

We have thus proved the following

$$\begin{aligned}
&\int_D \|q_{s_\ell,h_\ell}^{\boldsymbol{y}} - q_{s_{\ell-1},h_{\ell-1}}^{\boldsymbol{y}}\|_{\mathcal{W}_{s_\ell,\boldsymbol{\gamma}}}^2 \mathrm{d}x \\
&\quad \leqslant \left(h_{\ell-1}c_2(\|z\|_{V'} + \|\widehat{u}\|_{V'})C_P\exp(\|\bar{Z}\|_\infty)\right)^2 \\
&\qquad \times \sum_{\mathfrak{u}\subseteq\{1:s_\ell\}} \frac{1}{\gamma_{\mathfrak{u}}}\left(\frac{(|\mathfrak{u}|+2)!(|\mathfrak{u}|+6)}{3(\ln 2)^{|\mathfrak{u}|}}\right)^2\left(\prod_{j\in\mathfrak{u}}\frac{\tilde{b}_j^2}{\alpha_j - b_j}\right)\exp\left(\frac{81}{2}\|\boldsymbol{b}\|_2^2 + 2\frac{9}{\sqrt{2\pi}}\|\boldsymbol{b}\|_1\right),
\end{aligned}$$

as required. □

Without a careful choice of the weight parameters $\gamma_{\mathfrak{u}}$, the quantity $C_{s_\ell,\boldsymbol{\gamma}}$ might grow with increasing s_ℓ. To ensure that $C_{s_\ell,\boldsymbol{\gamma}}$ is bounded independently of s_ℓ, we choose the weight parameters accordingly. This requires an assumption on the boundedness of $\|\boldsymbol{b}\|_p$, which is also made in [72, Section 3.4], where it is discussed in detail.

Lemma 7.2.16. *Let N be a power of a prime number and let the assumptions of the preceding Theorem hold. Moreover, let $\lambda \in (\frac{1}{2}, 1]$ and assume that $\|\boldsymbol{b}\|_p$ is uniformly bounded with respect to s_ℓ for $p = 2\lambda/(1+\lambda)$. Then, for a particular choice of the weights $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$, there is a constant $C(\lambda) > 0$ such that*

$$\mathcal{V}_\ell \leqslant \frac{1}{R_\ell} h_\ell^2 C(\lambda) N^{-\frac{1}{\lambda}} .$$

Proof. Since N is a prime power, we have that $1/\varphi_{\mathrm{tot}}(N) \leqslant 2/N$. Due to the preceding Theorem it is sufficient to find an upper bound on $C_{s_\ell,\boldsymbol{\gamma}}$ that is independent of s_ℓ. To this end we choose the weights $\boldsymbol{\gamma}$ to minimize $C_{s_\ell,\boldsymbol{\gamma}}$. By [70, lemma 18] the "product and order dependent" (POD) minimizer γ^* of $C_{s_\ell,\boldsymbol{\gamma}}$ is given by

$$\gamma^* = \left(\left(\frac{(|\mathfrak{u}|+2)!(|\mathfrak{u}|+6)}{3(\ln 2)^{|\mathfrak{u}|}} \right)^2 \prod_{j\in\mathfrak{u}} \frac{\tilde{b}_j^2}{(\alpha_j - b_j)\varrho_j(\lambda)} \right)^{\frac{1}{1+\lambda}} .$$

One can show that

$$C_{s_\ell,\gamma^*} = S_\lambda^{1+\frac{1}{\lambda}}, \qquad \text{where} \qquad S_\lambda = \sum_{\mathfrak{u}\subseteq\{1:s_\ell\}} \left[\left(\frac{(|\mathfrak{u}|+2)!(|\mathfrak{u}|+6)}{3(\ln 2)^{|\mathfrak{u}|}} \right)^2 \prod_{j\in\mathfrak{u}} \frac{\tilde{b}_j^2 \varrho_j(\lambda)^{\frac{1}{\lambda}}}{\alpha_j - b_j} \right]^{\frac{\lambda}{1+\lambda}},$$

hence, it is sufficient to show that $S_\lambda < \infty$. To this end we choose the parameters α_j that minimize S_λ. We observe that all terms of S_λ are positive, thus minimizing S_λ, or equivalently C_{s_ℓ,γ^*}, with respect to the parameters $\{a_j\}_{j\geqslant 1}$ is equivalent to minimizing each of the functions $\frac{\varrho_j(\lambda)^{\frac{1}{\lambda}}}{\alpha_j - b_j}$ with respect to α_j. Due to (7.59), $\varrho_j(\lambda)^{\frac{1}{\lambda}} = c\exp(\alpha_j^2/\eta^*)$, for some constant c independent of α_j and for $\eta^* = (2\lambda-1)/(4\lambda)$, leads to

$$\alpha_j = \frac{1}{2}\left(b_j + \sqrt{b_j^2 + 1 - \frac{1}{2\lambda}} \right) \tag{7.63}$$

for the minimizer, see [70, Corollary 21]. Since $\|\boldsymbol{b}\|_p$ is bounded, we also have $\|\boldsymbol{b}\|_\infty \leqslant b_{\max}$ for all s, i.e., $b_j \leqslant b_{\max}$ for all $j = 1, \ldots, s_\ell$ and all s_ℓ. We denote by $\alpha_{\max}$ the value of (7.63) with b_j replaced by $b_{\max}$. We have $\alpha_j \leqslant \alpha_{\max}$ for all $j = 1, \ldots, s_\ell$ and all s_ℓ, and $\alpha_j - b_j \geqslant \alpha_{\max} - b_{\max}$. Furthermore, $\varrho_j(\lambda) \leqslant \varrho_{\max}(\lambda)$ for all j and all s, where $\varrho_{\max}(\lambda)$ is the value of (7.59) with α_j replaced by $\alpha_{\max}$.
From the definition of $\tilde{b}_j$ we see that $\tilde{b}_j \leqslant \sqrt{\pi} 2 b_j$, so by setting $\lambda = \frac{p}{2-p}$ and $\tau_\lambda := \frac{4\pi\varrho_{\max}(\lambda)^{\frac{1}{\lambda}}}{(\alpha_{\max}-b_{\max})3(\ln 2)^2}$, we have

$$\begin{aligned}
S_\lambda \leqslant \sum_{\mathfrak{u}\subseteq\{1:s_\ell\}} \big((|\mathfrak{u}|+2)!(|\mathfrak{u}|+6)\big)^p \prod_{j\in\mathfrak{u}} (\tau_\lambda b_j^2)^{\frac{p}{2}} &= \sum_{k=0}^{s_\ell} \big((k+2)!(k+6)\big)^p \sum_{\mathfrak{u}\subseteq\{1:s_\ell\},|\mathfrak{u}|=k} \prod_{j\in\mathfrak{u}} (\tau_\lambda b_j^2)^{\frac{p}{2}} \\
&\leqslant \sum_{k=0}^{s_\ell} \frac{\big((k+2)!(k+6)\big)^p}{k!} \tau_\lambda^{\frac{p}{2}k} \left(\sum_{j=1}^{s_\ell} b_j^p \right)^k \\
&\leqslant \sum_{k=0}^{\infty} \frac{\big((k+2)!(k+6)\big)^p}{k!} \tau_\lambda^{\frac{p}{2}k} \|\boldsymbol{b}\|_p^{pk} < \infty .
\end{aligned}$$

The finiteness follows by the ratio test, because $p < 1$. □

8 One-shot learning of surrogates

The reduced formulation (3.25) of the optimal control problem (3.23) subject to (3.24) is fundamentally based on the assumption that the forward problem can be solved exactly in each iteration, i.e., an existing algorithm for the solution of the state equation is embedded into an optimization loop. Thereby it is usually preferable to compute the gradient using a sensitivity or adjoint approach, cf., Chapter 4. However, the main drawback of this approach is that it requires the repeated costly solution of the (possibly nonlinear) state equation, even in the initial stages when the control variables are still far from their optimal value. This drawback can be partially overcome by carrying out the early optimization steps with a coarsely discretized PDE and/or only few samples from the space of parameters, cf., Section 7.2.
In this section, we will follow a different approach, which solves the optimization problem and the forward problem simultaneously by treating both, the design and the state variables, as optimization variables. Various names for the simultaneous solution of the design and state equation exist: all-at-once, one-shot method, piggy-back iterations etc., see, e.g., [14].
To be more precise, the state and the control variables are coupled through the constraint, which is kept explicitly during the optimization. To this end, we will consider the residual of (3.24)

$$e(u, z) = \mathcal{A}u - \mathcal{B}z\,,$$

where $e(u, z) : L^q_\mu(U, V) \times \mathcal{Z} \to L^q(U, W')$. In the following we will focus on penalty methods and refer to [92] and the references therein for other penalization strategies. A penalty method solves a constrained optimization problem by solving a sequence of unconstrained problems. Using, for instance, a quadratic penalty method in the present context, one aims to find a sequence of minimizers (z_k, u_k), given by

$$(z_k, u_k) = \arg\min_{z_k, u_k} \left(J(u_k, z_k) + \frac{\lambda_k}{2} \|e(u_k, z_k)\|^2_{L^q_\mu(U,W')} \right),$$

that converges to the minimizer $(z^*, u(z^*))$ of the constrained problem (3.25). The disadvantage of penalty methods is that the penalty parameter λ_k needs to be sent to infinity which renders the resulting k-th problem increasingly ill-conditioned. This problem can be avoided by using exact penalty methods, which will be subject to future work.
In the second part of this chapter, we will transer the developed ideas for surrogates in one-shot optimization to the setting of Bayesian inverse problems, see Section 8.2.

8.1 Surrogates in one-shot optimization under uncertainty

In this section we revisit the optimal control problem described in Section 4.3. We are interested in the situation, where a high resolution is needed for accurate approximations

of the solution of the operator equation, or a PDE solution, repectively. In this case the empirical approximation of the risk measure is of high computational costs as for each data point the underlying operator equation, or PDE model, needs to be solved.
Despite recent advances in PDE-constrained optimization under uncertainty problems, the incorporation of uncertainty in form of random parameters or random fields is still not feasible for many PDE models due to the significant increase in the computational complexity of the resulting optimization or control problems. The use of surrogate models, i.e., the replacement of the computational expensive solution of the forward model by approximations which are usually cheap to evaluate, is thus a promising direction in order to reduce the overall computational effort.
However, the surrogates need to be trained or calibrated in advance. In particular, in the context of optimization under uncertainty, a surrogate is needed for every feasible control in order to perform, e.g., for the numerical computation of the optimal control. One promising remedy to this issue lies in one-shot approaches, see e.g., [78, 144] and [73], where one-shot ideas are successfully generalized for the training of so-called residual neural networks.
Our framework is based on one-shot optimization approaches [14], where we reformulate the constrained optimal control problem as an unconstrained one via a penalization method. More precisely, in order to force the feasibility with respect to the model constraints, we include a penalty parameter allowing for an increasing weight on the penalization term. This setting allows the straightforward incorporation of surrogate. We replace the optimization with respect to the infinite-dimensional PDE solution by a parameterized family of functions, where the resulting optimization task is with respect to the parameters describing the surrogates. Examples of surrogates include polynomial series representations, neural networks, Gaussian process approximations and low rank approximations. We discuss various choices in Section 8.1.2. However, the suggested approach is not limited to the surrogates discussed here. Furthermore, we note that from a Bayesian perspective, this parameter controls the model error, i.e., increasing the penalization parameter corresponds to vanishing model noise, see Section 8.2 below.
In this section, we analyze the dependence of the optimization error on the number of data points as well as on the weight on the penalization. Moreover, we propose a stochastic gradient descent method in order to implement the resulting empirical risk minimization problem. In this section, we make the following contributions:

- We formulate a penalized empirical risk minimization problem and provide a consistency result in terms of large data limit as well as increasing penalty parameters. More precisely, we split the error in an error term decreasing with number of data points independently of the penalty parameter as well as in an error term decreasing in the strength of penalization independently of the number of data points.

- We formulate a stochastic gradient descent method in order to solve the penalized risk minimization problem where we allow an adaptive increase of the penalty parameter avoiding numerical instabilities due to high variance. Under suitable assumptions we prove convergence of the proposed stochastic gradient descent method. We verify the assumptions for linear surrogates.

- We test our proposed approach numerically, where we apply a linear as well as a nonlinear surrogate model. The linear surrogate model is based on a polynomial expansion, while the nonlinear surrogate model is described as a neural network.

8.1.1 Problem formulation

Let us briefly revisit the optimal control problem described in Section 4.3, where we choose the expected value as a risk measure. Our goal of computation is the following optimal control problem

$$\min_{z\in\mathcal{Z}_{ad},u\in\mathcal{Y}_{ad}} J(u,z)\,,\quad J(u,z) := \frac{1}{2}\int_U \|\mathcal{Q}u-\widehat{u}\|_{\mathfrak{J}}^2\,\mathrm{d}\mu(\boldsymbol{y}) + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2\,, \tag{8.1}$$

subject to the parametric linear operator equation in $L^p_\mu(U,W')$

$$\mathcal{A}u = \mathcal{B}z\,, \tag{8.2}$$

for $p=2$, a Hilbert space $\mathcal{Z}$ with $\mathcal{Z}_{ad}\subset\mathcal{Z}$, $\mathcal{Y}_{ad}\subset L^2_\mu(U,V)$, and a Hilbert space $\mathfrak{J}$, $\widehat{u}\in\mathfrak{J}$, $\mathcal{Q}\in\mathcal{L}(V,\mathfrak{J})$, $\mathcal{B}\in\mathcal{L}(\mathcal{Z},W')$. In particular, the operators $\mathcal{B}$ and $\mathcal{Q}$ are not dependent on $\boldsymbol{y}$ and thus can be uniformly bounded for all $\boldsymbol{y}$, i.e., $\|\mathcal{B}\|_{\mathcal{L}(\mathcal{Z},W')}\leqslant C_1$ and $\|\mathcal{Q}\|_{\mathcal{L}(V,\mathfrak{J})}\leqslant C_2$ for some $C_1,C_2>0$ and all $\boldsymbol{y}\in U$. This implies in particular, that $\mathcal{B}z\in L^p_\mu(U,W')$ for all p and all deterministic controls $z\in\mathcal{Z}$ and $\mathcal{Q}u\in L^2_\mu(U,\mathfrak{J})$ for all $u\in L^2_\mu(U,V)$. Moreover, we assume that $\mathcal{Q}$ and $\mathcal{B}$ have bounded inverse, i.e., $\|\mathcal{B}^{-1}\|_{\mathcal{L}(W',\mathcal{Z})}\leqslant C_3$ and $\|\mathcal{Q}^{-1}\|_{\mathcal{L}(\mathfrak{J},V)}\leqslant C_4$ for $C_3,C_4>0$. In particular, we assume that $\mathcal{A}$ is a boundedly invertible operator as described in Section 3.7.

We refer to Theorem 3.7.4 for the existence and uniqueness of solutions of the optimal control problem and to Theorem 4.2.11 for the optimality conditions. Moreover, we recall from Remark 3.7.5 that the solution of the optimal control problem remains unaffected by the choice of stating the constraint (8.2) in $L^2_\mu(U,W')$ or equivalently for all $\boldsymbol{y}\in U$ in W'.

In the previous chapters we considered the optimal control problems in their reduced formulations, see, e.g., (4.15), (4.36), or (4.52), assuming that the forward problem can be solved exactly in each iteration. Hence, for the actual computation, an existing algorithm for the solution of the state equation is embedded into an optimization loop. This approach requires the repeated costly solution of the state equation, even in the initial stages when the design variables are still far from their optimal value.

In this section, we will follow a one-shot approach, which solves the optimization problem and the forward problem simultaneously by treating both, the control and the state variables, as optimization variables. The state and the control variables are coupled through the PDE constraint, which is kept explicitly as a side constraint during the optimization. Specifically, we define the residual $e(u,z): L^2_\mu(U,V)\times\mathcal{Z}\to L^2_\mu(U,W')$ of (8.2) as

$$e(u,z) := \mathcal{A}u-\mathcal{B}z\,,$$

and employ a quadratic penalty method. In particular, we solve the constrained optimization problem (8.1) subject to (8.2) by solving a sequence of unconstrained optimization problems, i.e., we aim to find a sequence of (unique, global) minimizers (z_k,u_k), given by

$$(z_k,u_k) = \operatorname*{arg\,min}_{z_k,u_k}\left(J(u_k,z_k)+\frac{\lambda_k}{2}\|e(u_k,z_k)\|^2_{L^2_\mu(U,W')}\right), \tag{8.3}$$

that converges to the minimizer $(z^*,u(z^*))$ of the constrained problem (8.1) subject to (8.2).

In the following subsection there is a detailed presentation of different surrogates that might later be used as substitute for u_k in (8.3).

8.1.2 Surrogates

In many applications in the field of uncertainty quantification the forward model is computationally expensive to solve. Consequently, replacing the solution of the forward model by a surrogate, that is cheap to evaluate, can be a tremendous advantage.
For instance, neural networks (NN) have been successfully applied to various classes of PDEs, cp. e.g., [11, 57, 82, 120, 126, 149, 163] and also as approximation to the underlying model [44, 122]. For parametric PDEs, generalized polynomial chaos expansion haven been extensively studied, cf., [30] for an overview on approximation results. Recently, Gaussian processes haven been suggested for solving general nonlinear PDEs [26]. Here, we propose a general framework, which allows to include all different surrogate models in a one-shot approach.
In the next sections we will analyze the optimization problem in which the parametric mapping is replaced by a surrogate, i.e., the mapping

$$u^{\boldsymbol{y}} : U \to V$$

is replaced by a surrogate

$$u(\boldsymbol{\theta}, \boldsymbol{y}) : \Theta \times U \to V$$

where the $\boldsymbol{\theta} \in \Theta$ are the parameters of the surrogate.
Possible surrogates include for instance

- a power series of the form

$$u(\boldsymbol{\theta}, \boldsymbol{y}) = \sum_{\boldsymbol{\nu} \in \mathcal{F}} \boldsymbol{\theta}_{\boldsymbol{\nu}} \boldsymbol{y}^{\boldsymbol{\nu}} \tag{8.4}$$

- an orthogonal series of the form

$$u(\boldsymbol{\theta}, \boldsymbol{y}) = \sum_{\boldsymbol{\nu} \in \mathcal{F}} \boldsymbol{\theta}_{\boldsymbol{\nu}} P_{\boldsymbol{\nu}} \,, \quad P_{\boldsymbol{\nu}} := \prod_{j \geqslant 1} P_{\nu_j}(y_j) \,, \tag{8.5}$$

  where P_k is the Legendre polynomial of degree k defined on $[-1, 1]$ and normalized with respect to the uniform measure, i.e., such that $\int_{-1}^{1} |P_k(t)|^2 \frac{\mathrm{d}t}{2} = 1$.

- a neural network $u(\boldsymbol{\theta}, \boldsymbol{y}) : \Theta \times U \to V$, $(\boldsymbol{\theta}, \boldsymbol{y}) \mapsto u(\boldsymbol{\theta}, \boldsymbol{y})$ with $L \in \mathbb{N}$ layers, defined by the recursion

$$\begin{aligned} \boldsymbol{x}_0 &:= \boldsymbol{y} \,, \\ \boldsymbol{x}_\ell &:= \sigma(\boldsymbol{W}_\ell \boldsymbol{x}_{\ell-1} + \boldsymbol{b}_\ell) \,, \quad \text{for } \ell = 1, \dots, L-1 \,, \\ u(\boldsymbol{\theta}, \boldsymbol{y}) &:= \boldsymbol{W}_L \boldsymbol{x}_{L-1} + \boldsymbol{b}_L \,. \end{aligned} \tag{8.6}$$

  Here the parameters $\boldsymbol{\theta} \in \Theta := \times_{\ell=1}^{L} (\mathbb{R}^{N_\ell \times N_{\ell-1}} \times \mathbb{R}^{N_\ell})$ are a sequence of matrix-vector tuples

$$\boldsymbol{\theta} = \big((W_\ell, b_\ell)\big)_{\ell=1}^{L} = \big(\boldsymbol{W}_1, \boldsymbol{b}_1), (\boldsymbol{W}_2, \boldsymbol{b}_2), \dots, (\boldsymbol{W}_L, \boldsymbol{b}_L)\big) \,,$$

  and the activation function σ is applied component-wise to vector-valued inputs.

- Gaussian process or kernel based approximations. Recently, a general framework for the approximation of solution of nonlinear pdes has been proposed in [26]. The authors demonstrate the efficiency of Gaussian processes for nonlinear problems and derive a rigorous convergence analysis. We refer to [26] for more details, in particular also to the references therein.

- reduced basis or low rank approaches, which haven been demonstrated to efficiently approximate the solution of the forward problem even in high- or infinte-dimensional settings, see e.g., [8, 134].

There has been a lot of research towards efficient surrogates, in particular in the case of parametric PDEs and the above list is by far not exhaustive. We provide in the following a general framework to train surrogates simultaneously with the optimization step and illustrate the ansatz in numerical experiments for polynomial chaos and neural network approximations.
Based on the smoothness of the underlying function, approximation results of the above surrogates can be stated. To this end, we recall that the solution $u(\boldsymbol{y})$ of a parametric linear operator equation (3.19) is an analytic function with respect to the parameters $\boldsymbol{y}$, if the linear operators $A(\boldsymbol{y}) \in \mathcal{L}(V, W')$ are isomorphisms and as long as the operator A and the right-hand side z are parameterized in an analytical way, see e.g., [165, Theorem 1.2.37], or Theorem 4.6.13, which in addition provides bounds on the partial derivatives with respect to the parameters. Moreover, recall that analytic functions between Banach spaces admit holomorphic extensions, i.e., for analytic $f : U \to Y$ between two real Banach spaces X and Y with $U \subseteq X$ open, there exists an open set $\widetilde{U} \subseteq X_{\mathbb{C}}$[23] and a holomorphic extension $\widetilde{f} : \widetilde{U} \to Y_{\mathbb{C}}$ such that $U \subseteq \widetilde{U}$ and $\widetilde{f}|_U = f$, see [165, Proposition 1.2.33]. To quantify the smoothness of the underlying function we will use the notion of $(\boldsymbol{b}, \epsilon)$-holomorphy of a function, which is a sufficient criterion for many approximation results, see [149] and the referenes therein: Given a monotonically decreasing sequence $\boldsymbol{b} = (b_j)_{j\in\mathbb{N}}$ of positive real numbers that satisfies $\boldsymbol{b} \in \ell^p(\mathbb{N})$ for some $p \in (0, 1]$, a continuous function $u^{\boldsymbol{y}} : U \to V$ is called $(\boldsymbol{b}, \epsilon)$-holomorphic if for any sequence $\boldsymbol{\rho} := (\rho_j)_{j\geqslant 1} \in [1, \infty)^{\mathbb{N}}$, satisfying

$$\sum_{j\geqslant 1} (\rho_j - 1) b_j \leqslant \epsilon \,,$$

for some $\epsilon > 0$, there exists a complex extension $\widetilde{u} : B_{\boldsymbol{\rho}} \to V_{\mathbb{C}}$ of u, where $B_{\boldsymbol{\rho}} := \times_{j\in\mathbb{N}} B_{\rho_j}$, with $\widetilde{u}(\boldsymbol{y}) = u(\boldsymbol{y})$ for all $\boldsymbol{y} \in U$, such that $w \mapsto \widetilde{u}(\boldsymbol{w}) : B_{\boldsymbol{\rho}} \to V_{\mathbb{C}}$ is holomorphic as a function in each variable $w_j \in B_{\rho_j}$, $j \in \mathbb{N}$ with uniform bound

$$\sup_{\boldsymbol{w}\in B_{\boldsymbol{\rho}}} \|u(\boldsymbol{w})\|_{V_{\mathbb{C}}} \leqslant C \,.$$

The sequence $\boldsymbol{b}$ determines the size of the domains of the holomorphic extension, i.e., the faster the decay in $\boldsymbol{b}$, the faster the radii ρ_j may increase. Furthermore, the summability exponent p of the sequence $\boldsymbol{b} \in \ell^p(\mathbb{N})$ will determine the algebraic convergence rates of the surrogates below.
From [30, Corollary 3.11] we know that a $(\boldsymbol{b}, \epsilon)$-holomorphic function admits an unconditionally convergent Taylor generalized polynomial chaos expansion, i.e., the series in (8.4)

[23] For a real Banach space V, its complexification is the space $V_{\mathbb{C}} := V + iV$ with the Taylor norm $\|v + iw\|_{V_{\mathbb{C}}} := \sup_{t\in[0,2\pi)} \|\cos(t)v - \sin(t)w\|_V$ for all $v, w \in V$ and i denoting the imaginary unit.

with coefficients $\boldsymbol{\theta}_{\boldsymbol{\nu}} := \frac{1}{\boldsymbol{\nu}!}\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} u^{\boldsymbol{y}}|_{y=0}$ converges unconditionally towards $u^{\boldsymbol{y}}$ in $L^\infty(U,V)$. Moreover, let Λ_s be the set of indices that correspond to the s largest $\|\boldsymbol{\theta}_{\boldsymbol{\nu}}\|_V$, then we have

$$\sup_{\boldsymbol{y}\in U} \|u^{\boldsymbol{y}} - \sum_{\boldsymbol{\nu}\in\Lambda_s} \boldsymbol{\theta}_{\boldsymbol{\nu}} \boldsymbol{y}^{\boldsymbol{\nu}}\|_V \leqslant C(s+1)^{-\frac{1}{p}+1},$$

with $C = \|(\|\boldsymbol{\theta}_{\boldsymbol{\nu}}\|_V)_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^p} < \infty$.
Furthermore, we known from [30, Corollary 3.10] that a $(\boldsymbol{b},\epsilon)$-holomorphic function admits an unconditionally convergent Legendre series expansion, i.e., the series in (8.5) with coefficients $\boldsymbol{\theta}_{\boldsymbol{\nu}} := \int_U u^{\boldsymbol{y}} L_{\boldsymbol{\nu}}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}$ converges unconditionally towards $u^{\boldsymbol{y}}$ in $L^2_\mu(U,V)$ with

$$\|u - \sum_{\boldsymbol{\nu}\in\Lambda_n} \boldsymbol{\theta}_{\boldsymbol{\nu}} P_{\boldsymbol{\nu}}\|_{L^2_\mu(U,V)} \leqslant C(s+1)^{-\frac{1}{p}+\frac{1}{2}},$$

where $C = \|(\|\boldsymbol{\theta}_{\boldsymbol{\nu}}\|_V)_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^p} < \infty$ and Λ_s denotes the indices with the s largest $\|\boldsymbol{\theta}_{\boldsymbol{\nu}}\|_V$.
More recent results [149] show that $(\boldsymbol{b},\epsilon)$-holomorphic functions, i.e., the parametric solution manifold $U \ni \boldsymbol{y} \mapsto u^{\boldsymbol{y}} \in V$, can be expressed by a neural network of finite size. In [149] the authors illustrate this for the elliptic example (see Section 4.1) for $d = 1$ under the additional regularity assumptions that $z \in L^2(D)$ and $a(\boldsymbol{y}) \in W^{1,\infty}(D)$ for all $\boldsymbol{y} \in U$. Therefore, let $0 < q_V \leqslant q_X < 2$ and denote $p_V := (1/q_V + 1/2)^{-1} \in (0,1)$ and $p_X := (1/q_X + 1/2)^{-1} \in (0,1)$. Let $\beta_V := (\beta_{V,j})_{j\in\mathbb{N}} \in (0,1)^{\mathbb{N}}$ and $\beta_X := (\beta_{X,j})_{j\in\mathbb{N}} \in (0,1)^{\mathbb{N}}$ be monotonically decreasing sequences such that $\beta_V \in \ell^{q_V}(\mathbb{N})$ and $\beta_X \in \ell^{q_X}(\mathbb{N})$ and such that

$$\left\|\frac{\sum_{j\in\mathbb{N}} \beta_{V,j}^{-1}|\psi_j(\cdot)|}{a_0(\cdot)}\right\|_{L^\infty(D)} < 1, \quad \left\|\frac{\sum_{j\in\mathbb{N}} \beta_{X,j}^{-1}|\psi_j(\cdot)|}{a_0(\cdot)}\right\|_{L^\infty(D)} < 1, \quad \left\|\sum_{j\in\mathbb{N}} \beta_{X,j}^{-1}|\psi_j'(\cdot)|\right\|_{L^\infty(D)} < \infty.$$

Then (see, [149, Theorem 4.8]) there is a constant $C > 0$ such that for every $s \in \mathbb{N}$, there exists a ReLU neural network (i.e., a neural network (8.6) with activation function $\sigma(\boldsymbol{x}) = \max(0,\boldsymbol{x})$) denoted by $u(\boldsymbol{\theta},\boldsymbol{y})$ with $s+1$ input units and for a number $\mathcal{N} \geqslant s$ with $r = \min(1,(1+p_V^{-1})/(1+p_V^{-1}-p_X^{-1}))$, it holds

$$\sup_{\boldsymbol{y}\in U} \|u^{\boldsymbol{y}} - u(\boldsymbol{\theta},\boldsymbol{y})\|_V \leqslant C\mathcal{N}^{-r}.$$

Furthermore, for any $s \in \mathbb{N}$ the size and depth of the neural network can be bounded by

$$\begin{aligned}\operatorname{size}(u(\boldsymbol{\theta},\boldsymbol{y})) &\leqslant C(1+\mathcal{N}\log(\mathcal{N})\log(\log(\mathcal{N})))\\ \operatorname{depth}(u(\boldsymbol{\theta},\boldsymbol{y})) &\leqslant C(1+\log(\mathcal{N})\log(\log(\mathcal{N}))),\end{aligned}$$

where the size of neural network is defined as the total number of nodes plus the total number of nonzero weights $\operatorname{size}(u(\boldsymbol{\theta},\boldsymbol{y})) := |\{(i,j,\ell) : (W_{i,j})_\ell \neq 0\}| + \sum_{\ell=0}^{L} N_\ell$ and the depth of a neural network $\operatorname{depth}(u(\boldsymbol{\theta},\boldsymbol{y})) = L-1$ is the number of hidden layers. Setting $\boldsymbol{b} := (\|\psi_j\|_{L^\infty(D)})_{j\in\mathbb{N}}$ and assuming in addition to (AE2)-(AE5) that $\boldsymbol{b} \in \ell^p(\mathbb{N})$ for some $p \in (0,1)$, the parametric solution $u^{\boldsymbol{y}}$ of the uniformly elliptic problem (4.8) is $(\boldsymbol{b},\epsilon)$-holomorphic, see e.g., [30, 149]. We conclude that, under these additional assumptions, the convergence results of the polynomial expansions and the approximation result of the neural network apply to the elliptic PDE problem in Section 4.1. We note also that the series expansions (8.4) and (8.5) are linear in its parameters $\boldsymbol{\theta}$, whereas the neural network is nonlinear in its parameters due to the nonlinear activation function σ.

8.1.3 Consistency analysis

In our consistency analysis, we are going to analyse the proposed penalty method, see (8.3), with respect to the penalty parameter λ_k and the number of i.i.d. data points n, denoted as $(\boldsymbol{y}^i)_{i=1}^n$, which are used to approximate the expected values with respect to $\boldsymbol{y}$. Thereby, we assume that the state $u_k^{\boldsymbol{y}}$ has been parameterized by a surrogate $u(\boldsymbol{\theta}, \boldsymbol{y})$, see Section 8.1.2, and the penalty parameters $(\lambda_k)_{k\in\mathbb{N}}$ are monotonically increasing to infinity. In particular, we try to connect the following optimization problems:

(cRM) The original constrained risk minimization (cRM) problem

$$\min_{z,\boldsymbol{\theta}} \ \frac{1}{2}\mathbb{E}_{\boldsymbol{y}}[\|\mathcal{Q}u(\boldsymbol{\theta},\boldsymbol{y}) - \widehat{u}\|_{\mathfrak{Y}}^2] + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2$$

subjected to

$$\mathbb{E}_{\boldsymbol{y}}[\|e(u(\boldsymbol{\theta},\boldsymbol{y}),z)\|_{W'}^2] = 0.$$

We assume there exists a unique solution of this problem, which we will denote by $(z_\infty^*, \boldsymbol{\theta}_\infty^*)$.

(pRM) The penalized risk minimization (pRM) problem

$$\min_{z,\boldsymbol{\theta}} \ \frac{1}{2}\mathbb{E}_{\boldsymbol{y}}[\|\mathcal{Q}u(\boldsymbol{\theta},\boldsymbol{y}) - \widehat{u}\|_{\mathfrak{Y}}^2] + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2 + \frac{\lambda_k}{2}\mathbb{E}_{\boldsymbol{y}}[\|e(u(\boldsymbol{\theta},\boldsymbol{y}),z)\|_{W'}^2]. \tag{8.7}$$

We assume there exists a unique solution denoted by $(z_\infty^k, \boldsymbol{\theta}_\infty^k)$.

(pERM) The penalized empirical risk minimization (pERM) problem

$$\min_{z,\boldsymbol{\theta}} \ \frac{1}{2n}\sum_{i=1}^n \|\mathcal{Q}u(\boldsymbol{\theta},\boldsymbol{y}^i) - \widehat{u}\|_{\mathfrak{Y}}^2 + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2 + \frac{\lambda_k}{2}\frac{1}{n}\sum_{i=1}^n \|e(u(\boldsymbol{\theta},\boldsymbol{y}),z)\|_{W'}^2 \tag{8.8}$$

We assume there exists a unique solution denoted by $(z_n^k, \boldsymbol{\theta}_n^k)$.

For simplicity, in the following we denote $x = (z, \boldsymbol{\theta}) \in \mathcal{X} := \mathcal{Z} \times \mathbb{R}^d$ and define the functions

$$\begin{aligned} f &: \mathcal{X} \times U \to \mathbb{R}_+, \quad \text{with} \quad f(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2}\|\mathcal{Q}u(\boldsymbol{\theta},\boldsymbol{y}) - \widehat{u}\|_{\mathfrak{Y}}^2 + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2, \\ g &: \mathcal{X} \times U \to \mathbb{R}_+, \quad \text{with} \quad g(\boldsymbol{x},\boldsymbol{y}) = \|e(u(\boldsymbol{\theta},\boldsymbol{y}),z)\|_{W'}^2, \end{aligned}$$

where we assume here and in the following that W' is a Hilbert space, which implies that $L_\mu^2(U, W')$ is a Hilbert space.
Since the dimension truncation error can be controlled by Theorem 5.4.4, we neglect this error contribution in the present consistency analysis.
To simplify the notation we work in the following of this section with gradients instead of Fréchet derivatives. The gradient of a functional $\hat{J} : \mathcal{Z} \to \mathbb{R}$ is the unique representer in $\mathcal{Z}$ of the Fréchet derivative $\hat{J}'(z)$ of $\hat{J}$, which belongs to $\mathcal{Z}'$, i.e., $\hat{J}'(z) = R_{\mathcal{Z}}\nabla\hat{J}$, where $R_Z : Z \to Z'$ denotes the Riesz operator in the Hilbert space $\mathcal{Z}$ given by $\langle z_1, R_{\mathcal{Z}} z_2\rangle_{\mathcal{Z},\mathcal{Z}'} = \langle z_1, z_2\rangle_{\mathcal{Z},\mathcal{Z}}$.

Convergence of pRM to cRM

We start with the error dependence on the penalty parameter λ_k. The following is a long known result (see e.g., [127, Theorem 1]) providing unique existence of solutions as well as convergence towards the unconstrained problem for increasing penalty parameter λ_k.

Theorem 8.1.1. *Let H_1 and H_2 be two Hilbert spaces and let $f(\boldsymbol{x})$ be a functional on H_1 and the constraint $h(x)$ be an operator from H_1 into H_2. Moreover, suppose*

- *there exists a unique global minimizer $x^* \in \mathcal{X}$ of the problem*

$$\min_{x \in \mathcal{X}} f(\boldsymbol{x}) \quad \text{s.t. } h(x) = 0 \text{ in } H_2 \,.$$

- *that $\nabla_x f(x), \nabla_x^2 f(x)$ and $\nabla_x h(x), \nabla_x^2 h(x)$ exist with*

$$\begin{aligned} &\|\nabla_x^2 f(x) - \nabla_x^2 f(y)\|_{\mathcal{L}(H_1, \mathcal{L}(H_1, \mathbb{R}))} \leqslant L_1 \|x - y\|_{H_1} \\ \text{and} \quad &\|\nabla_x^2 h(x) - \nabla_x^2 h(y)\|_{\mathcal{L}(H_1, \mathcal{L}(H_1, H_2))} \leqslant L_2 \|x - y\|_{H_1} \,. \end{aligned}$$

- *the linear operator $\nabla_x h(\boldsymbol{x}^x)$ is non-degenerate, i.e., $\|(\nabla_x h(x^*))^* y\|_{H_2} \geqslant c\|y\|_{H_2}$ for $c > 0$ and for all $y \in H_2$.*

- *the self-adjoint operator $\nabla_x^2 L(x^*, y^*)$ is positive definite, i.e., $\langle \nabla_x^2 L(x^*, y^*)\tilde{x}, \tilde{x} \rangle \geqslant m\|\tilde{x}\|_{H_1}^2$ for $m > 0$ and all $\tilde{x} \in H_1$. Here, the functional L denotes the Lagrangian, y^* denotes the Lagrange multiplier corresponding to x^*, and the Lagrange multiplier rule is applicable because of the first three assumptions.*

Then, for sufficiently large $\lambda_k > 0$, there exists a unique minimizer x_k^* of the problem

$$\min_{x \in H_1} f(x) + \frac{\lambda_k}{2} \|h(x)\|_{H_2}^2$$

which satisfies

$$\|x_k^* - x^*\|_{H_1} \leqslant \frac{C}{2\lambda_k} \|y^*\|_{H_2} \quad \text{and} \quad \|\lambda_k h(x_k^*) - y^*\|_{H_1} \leqslant \frac{C}{2\lambda_k} \|y^*\|_{H_2} \,.$$

This theorem holds in infinite-dimensional Hilbert spaces H_1 and H_2, in that case the derivatives with respect to $x \in H_1$ in the theorem are Fréchet derivatives. For our problem at hand with $\boldsymbol{x} = (z, \boldsymbol{\theta}) \in \mathcal{X} = \mathcal{Z} \times \mathbb{R}^d$ the assumptions need to be satisfied for $f(\boldsymbol{x}) := \frac{1}{2}\mathbb{E}_{\boldsymbol{y}}[\|\mathcal{Q}u(\boldsymbol{\theta}, \boldsymbol{y}) - \widehat{u}\|_{\mathfrak{J}}^2] + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2$, $h(\boldsymbol{x}) = e(u(\boldsymbol{\theta}, \boldsymbol{y}, z))$ and $g(\boldsymbol{x}) := \|h(\boldsymbol{x})\|_{W'}^2$ based on the spaces $H_1 = \mathcal{X}$ and $H_2 = L_\mu^2(U, W')$. Here we need the assumption that W' is a Hilbert space, such that $L_\mu^2(U, W')$ is a Hilbert space. In this case the H_1-norm is just the norm on $\mathcal{X}$, e.g. $\|(z, \boldsymbol{\theta})\|_{\mathcal{X}} = (\|z\|_{\mathcal{Z}}^2 + \|\boldsymbol{\theta}\|_2^2)^{1/2}$, and the H_2-norm is $\|\cdot\|_{L_\mu^2(U,W')} := (\mathbb{E}_{\boldsymbol{y}}[\|\cdot\|_{W'}^2])^{1/2}$. If a surrogate satisfies the assumptions of the preceding theorem, the convergence of the minimizers of the (pRM) problem to the minimizer of the (cRM) problem is guaranteed.

Lemma 8.1.2. *Suppose that f and g satisfy the assumptions of Theorem 8.1.1. Then the solution of the (cRM) problem converges to the solution of the (cRM) problem, in the sense that there exists $C_1 > 0$ independent of n such that*

$$\|(z_\infty^k, \boldsymbol{\theta}_\infty^k) - (z_\infty^*, \boldsymbol{\theta}_\infty^*)\|_{\mathcal{X}}^2 \leqslant \frac{C_1}{\lambda_k^2}.$$

Convergence of pERM to pRM

The following result describes the error arising due to the empirical approximation of the risk function uniformly in the penalization.

Lemma 8.1.3. *Suppose that f is convex and g is strongly convex, i.e., $\nabla^2_{\boldsymbol{x}} g(\boldsymbol{x}, \boldsymbol{y}) > m \cdot \mathcal{I}$ for all $x \in \mathcal{Z} \times \mathbb{R}^d$ and $\boldsymbol{y} \in U$. Let $\lambda_0 = 1$ and assume that*

$$\operatorname{Tr}(\operatorname{Cov}_{\boldsymbol{y}}(\nabla_{\boldsymbol{x}}(f(\boldsymbol{x}, \boldsymbol{y})) + \frac{\lambda_0}{2} g(\boldsymbol{x}, \boldsymbol{y}))) < \infty.$$

Then the solution of the (pERM) problem converges uniformly in λ_k to the solution of the (pRM) problem, in the sense that there exists a constant $C_2 > 0$ independent of λ_k such that

$$\mathbb{E}_{\boldsymbol{y}}[\|(z_n^k, \boldsymbol{\theta}_n^k) - (z_\infty^k, \boldsymbol{\theta}_\infty^k)\|^2_{\mathcal{X}}] \leqslant \frac{C_2}{n}.$$

Proof. Under the above assumption the objective function in (8.8) is strongly convex. The unique solution $\boldsymbol{x}_n^k$ satisfies

$$\frac{1}{n} \sum_{i=1}^n \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_n^k, \boldsymbol{y}^i) + \frac{\lambda_k}{2} \frac{1}{n} \sum_{i=1}^n \nabla_{\boldsymbol{x}} g(\boldsymbol{x}_n^k, \boldsymbol{y}^i) = 0.$$

Similarly, the unique minimzer of (8.7) is characterized by

$$\nabla_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{y}}[f(\boldsymbol{x}_\infty^k, \boldsymbol{y})] + \frac{\lambda_k}{2} \nabla_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{y}}[g(\boldsymbol{x}_\infty^k, \boldsymbol{y})] = 0.$$

We are now interested in the discrepancy of $\boldsymbol{x}_n^k$ and $\boldsymbol{x}_\infty^k$. We define the functions

$$\Psi^k(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{y}}[f(\boldsymbol{x}, \boldsymbol{y})] + \frac{\lambda_k}{2} \mathbb{E}_{\boldsymbol{y}}[g(\boldsymbol{x}, \boldsymbol{y})]$$

and its empirical approximation

$$\Psi_n^k(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^n f(\boldsymbol{x}, \boldsymbol{y}^i) + \frac{\lambda_k}{2} \frac{1}{n} \sum_{i=1}^n g(\boldsymbol{x}, \boldsymbol{y}^i).$$

By the strong convexity of Ψ_n^k it follows that

$$\begin{aligned} \|\boldsymbol{x}_n^k - \boldsymbol{x}_\infty^k\|^2_{\mathcal{X}} &\leqslant \frac{1}{m \cdot \frac{\lambda_k}{2}} \langle \boldsymbol{x}_n^k - \boldsymbol{x}_\infty^k, \nabla_{\boldsymbol{x}} \Psi_n^k(\boldsymbol{x}_n^k) - \nabla_{\boldsymbol{x}} \Psi_n^k(\boldsymbol{x}_\infty^k) \rangle_{\mathcal{X}} \\ &= \frac{1}{m \cdot \frac{\lambda_k}{2}} \langle \boldsymbol{x}_n^k - \boldsymbol{x}_\infty^k, \nabla_{\boldsymbol{x}} \Psi^k(\boldsymbol{x}_\infty^k) - \nabla \Psi_n^k(\boldsymbol{x}_\infty^k) \rangle_{\mathcal{X}} \end{aligned}$$

using the stationarity of the minimizers. Applying the Cauchy–Schwarz inequality leads to

$$\|\boldsymbol{x}_n^k - \boldsymbol{x}_\infty^k\|_{\mathcal{X}} \leqslant \frac{1}{m \cdot \frac{\lambda_k}{2}} \|\nabla \Psi_n^k(\boldsymbol{x}_\infty^k) - \nabla_{\boldsymbol{x}} \Psi^k(\boldsymbol{x}_\infty^k)\|_{\mathcal{X}'}.$$

Next, we note that for $\psi^k(\boldsymbol{x}) := f(\boldsymbol{x}, \boldsymbol{y}) + \frac{\lambda_k}{2} g(\boldsymbol{x}, \boldsymbol{y})$

$$\|\nabla \Psi_n^k(\boldsymbol{x}_\infty^k) - \nabla_{\boldsymbol{x}} \Psi^k(\boldsymbol{x}_\infty^k)\|^2_{\mathcal{X}'}$$

$$= \operatorname{Tr}\left(\left(\frac{1}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{x}}\psi^k(\boldsymbol{x}_\infty^k,\boldsymbol{y}^i) - \mathbb{E}_{\boldsymbol{y}}[\nabla_{\boldsymbol{x}}\psi^k(\boldsymbol{x}_\infty^k,\boldsymbol{y})]\right)\right.$$
$$\left.\otimes\left(\frac{1}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{x}}\psi^k(\boldsymbol{x}_\infty^k,\boldsymbol{y}^i) - \mathbb{E}_{\boldsymbol{y}}[\nabla_{\boldsymbol{x}}\psi^k(\boldsymbol{x}_\infty^k,\boldsymbol{y})]\right)\right)$$

and by taking the expectation

$$\mathbb{E}[\|\nabla\Psi_n^k(\boldsymbol{x}_\infty^k) - \nabla_{\boldsymbol{x}}\Psi^k(\boldsymbol{x}_\infty^k)\|_{\mathcal{X}'}^2] = \frac{1}{n}\operatorname{Tr}(\operatorname{Cov}(\nabla_{\boldsymbol{x}}\psi^k(\boldsymbol{x}_\infty^k,\boldsymbol{y}))).$$

It holds that

$$\begin{aligned}
\operatorname{Tr}(\operatorname{Cov}(\nabla_{\boldsymbol{x}}\psi^k(\boldsymbol{x},\boldsymbol{y}))) &= \operatorname{Tr}(\operatorname{Cov}(\nabla_{\boldsymbol{x}}f(\boldsymbol{x},\boldsymbol{y}) + \frac{\lambda_k}{2}\nabla_{\boldsymbol{x}}g(\boldsymbol{x},\boldsymbol{y})))\\
&= \operatorname{Tr}(\operatorname{Cov}(\nabla_{\boldsymbol{x}}f(\boldsymbol{x},\boldsymbol{y})) + \frac{\lambda_k}{2}\operatorname{Cov}(\nabla_{\boldsymbol{x}}f(\boldsymbol{x},\boldsymbol{y}),\nabla_{\boldsymbol{x}}g(\boldsymbol{x},\boldsymbol{y}))\\
&\quad + \frac{\lambda_k}{2}\operatorname{Cov}(\nabla_{\boldsymbol{x}}g(\boldsymbol{x},\boldsymbol{y}),\nabla_{\boldsymbol{x}}f(\boldsymbol{x},\boldsymbol{y})) + \frac{\lambda_k^2}{4}\operatorname{Cov}(\nabla_{\boldsymbol{x}}g(\boldsymbol{x},\boldsymbol{y})))\\
&\leqslant \operatorname{Tr}(\max\{1,\frac{\lambda_k^2}{4}\}(\operatorname{Cov}(\nabla_{\boldsymbol{x}}f(\boldsymbol{x},\boldsymbol{y})) + \operatorname{Cov}(\nabla_{\boldsymbol{x}}f(\boldsymbol{x},\boldsymbol{y}),\nabla_{\boldsymbol{x}}g(\boldsymbol{x},\boldsymbol{y}))\\
&\quad + \operatorname{Cov}(\nabla_{\boldsymbol{x}}g(\boldsymbol{x},\boldsymbol{y}),\nabla_{\boldsymbol{x}}f(\boldsymbol{x},\boldsymbol{y})) + \operatorname{Cov}(\nabla_{\boldsymbol{x}}g(\boldsymbol{x},\boldsymbol{y})))\\
&= \max\{1,\frac{\lambda_k^2}{4}\}\operatorname{Tr}(\operatorname{Cov}(\nabla_{\boldsymbol{x}}\psi^0(\boldsymbol{x},\boldsymbol{y})))
\end{aligned}$$

with $\lambda_0 = 1$. Finally, we obtain the bound

$$\mathbb{E}[\|\boldsymbol{x}_n^k - \boldsymbol{x}_\infty^k\|_{\mathcal{X}}^2] \leqslant C_{m,\lambda_k}\frac{1}{n}\operatorname{Tr}(\operatorname{Cov}(\nabla_{\boldsymbol{x}}\psi^0(\boldsymbol{x},\boldsymbol{y}))),$$

where

$$C_{m,\lambda_k} := \frac{1}{m^2}\frac{\max\{1,\frac{\lambda_k^2}{4}\}}{\frac{\lambda_k^2}{4}} \leqslant \frac{1}{m^2}.$$

□

8.1.4 Convergence of pERM to cRM

Finally, we are ready to prove consistency in the sense that solutions of the (pERM) converge to solutions of the original (cRM). We can use Lemma 8.1.3 and Lemma 8.1.2 by applying

$$\mathbb{E}[\|(z_n^k,\boldsymbol{\theta}_n^k) - (z_\infty^*,\boldsymbol{\theta}_\infty^*)\|_{\mathcal{X}}^2] \leqslant 2\underbrace{\mathbb{E}[\|(z_n^k,\boldsymbol{\theta}_n^k) - (z_\infty^k,\boldsymbol{\theta}_\infty^k)\|_{\mathcal{X}}^2]}_{\text{pERM to pRM}} + 2\underbrace{\|(z_\infty^k,\boldsymbol{\theta}_\infty^k) - (z_\infty^*,\boldsymbol{\theta}_\infty^*)\|_{\mathcal{X}}^2}_{\text{pRM to cRM}}.$$

Theorem 8.1.4. *Suppose that f and g satisfy the assumptions of Lemma 8.1.2 and Lemma 8.1.3. Then the the solution $(z_n^k,\boldsymbol{\theta}_n^k)$ is consistent in the sense that there exists $C_1, C_2 > 0$ such that*

$$\mathbb{E}[\|(z_n^k,\boldsymbol{\theta}_n^k) - (z_\infty^*,\boldsymbol{\theta}_\infty^*)\|_{\mathcal{X}}^2] \leqslant \frac{C_1}{\lambda_k^2} + \frac{C_2}{n}.$$

For a surrogate that is linear in its parameters, i.e., $u(\boldsymbol{\theta}, z) = B^{\boldsymbol{y}}\boldsymbol{\theta}$, the first assumptions of Theorem 8.1.1 (and thus Lemma 8.1.2) follows from the strict convexity of f. The second assumption is clearly satisfied since for a linear surrogate, the constraint h is linear and hence the objective f is quadratic. The third assumption is true if we have for all $y \in L^2_\mu(U, W')$ that

$$\mathbb{E}[\|(\nabla_{(\boldsymbol{\theta},z)} h(\boldsymbol{\theta}^*, z^*))^* y\|^2_{\mathcal{X}}] \geqslant c\|y\|^2_{L^2_\mu(U,W')} \,.$$

In the setting with linear surrogate the operator $(\nabla_{(\boldsymbol{\theta},z)} h(\boldsymbol{\theta}^*, z^*))^* : L^2_\mu(U, W') \to \mathcal{X}$ simplifies to $((\mathcal{A}^{\boldsymbol{y}} B^{\boldsymbol{y}})^*, -\mathcal{B}^*)^\top$, such that $\mathbb{E}[\|((\mathcal{A}^{\boldsymbol{y}} B^{\boldsymbol{y}})^*, -\mathcal{B}^*)^\top y\|^2_{\mathcal{X}}] = \mathbb{E}[\|(\mathcal{A}^{\boldsymbol{y}} B^{\boldsymbol{y}})^* y\|^2_{\mathbb{R}^d} + \|\mathcal{B}^* y\|^2_{\mathcal{Z}}] \geqslant \mathbb{E}[(a^2_{\min}\sigma_{\min}(B^{\boldsymbol{y}}(B^{\boldsymbol{y}})^*) + \sigma_{\min}(\mathcal{B}\mathcal{B}^*))\|y\|^2_{L^2_\mu(U,W')}]$. Here $\sigma_{\min}(\mathcal{B}\mathcal{B}^*) > 0$ since $\mathcal{B}$ has bounded inverse. Furthermore, from the linearity of the constraint follows that the Hessian of the Lagrangian simplifies to the Hessian of the objective function $\nabla^2_{\boldsymbol{x}} f(\boldsymbol{\theta}, z) = \operatorname{diag}(\mathbb{E}[(B^{\boldsymbol{y}})^* B^{\boldsymbol{y}}], \alpha \cdot \mathcal{B}^*\mathcal{B})$. The fourth condition is thus satisfied if $\sigma_{\min}(\mathbb{E}[(B^{\boldsymbol{y}})^* B^{\boldsymbol{y}}]) \geqslant M$ for some $M > 0$ and $\alpha > 0$. If $\sigma_{\min}(\mathbb{E}[(B^{\boldsymbol{y}})^* B^{\boldsymbol{y}}]) = 0$ the fourth condition can still be satisfied by introducing a quadratic penalty on the surrogate parameters in the objective function.

8.1.5 Stochastic gradient descent for pRM problems

In order to solve the (pRM) problem we propose to apply the stochastic gradient descent (SGD) method. This means, instead of solving the (pERM) problem offline for large but fixed number of data n, we solve the (pRM) online. Therefore, we further propose to adaptively increase the penalty parameter λ_k within the SGD.
We first formulate a general convergence result for the penalized SGD method, which we then apply to verify the convergence in the setting of our PDE-constrained optimization problem given by the (cRM) problem.

Algorithm 7 Penalized stochastic gradient descent method with adaptive penalty parameter.

Require: x_0, $\beta = (\beta_k)_{k=1}^n$, $(\lambda_k)_{k=1}^n$, i.i.d. sample $(\boldsymbol{y}^k)_{k=1}^n \sim \boldsymbol{y}$.
1: **for** $k = 0, 1, \ldots, n-1$ **do**
2: $\quad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \beta_k \nabla_{\boldsymbol{x}}[f(\boldsymbol{x}_k, \boldsymbol{y}^k) + \lambda_k g(\boldsymbol{x}_k, \boldsymbol{y}^k)]$
3: **end for**

The sequence of step sizes β is assumed to satisfy the Robbins-Monro condition

$$\sum_{j=1}^{\infty} \beta_k = \infty, \quad \sum_{j=1}^{\infty} \beta_k^2 < \infty,$$

which means that β_k converges to zero, but not too fast [132]. In the following theorem we present sufficient conditions under which the resulting estimate $\boldsymbol{x}_n$ from Algorithm 7 converges to the solution of the (pRM) with penalty parameter choice $\bar{\lambda} \gg 0$, i.e., to

$$\boldsymbol{x}^* \in \arg\min_{x \in \mathcal{X}} \; \Psi_{\bar{\lambda}}(\boldsymbol{x}), \quad \Psi_{\bar{\lambda}}(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{y}}[f(\boldsymbol{x}, \boldsymbol{y}) + \frac{\bar{\lambda}}{2} g(\boldsymbol{x}, \boldsymbol{y})].$$

Theorem 8.1.5. *We assume that the objective function satisfies*

$$\langle x - \boldsymbol{x}^*, \nabla_{\boldsymbol{x}} \Psi_{\bar{\lambda}}(\boldsymbol{x}) \rangle_{\mathcal{X}} > c\|x - \boldsymbol{x}^*\|^2_{\mathcal{X}} \tag{8.9}$$

for all $x \in \mathcal{X}$ and some $c > 0$ and that for each λ_k we have

$$\mathbb{E}_{\boldsymbol{y}}[\|\nabla_{\boldsymbol{x}}[f(\boldsymbol{x},\boldsymbol{y}) + \lambda_k g(\boldsymbol{x},\boldsymbol{y})\|_{\mathcal{X}}^2] < a_k + b_k\|x - \boldsymbol{x}^*\|_{\mathcal{X}}^2, \tag{8.10}$$

where (a_k) and (b_k) are monotonically increasing with $a_0, b_0 > 0$ and $a_k \leqslant \bar{a},\ b_k \leqslant \bar{b}$. Furthermore, we assume that the discrepancy of the penalized stochastic gradients can be bounded locally by

$$\sup_{x\in\mathcal{X},\ \|x\|_{\mathcal{X}}\leqslant R} \|\mathbb{E}_{\boldsymbol{y}}[(\lambda_k - \bar{\lambda})\nabla_{\boldsymbol{x}} g(\boldsymbol{x},\boldsymbol{y})]\|_{\mathcal{X}}^2 \leqslant \kappa_1(R)|\lambda_k - \bar{\lambda}|^2, \tag{8.11}$$

for some $\kappa_1(R) > 0$, $R > 0$. Suppose that $|\lambda_k - \bar{\lambda}|^2$ is monotonically decreasing and $\beta_k \leqslant c/b_k$, then it holds true that

$$\mathbb{E}[\mathbb{1}_{\{\|\boldsymbol{x}_k\|\leqslant R\}}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_{\mathcal{X}}^2] \leqslant \left(\mathbb{E}[\|x_0 - \boldsymbol{x}^*\|_{\mathcal{X}}^2 + 2\bar{a}\sum_{j=1}^{\infty}\beta_j^2\right) C_n + \frac{2\kappa_1(R)}{c^2}|\lambda_0 - \bar{\lambda}|^2,$$

with

$$C_n := \min_{k\leqslant n}\max\{\prod_{j=k+1}^{n}(1 - c\beta_j), \frac{\bar{a}}{c}\beta_k\}$$

converging to zero for $n \to \infty$. Further, for an adaptive choice of the penalty parameter λ_k such that $\frac{2\kappa_1(R)}{c^2}|\lambda_k - \bar{\lambda}|^2 \leqslant D\beta_k$ we obtain

$$\mathbb{E}[\mathbb{1}_{\{\|\boldsymbol{x}_k\|\leqslant R\}}(\boldsymbol{x}_k)\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_{\mathcal{X}}^2] \leqslant \left(\mathbb{E}[\|x_0 - \boldsymbol{x}^*\|_{\mathcal{X}}^2 + 2(\bar{a} + \frac{D}{c})\sum_{j=1}^{\infty}\beta_j^2\right) C_n.$$

Proof. The proof is based on a Gronwall-type argument and similar to the proof of Proposition 3.3 in [20] and can be found in [79]. □

Remark 8.1.6. *We note that the restriction boundedness of $\|\boldsymbol{x}_k\|_{\mathcal{X}} \leqslant R$ is a techniqual reason for the proof and can be forced through a projection onto the ball $B_R = \{\boldsymbol{x} \in \mathcal{X} \mid \|\boldsymbol{x}\|_{\mathcal{X}} \leqslant R\}$ by*

$$\mathcal{P}_R : \mathcal{X} \to B_R, \quad \textit{with} \quad \mathcal{P}_R(\boldsymbol{x}) = \arg\min_{\boldsymbol{x}'\in B_R}\|\boldsymbol{x} - \boldsymbol{x}'\|_{\mathcal{X}}.$$

The projected stochastic gradient descent method then evolves through the update

$$\boldsymbol{x}_{k+1} = \mathcal{P}_R\left(\boldsymbol{x}_k - \beta_k\nabla_{\boldsymbol{x}}[f(\boldsymbol{x}_k,\boldsymbol{y}^k) + \lambda_k g(\boldsymbol{x}_k,\boldsymbol{y}^k)]\right).$$

The above proof remains the same since the projection operator is nonexpansive in the sense that

$$\begin{aligned}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|_{\mathcal{X}} &= \|\mathcal{P}_R\left(\boldsymbol{x}_k - \beta_k\nabla_{\boldsymbol{x}}[f(\boldsymbol{x}_k,\boldsymbol{y}^k) + \lambda_k g(\boldsymbol{x}_k,\boldsymbol{y}^k)]\right) - \mathcal{P}_R(\boldsymbol{x}^*)\|_{\mathcal{X}}^2\\ &\leqslant \|\boldsymbol{x}_k - \beta_k\nabla_{\boldsymbol{x}}[f(\boldsymbol{x}_k,\boldsymbol{y}^k) + \lambda_k g(\boldsymbol{x}_k,\boldsymbol{y}^k)] - \boldsymbol{x}^*\|_{\mathcal{X}}^2.\end{aligned}$$

Moreoever, the presented convergence result in Theorem 8.1.5 indicates how to control the ratio between the sequence of step sizes (β_k) and the penalty parameters (λ_k) based on the dependence of $\kappa(R)$ on $R > 0$.

8.1.6 Application to linear surrogate models

In this section we verify that a surrogate, that is linear in its parameters, satisfies the assumptions of Theorem 8.1.5. Therefore, we assume in this section that the surrogate is of the following form:

$$u(\boldsymbol{\theta}, \boldsymbol{y}) := B^{\boldsymbol{y}}\boldsymbol{\theta} \tag{8.12}$$

for surrogate parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{y}$-dependent operator $B^{\boldsymbol{y}} : \mathbb{R}^d \to V$. The following two lemmas will help to prove this result:

Lemma 8.1.7. *Let $g(\boldsymbol{x}, \boldsymbol{y}) := \|e(u(\boldsymbol{\theta}, \boldsymbol{y}), z)\|^2$, with $e(u(\boldsymbol{\theta}, \boldsymbol{y}), z) := A^{\boldsymbol{y}}u(\boldsymbol{\theta}, \boldsymbol{y}) - \mathcal{B}z = A^{\boldsymbol{y}}B^{\boldsymbol{y}}\boldsymbol{\theta} - \mathcal{B}z$, where $\boldsymbol{x} = (\boldsymbol{\theta}, z)$, and with bounded largest eigenvalue $\sigma_{\max}(A^{\boldsymbol{y}}) \leqslant a_{\max}$ for all $\boldsymbol{y} \in U$. Then it holds that*

$$\|\nabla_{\boldsymbol{x}} g(\boldsymbol{x}, \boldsymbol{y})\|_{\mathcal{X}}^2 \leqslant 4(a_{\max}^2 \sigma_{\max}(B^{\boldsymbol{y}}(B^{\boldsymbol{y}})^*) + \sigma_{\max}(\mathcal{B}\mathcal{B}^*))\, g(\boldsymbol{x}, \boldsymbol{y})\,.$$

Proof. We have

$$\begin{aligned}
\|\nabla_{\boldsymbol{x}} g(\boldsymbol{x}, \boldsymbol{y})\|_{\mathcal{X}}^2 &= \|2(A^{\boldsymbol{y}}B^{\boldsymbol{y}})^*(A^{\boldsymbol{y}}B^{\boldsymbol{y}}\boldsymbol{\theta} - \mathcal{B}z)\|^2 + \|2\mathcal{B}^*(\mathcal{B}z - A^{\boldsymbol{y}}B^{\boldsymbol{y}}\boldsymbol{\theta}\|_{\mathcal{Z}}^2 \\
&\leqslant 4(a_{\max}^2 \sigma_{\max}(B^{\boldsymbol{y}}(B^{\boldsymbol{y}})^*) + \sigma_{\max}(\mathcal{B}\mathcal{B}^*))\|A^{\boldsymbol{y}}B^{\boldsymbol{y}}\boldsymbol{\theta} - \mathcal{B}z\|_{L^2_\mu(U,W')}^2 \\
&= 4(a_{\max}^2 \sigma_{\max}(B^{\boldsymbol{y}}(B^{\boldsymbol{y}})^*) + \sigma_{\max}(\mathcal{B}\mathcal{B}^*)) g(\boldsymbol{x}, \boldsymbol{y})\,.
\end{aligned}$$

□

Lemma 8.1.8. *Let $g(\boldsymbol{x}, \boldsymbol{y}) := \|e(u(\boldsymbol{\theta}, \boldsymbol{y}), z)\|_{L^2_\mu(U,W')}^2$, with $e(u(\boldsymbol{\theta}, \boldsymbol{y}), z) := A^{\boldsymbol{y}}u(\boldsymbol{\theta}, \boldsymbol{y}) - \mathcal{B}z = A^{\boldsymbol{y}}B^{\boldsymbol{y}}\boldsymbol{\theta} - \mathcal{B}z$, where $\boldsymbol{x} = (\boldsymbol{\theta}, z)$, and with bounded largest eigenvalue $\sigma_{\max}(A^{\boldsymbol{y}}) \leqslant a_{\max}$ for all $\boldsymbol{y} \in U$. Then it holds that*

$$g(\boldsymbol{x}, \boldsymbol{y}) \leqslant 2(a_{\max}^2 \sigma_{\max}((B^{\boldsymbol{y}})^* B^{\boldsymbol{y}}) + \sigma_{\max}(\mathcal{B}^*\mathcal{B}))\|x\|_{\mathcal{X}}^2\,.$$

Proof.

$$\begin{aligned}
g(\boldsymbol{x}, \boldsymbol{y}) = \|A^{\boldsymbol{y}}B^{\boldsymbol{y}}\boldsymbol{\theta} - \mathcal{B}z\|_{L^2_\mu(U,W')}^2 &\leqslant 2(\|A^{\boldsymbol{y}}B^{\boldsymbol{y}}\boldsymbol{\theta}\|_{L^2_\mu(U,W')}^2 + \|\mathcal{B}z\|_{L^2_\mu(U,W')}^2) \\
&\leqslant 2(a_{\max}^2 \sigma_{\max}((B^{\boldsymbol{y}})^* B^{\boldsymbol{y}})\|\boldsymbol{\theta}\|_{\mathbb{R}^d}^2 + \sigma_{\max}(\mathcal{B}^*\mathcal{B}))\|z\|_{\mathcal{Z}}^2) \\
&\leqslant 2(a_{\max}^2 \sigma_{\max}((B^{\boldsymbol{y}})^* B^{\boldsymbol{y}}) + \sigma_{\max}(\mathcal{B}^*\mathcal{B}))\|x\|_{\mathcal{X}}^2\,.
\end{aligned}$$

□

Theorem 8.1.9. *Let $\alpha > 0$ and $\sigma_{\min}(\mathbb{E}[(B^{\boldsymbol{y}})^* B^{\boldsymbol{y}}]) > 0$, then a surrogate of the form (8.12) satisfies the assumptions of Theorem 8.1.5.*

Proof. Firstly, we show that

$$\langle x - \boldsymbol{x}^*, \nabla_{\boldsymbol{x}} \Psi_{\bar{\lambda}}(\boldsymbol{x})\rangle_{\mathcal{X}} > c\|x - \boldsymbol{x}^*\|_{\mathcal{X}}^2$$

is true for a constant $c > 0$ and all $x \in \mathcal{X}$. To verify this assumption, we first show that $\Psi_{\bar{\lambda}}$ is c-strongly convex. The c-strong convexity is equivalent to

$$\langle \boldsymbol{x} - \boldsymbol{x}^*, \nabla_{\boldsymbol{x}} \Psi_{\bar{\lambda}}(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \Psi_{\bar{\lambda}}(\boldsymbol{x}^*)\rangle_{\mathcal{X}} \geqslant c\|\boldsymbol{x} - \boldsymbol{x}^*\|_{\mathcal{X}}^2\,.$$

Note that

$$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x},\boldsymbol{y}) = (\mathcal{Q}B^{\boldsymbol{y}})^*(\mathcal{Q}B^{\boldsymbol{y}}\boldsymbol{\theta} - \widehat{u}) \qquad \nabla_{\boldsymbol{\theta}} \frac{\lambda}{2} g(\boldsymbol{x},\boldsymbol{y}) = \lambda(A^{\boldsymbol{y}}B^{\boldsymbol{y}})^*(A^{\boldsymbol{y}}B^{\boldsymbol{y}}\boldsymbol{\theta} - \mathcal{B}z)$$
$$\nabla_z f(\boldsymbol{x},\boldsymbol{y}) = \alpha z \qquad \nabla_z \frac{\lambda}{2} g(\boldsymbol{x},\boldsymbol{y}) = -\lambda\mathcal{B}^*(A^{\boldsymbol{y}}B^{\boldsymbol{y}}\boldsymbol{\theta} - \mathcal{B}z)$$

Using the linearity of the surrogate in its parameters, i.e., $u(\boldsymbol{\theta},\boldsymbol{y}) := B^{\boldsymbol{y}}\boldsymbol{\theta}$, we obtain

$$\begin{aligned}
&\langle \boldsymbol{x}-\boldsymbol{x}^*, \nabla_{\boldsymbol{x}}\Psi_{\bar\lambda}(\boldsymbol{x}) - \nabla_{\boldsymbol{x}^*}\Psi_{\bar\lambda}(\boldsymbol{x}^*)\rangle_{\mathcal{X}} \\
&= \big\langle (\boldsymbol{\theta}-\boldsymbol{\theta}^*, z-z^*), \mathbb{E}\big[((\mathcal{Q}B^{\boldsymbol{y}})^*(\mathcal{Q}B^{\boldsymbol{y}}) + \bar\lambda(A^{\boldsymbol{y}}B^{\boldsymbol{y}})^*(A^{\boldsymbol{y}}B^{\boldsymbol{y}}))(\boldsymbol{\theta}-\boldsymbol{\theta}^*) \\
&\quad - \bar\lambda(A^{\boldsymbol{y}}B^{\boldsymbol{y}})^*\mathcal{B}(z-z^*), (\alpha+\bar\lambda\mathcal{B}^*\mathcal{B})(z-z^*) - \bar\lambda\mathcal{B}^*A^{\boldsymbol{y}}B^{\boldsymbol{y}}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\big]\big\rangle_{\mathcal{X}} \\
&= \langle \boldsymbol{\theta}-\boldsymbol{\theta}^*), \mathbb{E}[(\mathcal{Q}B^{\boldsymbol{y}})^*(\mathcal{Q}B^{\boldsymbol{y}})](\boldsymbol{\theta}-\boldsymbol{\theta}^*)\rangle_{\mathbb{R}^d} + \bar\lambda\|A^{\boldsymbol{y}}B^{\boldsymbol{y}}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2_{L^2_\mu(U,W')} \\
&\quad - \langle \boldsymbol{\theta}-\boldsymbol{\theta}^*, \bar\lambda(A^{\boldsymbol{y}}B^{\boldsymbol{y}})^*\mathcal{B}(z-z^*)\rangle_{\mathbb{R}^d} \\
&\quad + \alpha\|z-z^*\|^2_{\mathcal{Z}} + \bar\lambda\|\mathcal{B}(z-z^*)\|^2_{L^2_\mu(U,W')} - \langle z-z^*, \bar\lambda\mathcal{B}^*A^{\boldsymbol{y}}B^{\boldsymbol{y}}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\rangle_{\mathcal{Z}} \\
&= \langle \boldsymbol{\theta}-\boldsymbol{\theta}^*, \mathbb{E}[(\mathcal{Q}B^{\boldsymbol{y}})^*(\mathcal{Q}B^{\boldsymbol{y}})](\boldsymbol{\theta}-\boldsymbol{\theta}^*)\rangle_{\mathbb{R}^d} + \alpha\|z-z^*\|^2_{\mathcal{Z}} \\
&\quad + \bar\lambda\|\mathcal{B}(z-z^*) - A^{\boldsymbol{y}}B^{\boldsymbol{y}}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2_{L^2_\mu(U,W')} \\
&\geqslant \sigma_{\min}(\mathcal{Q}^*\mathcal{Q})\langle \boldsymbol{\theta}-\boldsymbol{\theta}^*, \mathbb{E}[(B^{\boldsymbol{y}})^*B^{\boldsymbol{y}}](\boldsymbol{\theta}-\boldsymbol{\theta}^*)\rangle_{\mathbb{R}^d} + \alpha\|z-z^*\|^2_{\mathcal{Z}} \\
&\geqslant c\|\boldsymbol{x}-\boldsymbol{x}^*\|^2_{\mathcal{X}}\,,
\end{aligned}$$

where $c = \min(\sigma_{\min}(\mathcal{Q}^*\mathcal{Q})\sigma_{\min}(\mathbb{E}[(B^{\boldsymbol{y}})^*B^{\boldsymbol{y}}]), \alpha)$. Note that $(\sigma_{\min}(\mathcal{Q}'\mathcal{Q}) > 0$ since $\mathcal{Q}$ is assumed to have a bounded inverse. Since $\alpha > 0$ and $\sigma_{\min}(\mathbb{E}[(B^{\boldsymbol{y}})^*B^{\boldsymbol{y}}]) > 0)$ by assumption, the assertion is true since $\boldsymbol{x}^*$ is a stationary point of $\Psi_{\bar\lambda}(\boldsymbol{x})$.
Secondly, we show that

$$\mathbb{E}[\|\nabla_{\boldsymbol{x}}\big(f(\boldsymbol{x},\boldsymbol{y}) + \lambda_k g(\boldsymbol{x},\boldsymbol{y})\big)\|^2_{\mathcal{X}}] \leqslant a_k + b_k\|x-\boldsymbol{x}^*\|^2_{\mathcal{X}}\,.$$

For a stationary point $\boldsymbol{x}^*$ of $\Psi_{\bar\lambda}(\boldsymbol{x})$ have that

$$0 = \nabla_{\boldsymbol{x}}\big(f(\boldsymbol{x}^*,\boldsymbol{y}) + \bar\lambda g(\boldsymbol{x}^*,\boldsymbol{y})\big)$$

and thus

$$\begin{aligned}
\|\nabla_{\boldsymbol{x}}\big(f(\boldsymbol{x},\boldsymbol{y}) + \lambda_k g(\boldsymbol{x},\boldsymbol{y})\big)\|^2_{\mathcal{X}} &= \|\nabla_{\boldsymbol{x}}\big(f(\boldsymbol{x},\boldsymbol{y}) + \lambda_k g(\boldsymbol{x},\boldsymbol{y})\big) - \nabla_{\boldsymbol{x}}\big(f(\boldsymbol{x}^*,\boldsymbol{y}) + \bar\lambda g(\boldsymbol{x}^*,\boldsymbol{y})\big)\|^2_{\mathcal{X}} \\
&= \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x},\boldsymbol{y}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*,\boldsymbol{y}) + \lambda_k\nabla_{\boldsymbol{x}} g(\boldsymbol{x},\boldsymbol{y}) - \bar\lambda\nabla_{\boldsymbol{x}} g(\boldsymbol{x}^*,\boldsymbol{y})\|^2_{\mathcal{X}} \\
&\leqslant 2\|\nabla_{\boldsymbol{x}}(f(\boldsymbol{x},\boldsymbol{y}) - f(\boldsymbol{x}^*,\boldsymbol{y}))\|^2_{\mathcal{X}} \\
&\quad + 2\|\lambda_k\nabla_{\boldsymbol{x}} g(\boldsymbol{x},\boldsymbol{y}) - \bar\lambda\nabla_{\boldsymbol{x}} g(\boldsymbol{x}^*,\boldsymbol{y})\|^2_{\mathcal{X}}\,.
\end{aligned}$$

For the first summand we have $\nabla_{\boldsymbol{x}}\big(f(\boldsymbol{x},\boldsymbol{y}) - f(\boldsymbol{x}^*,\boldsymbol{y})\big) = \big((\mathcal{Q}B^{\boldsymbol{y}})^*(\mathcal{Q}B^{\boldsymbol{y}})(\boldsymbol{\theta}-\boldsymbol{\theta}^*), \alpha(z - z^*)\big)$ and thus

$$\mathbb{E}[2\|\nabla_{\boldsymbol{x}}\big(f(\boldsymbol{x},\boldsymbol{y}) - f(\boldsymbol{x}^*,\boldsymbol{y})\big)\|^2_{\mathcal{X}}] \leqslant \tilde c^2\|x-\boldsymbol{x}^*\|^2_{\mathcal{X}}\,,$$

with $\tilde c = 2\mathbb{E}[\sigma_{\max}(\mathcal{Q}^*\mathcal{Q})\sigma_{\max}((B^{\boldsymbol{y}})^*B^{\boldsymbol{y}})] + \alpha$. Moreover, for the second summand we have

$$\begin{aligned}
\nabla_{\boldsymbol{x}}\big(\lambda_k g(\boldsymbol{x},\boldsymbol{y}) - \bar\lambda g(\boldsymbol{x}^*,\boldsymbol{y})\big) = \big(&2(A^{\boldsymbol{y}}B^{\boldsymbol{y}})^*((A^{\boldsymbol{y}}B^{\boldsymbol{y}})(\lambda_k\boldsymbol{\theta} - \bar\lambda\boldsymbol{\theta}^*) \\
&- \mathcal{B}(\lambda_k z - \bar\lambda z^*)), 2\mathcal{B}^*(\lambda_k z - \bar\lambda z^*) - A^{\boldsymbol{y}}B^{\boldsymbol{y}}(\lambda_k\boldsymbol{\theta} - \bar\lambda\boldsymbol{\theta}^*)\big)
\end{aligned}$$

$$= \nabla_{\boldsymbol{x}} g(\lambda_k \boldsymbol{x} - \bar{\lambda}\boldsymbol{x}^*, \boldsymbol{y})\,.$$

We can use this together with Lemma 8.1.7 and Lemma 8.1.8 to obtain

$$\begin{aligned}
&\mathbb{E}[\|\nabla_{\boldsymbol{x}} g(\lambda_k \boldsymbol{x} - \bar{\lambda}\boldsymbol{x}^*, \boldsymbol{y})\|_{\mathcal{X}}^2] \\
&\qquad \leqslant \mathbb{E}[4(a_{\max}^2 \sigma_{\max}(B^{\boldsymbol{y}}(B^{\boldsymbol{y}})^*) + \sigma_{\max}(\mathcal{B}\mathcal{B}^*))\, g(\lambda_k x - \bar{\lambda}\boldsymbol{x}^*, \boldsymbol{y})] \\
&\qquad \leqslant \mathbb{E}[4(a_{\max}^2 \sigma_{\max}(B^{\boldsymbol{y}}(B^{\boldsymbol{y}})^*) + \sigma_{\max}(\mathcal{B}\mathcal{B}^*)) \\
&\qquad\quad \times 2(a_{\max}^2 \sigma_{\max}((B^{\boldsymbol{y}})^* B^{\boldsymbol{y}}) + \sigma_{\max}(\mathcal{B}^*\mathcal{B}))\|\lambda_k x - \bar{\lambda}\boldsymbol{x}^*\|_{\mathcal{X}}^2] \\
&\qquad = \big(8a_{\max}^4 \mathbb{E}[\sigma_{\max}((B^{\boldsymbol{y}})^* B^{\boldsymbol{y}})\sigma_{\max}(B^{\boldsymbol{y}}(B^{\boldsymbol{y}})^*)] + 8\sigma_{\max}(\mathcal{B}\mathcal{B}^*)\sigma_{\max}(\mathcal{B}^*\mathcal{B}) \\
&\qquad\quad + \sigma_{\max}(\mathcal{B}\mathcal{B}^*)(2a_{\max}^2 \mathbb{E}[\sigma_{\max}((B^{\boldsymbol{y}})^* B^{\boldsymbol{y}})]) \\
&\qquad\quad + 4a_{\max}^2 \mathbb{E}[\sigma_{\max}(B^{\boldsymbol{y}}(B^{\boldsymbol{y}})^*)]\sigma_{\max}(\mathcal{B}^*\mathcal{B})\big) \\
&\qquad\quad \times \|\lambda_k x - \bar{\lambda}\boldsymbol{x}^*\|_{\mathcal{X}}^2 \\
&\qquad = \big(8a_{\max}^4 \mathbb{E}[\sigma_{\max}((B^{\boldsymbol{y}})^* B^{\boldsymbol{y}})\sigma_{\max}(B^{\boldsymbol{y}}(B^{\boldsymbol{y}})^*)] + 8\sigma_{\max}(\mathcal{B}\mathcal{B}^*)\sigma_{\max}(\mathcal{B}^*\mathcal{B}) \\
&\qquad\quad + \sigma_{\max}(\mathcal{B}\mathcal{B}^*)(2a_{\max}^2 \mathbb{E}[\sigma_{\max}((B^{\boldsymbol{y}})^* B^{\boldsymbol{y}})]) \\
&\qquad\quad + 4a_{\max}^2 \mathbb{E}[\sigma_{\max}(B^{\boldsymbol{y}}(B^{\boldsymbol{y}})^*)]\sigma_{\max}(\mathcal{B}^*\mathcal{B})\big) \\
&\qquad\quad \times 2\big(\lambda_k^2 \|\boldsymbol{x} - \boldsymbol{x}^*\|_{\mathcal{X}}^2 + (\lambda_k - \bar{\lambda})^2 \|\boldsymbol{x}^*\|_{\mathcal{X}}^2\big)\,.
\end{aligned}$$

We conclude that

$$\mathbb{E}[\|\nabla_{\boldsymbol{x}}\big(f(\boldsymbol{x}, \boldsymbol{y}) + \lambda_k g(\boldsymbol{x}, \boldsymbol{y})\big)\|_{\mathcal{X}}^2] \leqslant a_k + b_k \|x - \boldsymbol{x}^*\|_{\mathcal{X}}^2$$

holds for $a_0 \geqslant a_k = 2C_{ab}(\lambda_k - \bar{\lambda})^2 \|x^*\|_{\mathcal{X}}^2$ and $b_k = 2\tilde{c}^2 + 2C_{ab}\lambda_k^2$, where

$$\begin{aligned}
C_{ab} = \big(&8a_{\max}^4 \mathbb{E}[\sigma_{\max}((B^{\boldsymbol{y}})^* B^{\boldsymbol{y}})\sigma_{\max}(B^{\boldsymbol{y}}(B^{\boldsymbol{y}})^*)] + 8\sigma_{\max}(\mathcal{B}\mathcal{B}^*)\sigma_{\max}(\mathcal{B}^*\mathcal{B}) \\
&+ \sigma_{\max}(\mathcal{B}\mathcal{B}^*)(2a_{\max}^2 \mathbb{E}[\sigma_{\max}((B^{\boldsymbol{y}})^* B^{\boldsymbol{y}})]) + 4a_{\max}^2 \mathbb{E}[\sigma_{\max}(B^{\boldsymbol{y}}(B^{\boldsymbol{y}})^*)]\sigma_{\max}(\mathcal{B}^*\mathcal{B})\big).
\end{aligned}$$

Thirdly, we show that

$$\sup_{\boldsymbol{x} \in \mathcal{X}, \|\boldsymbol{x}\|_{\mathcal{X}} \leqslant R} \|\mathbb{E}[(\lambda_k - \bar{\lambda})\nabla_{\boldsymbol{x}} g(\boldsymbol{x}, \boldsymbol{y})]\|^2 \leqslant \kappa_1(R)|\lambda_k - \bar{\lambda}|^2\,,$$

for some $\kappa_1 > 0$.
We observe that

$$\begin{aligned}
\|\mathbb{E}[(\lambda_k - \bar{\lambda})\nabla_{\boldsymbol{x}} g(\boldsymbol{x}, \boldsymbol{y})]\|_{\mathcal{X}}^2 &\leqslant |\lambda_k - \bar{\lambda}|^2 \|\mathbb{E}[\nabla_{\boldsymbol{x}} g(\boldsymbol{x}, \boldsymbol{y})]\|_{\mathcal{X}}^2 \\
&\leqslant |\lambda_k - \bar{\lambda}|^2 \mathbb{E}[\|\nabla_{\boldsymbol{x}} g(\boldsymbol{x}, \boldsymbol{y})\|_{\mathcal{X}}^2] \\
&\leqslant |\lambda_k - \bar{\lambda}|^2 \mathbb{E}[4(a_{\max}^2 \sigma_{\max}(B^{\boldsymbol{y}}(B^{\boldsymbol{y}})^*) + \sigma_{\max}(\mathcal{B}\mathcal{B}^*))\, g(\boldsymbol{x}, \boldsymbol{y})] \\
&\leqslant |\lambda_k - \bar{\lambda}|^2 \mathbb{E}[4(a_{\max}^2 \sigma_{\max}(B^{\boldsymbol{y}}(B^{\boldsymbol{y}})^*) + \sigma_{\max}(\mathcal{B}\mathcal{B}^*)) \\
&\qquad 2(a_{\max}^2 \sigma_{\max}((B^{\boldsymbol{y}})^* B^{\boldsymbol{y}}) + \sigma_{\max}(\mathcal{B}^*\mathcal{B}))\|x\|_{\mathcal{X}}^2]\,.
\end{aligned}$$

Thus the claim holds with $\kappa_1(R) = \mathbb{E}[4(a_{\max}^2 \sigma_{\max}(B^{\boldsymbol{y}}(B^{\boldsymbol{y}})^*) + \sigma_{\max}(\mathcal{B}\mathcal{B}^*))$
$\times 2(a_{\max}^2 \sigma_{\max}((B^{\boldsymbol{y}})^* B^{\boldsymbol{y}}) + \sigma_{\max}(\mathcal{B}^*\mathcal{B}))R^2]$. □

Remark 8.1.10. *The assumption $\sigma_{\min}(\mathbb{E}[(B^{\boldsymbol{y}})^* B^{\boldsymbol{y}}]) > 0$ in theorem 8.1.9 can be dropped if a quadratic regularization on the surrogate parameters $\boldsymbol{\theta}$ is employed.*

8.1.7 Numerical experiments

The model problem in our numerical experiments is the Poisson equation, (4.1) – (4.3), on the unit square $D = (0,1)^2$. We use piecewise-linear finite elements on a uniform triangular mesh with meshwidth $h = 1/8$. The random input field is modelled as

$$a^{\boldsymbol{y}}(\boldsymbol{x}) = a_0(\boldsymbol{x}) + \sum_{j=1}^{s} y_j \frac{1}{(\pi^2(k_j^2 + \ell_j^2) + \tau^2)^{\vartheta}} \sin(\pi x_1 k_j) \sin(\pi x_2 \ell_j)\,,$$

where $a_0(\boldsymbol{x}) = 0.00001 + \| \sum_{j=1}^{s} \frac{1}{(\pi^2(k_j^2+\ell_j^2)+\tau^2)^{\vartheta}} \sin(\pi x_1 k_j) \sin(\pi x_2 \ell_j) \|_{L^\infty(D)}$, $s = 4$, $\vartheta = 0.25$, $\tau = 3$, $(k_j, \ell_j)_{j\in\{1,\dots,s\}} \in \{1,\dots,s\}^2$ and $y_j \sim \mathcal{U}([-1,1])$ i.i.d. for all $j = 1,\dots,s$. The variance of the resulting PDE solution $u^{\boldsymbol{y}}$, with right-hand side $z(\boldsymbol{x}) = x_2^2 - x_1^2$, is illustrated in Figure 8.1 and Figure 8.2. The mean and standard deviation is estimated using 10^5 Monte Carlo samples.
In the following numerical experiments we solve the (pERM) problem

$$\min_{(z,\boldsymbol{\theta})} \frac{1}{n} \sum_{i=1}^{n} \|u(\boldsymbol{\theta}, \boldsymbol{y}) - \widehat{u}\|^2 + \frac{\alpha}{2}\|z\|^2 + \lambda_k \frac{1}{n} \sum_{i=1}^{n} \|A^{\boldsymbol{y}} u(\boldsymbol{\theta}, \boldsymbol{y}) - z\|^2\,.$$

where $\alpha = 0.5$ and the target state $\widehat{u}$ is given as $\widehat{u} = \Delta^{-1} 100(x_2^2 - x_1^2)$. We solve the optimization problem using the so-called ADAM algorithm as implemented in `tensorflow` (see, e.g., [103]) and the `scipy` implementation of the L-BFGS method (see, e.g., [49]). The initial guess for the optimization routines is $(\boldsymbol{z}_0, \boldsymbol{\theta}_0)$ with $\boldsymbol{z}_0 = (0,\dots,0) \in \mathbb{R}^n$ and $\boldsymbol{\theta}_0 = (1,\dots,1) \in \mathbb{R}^d$. In our experiments we compared to two different surrogate models: the orthogonal Legendre polynomials, which are linear in the parameters $\boldsymbol{\theta}$ and a neural network, which is nonlinear in the parameters $\boldsymbol{\theta}$.
Recall the polynomial expansion from (8.5): $u(\boldsymbol{\theta}, \boldsymbol{y}) = \sum_{\boldsymbol{\nu}\in\mathbb{N}_0^s}^{|\boldsymbol{\nu}|=\ell} \boldsymbol{\theta}_{\boldsymbol{\nu}} P_{\boldsymbol{\nu}}(\boldsymbol{y})$ of degree $\ell = 1, 2, 3$, with $P_{\boldsymbol{\nu}} = \prod_{k=1}^{s} P_{\nu_k}(y_k)$ and P_{ν_k} is the k-th order Legendre polynomial. The number

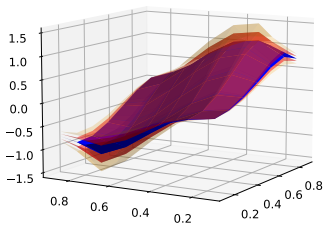


Figure 8.1: Mean of the states (blue) plus/minus 1 (red) and 2 (orange) standard deviations. Here only the values in the interior of the domain D are plotted for better illustration.

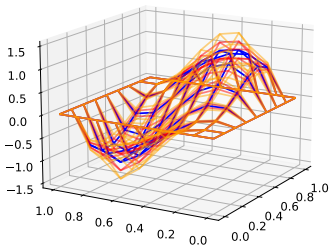


Figure 8.2: Mean of the states (blue) plus/minus 1 (red) and 2 (orange) standard deviations. The variance is zero on the boundary of the domain ∂D.

of parameters $\boldsymbol{\theta}$ increases rapidly as the order of the polynomials increases. In fact, $\boldsymbol{\theta} \in \mathbb{R}^{n_{\text{FEM}} \times n_{\text{Pol}}}$, where n_{FEM} denotes the number of degrees of freedoms of the finite element method and n_{Pol} denotes the number of polynomials given by $n_{\text{Pol}} = \frac{(\ell+s)!}{\ell! s!}$, i.e., $n_{\text{Pol}} = 15$ if $\ell = 2$ and $n_{\text{Pol}} = 35$ if $\ell = 3$ for $s = 4$. Consequently, the Legendre polynomial expansions have 245, 735, and 1715 parameters to be determined during the optimization.

A nonlinear surrogate we are testing is a neural network, as defined in (8.6) of size $[4, 9, 9, 9, 49]$, i.e., $(\boldsymbol{W}_1, \boldsymbol{b}_1) \in \mathbb{R}^{9\times 4} \times \mathbb{R}^9$, $(\boldsymbol{W}_2, \boldsymbol{b}_2) \in \mathbb{R}^{9\times 9} \times \mathbb{R}^9$, $(\boldsymbol{W}_3, \boldsymbol{b}_3) \in \mathbb{R}^{9\times 9} \times \mathbb{R}^9$, $(\boldsymbol{W}_4, \boldsymbol{b}_4) \in \mathbb{R}^{49\times 9} \times \mathbb{R}^9$, and thus with a total number of 715 parameters. The activation function we are using is the sigmoid function $\sigma(x) := 1/(1 + \exp(-x))$.

In our first experiment, we verify Lemma 8.1.3. To this end, we set $\lambda_k = 1$ for all k and solve the (pERM) problem multiple times for increasing sample size $n = 2^\ell$, with $\ell = 1, \dots, 13$. In this experiment the surrogate is a Legendre polynomial expansion of degree 2. The reference solution $(z_{\text{ref}}, \boldsymbol{\theta}_{\text{ref}})$ is computed by using $n_{\text{ref}} = 2^{14}$ Monte Carlo samples. The observed rate in Figure 8.3 aligns nicely with the predicted rate in Lemma 8.1.3.

Next, we verify Lemma 8.1.2. We fix the sample size $n = 100$ and solve the (pERM) problem for increasing penalty parameter λ_k. The surrogate in this experiment is again the Legendre polynomial expansion of degree 2. In Figure 8.4 we observe the rate predicted by Lemma 8.1.2. Here the reference solution is computed for $\lambda_k \approx 1.7 \cdot 10^6$. For numerical stability we regularize the problem in this experiment by adding the term $10^{-5}\|\boldsymbol{\theta}\|^2$ to the objective function of the (pERM) problem. In the experiments for Figure 8.3 and Figure 8.4 we used the L-BFGS method to solve the optimization problems.

As predicted by the theory, we also observe this rate in the following experiment, where we use the ADAM algorithm as implemented in `tensorflow` and increase λ_k linearly in each iteration k of the ADAM algorithm. The reference solution $(z_{\text{ref}}, u^{\boldsymbol{y}}_{\text{ref}})$ of the (cRM) problem is computed using the L-BFGS as implemented in `scipy`. We perform this experiment for the NN and the Legendre polynomial expansions of order $1, 2$ and 3. For each of the surrogates considered, we observe the expected rate of the error in the

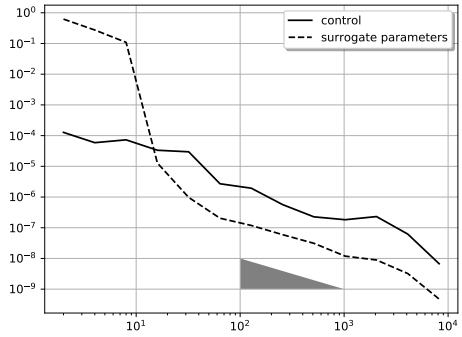


Figure 8.3: Convergence for increasing sample size. Squared error of the optimal controls $\|z - z_{\text{ref}}\|^2$ and squared error of the optimal surrogate parameters $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{ref}}\|^2$.

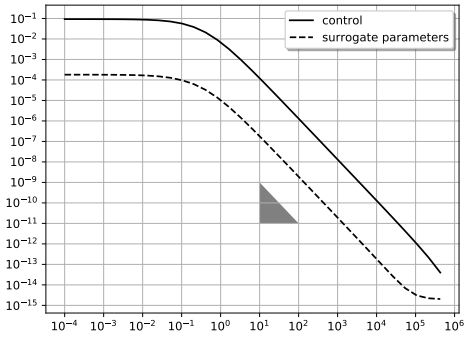


Figure 8.4: Convergence for increasing penalty parameter λ_k. Squared error of the optimal controls $\|z - z_{\text{ref}}\|^2$ and squared error of the optimal surrogate parameters $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{ref}}\|^2$.

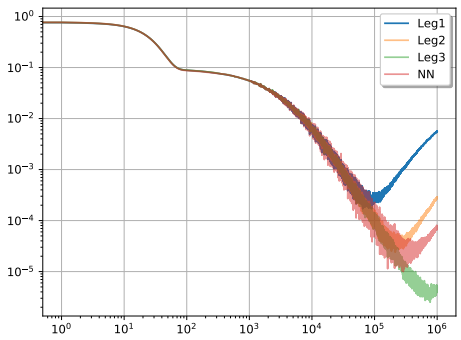


Figure 8.5: Mean squared error of control computed with surrogate and L-BFGS reference solution of the control $\|z - z_{\text{ref}}\|^2$

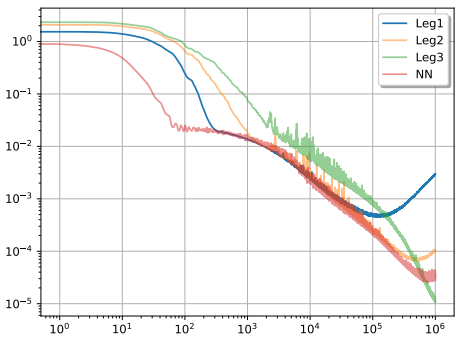


Figure 8.6: Mean squared error of surrogate and L-BFGS reference solution of the state $\mathbb{E}[\|u_{\boldsymbol{\theta}}^{\boldsymbol{y}} - u_{\text{ref}}^{\boldsymbol{y}}\|^2]$

control, see Figure 8.5. Clearly, this error is bounded from below by the approximation properties of the surrogates. In Figure 8.6 we observe the predicted rate only for the largest surrogate, the Legendre polynomial approximation of order 3.

In the same experiment we plot the model error and the difference of the surrogates to the target state $\widehat{u}$. We observe that the model error becomes smaller for surrogates with better approximation properties.

Moreover, due to the nonlinearity introduced by the activation function of the NN, our convergence theory does not apply to the problem with the NN surrogate. However, the numerical experiments are demonstrating that the NN can outperform the Legendre polynomials with a comparable number of optimization parameters.

Finally, we verify that the ADAM algorithm with adaptive choice of the penalty parameter converges to the solution of the (pERM) problem with large reference value $\bar{\lambda}$, see Theorem 8.1.5. We compute the reference solution $(z_{\text{ref}}, u_{\boldsymbol{\theta}_{\text{ref}}}^{\boldsymbol{y}})$ with $\bar{\lambda} = 100$ using the L-BFGS algorithm and plot the error of the ADAM algorithm with adaptive choice of the penalty

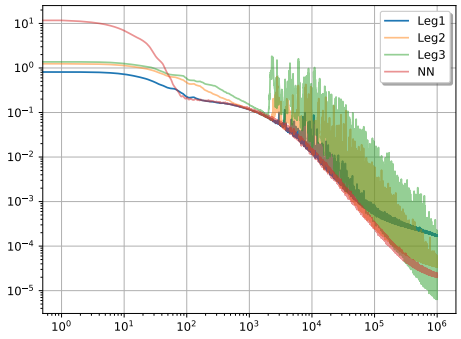


Figure 8.7: Mean squared residual $\mathbb{E}[\|A^{\boldsymbol{y}} u_{\boldsymbol{\theta}}^{\boldsymbol{y}} - z\|^2]$

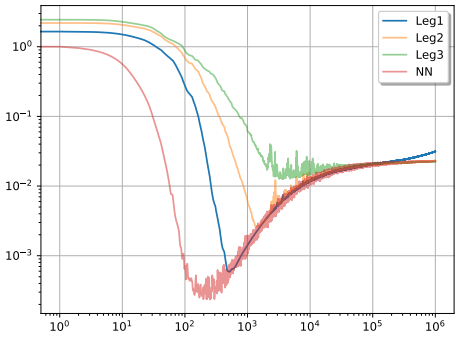


Figure 8.8: Mean squared error $\mathbb{E}[\|u_{\boldsymbol{\theta}}^{\boldsymbol{y}} - \widehat{u}\|^2]$ of the surrogate $u_{\boldsymbol{\theta}}^{\boldsymbol{y}}$ and the target state $\widehat{u}$

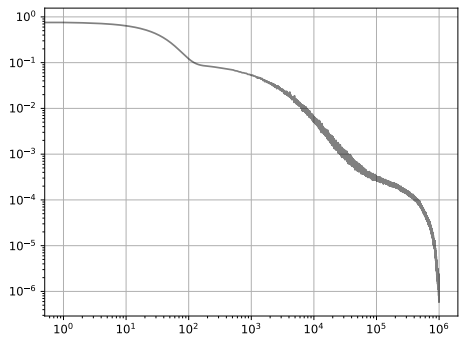


Figure 8.9: Squared error $\|z - z_{\text{ref}}\|^2$ of the control obtained with ADAM and adaptive choice of λ_k and the reference solution

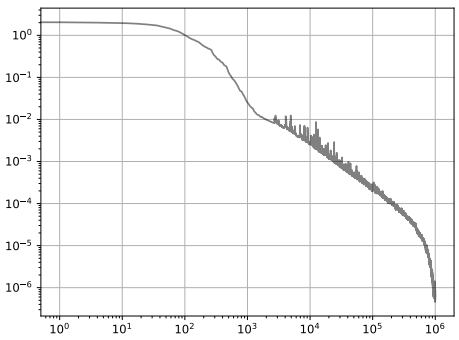


Figure 8.10: Mean squared error $\mathbb{E}[\|u_{\boldsymbol{\theta}}^{\boldsymbol{y}} - u_{\boldsymbol{\theta}_{\text{ref}}}^{\boldsymbol{y}}\|^2]$ of the surrogate obtained with ADAM and adaptive choice of λ_k and the reference surrogate

parameter λ_k against the iterations k of the ADAM algoritm. We observe convergence for both, the control and the state variable. The surrogate in this experiment is a Legendre polynomial expansion of degree 2.

8.2 Application to Bayesian inverse problems

Classical methods to solve inverse problems are based on the so-called reduced optimization approach of the (regularized) data misfit, which formulate the minimization problem as an unconstrained optimization problem using the solution operator of (8.13), see e.g., [45, 99]. This contrasts the so-called one-shot approaches, which solve the underlying model equation and the optimality conditions simultaneously. One-shot (or all-at-once) approaches are well established in the context of PDE-constrained optimization (see, e.g., [14]) and have recently been introduced to the setting of inverse problems, see [97, 98].
In the Bayesian setting, the connection between the maximum a posteriori estimation and the optimization approach to the inverse problem is well established in the finite-dimensional setting [96], as well as in the infinite-dimensional setting for certain prior classes, see [2, 28, 35, 83]. For details on the Bayesian approach to inverse problems we refer to [96, 153].
Recently, data-driven concepts have been applied to inverse problems in order to reduce the computational complexity in case of highly complex forward models and to improve models in case of limited understanding of the underlying processes [6]. For instance, neural networks have been successfully applied in the case of parametric holomorphic forward models [85] and in the case of limited knowledge of the underlying model [129, 130, 131, 151, 162]. For the training of neural networks, gradient-based methods are typically used [66]. The ensemble Kalman inversion (EKI) (see e.g., [91, 142, 143]) has been recently applied as gradient-free optimizer for the training of neural networks in [80, 109].
In this section, we transfer the idea from Section 8.1 to the setting of inverse problems using the connection between the maximum a posteriori estimation in Bayesian inverse

problems and the reduced optimization approach. More precisely, we approximate the solution of the forward problem using neural networks, i.e., the computationally intense solution of the forward problem is replaced by a neural network, which is cheap to evaluate. Since training of the neural network in advance for all possible outcomes of a quantity of interest can be challenging and requires a neural network with large expressive power, i.e., many parameters. In order to reduce the computational complexity, we propose to train the neural network simultaneously to solving the inverse problem in a one-shot framework. This approach has the potential to reduce the overall costs significantly since in the one-shot optimization the neural network is trained only for the optimal solution of the inverse problem, whereas it needs to be trained for all possible quantity of interests in the parameter space if it is trained in advance.
In this section, we mak the following contributions:

- We establish the connection between the inverse problem in a one-shot formulation and the Bayesian setting. In particular, the Bayesian viewpoint allows the incorporation of model uncertainty and provides a natural way to define the regularization parameters. In case of an exact forward model, the vanishing noise can be interpreted as a penalty method. This observation allows to establish convergence results of the solution of the the one-shot formulation to the corresponding (regularized) solution of the reduced optimization problem (with exact forward model). The numerical approximation of the forward problem is replaced by a neural network in the one-shot formulation, i.e., the neural network does not have to be trained in advance.

- We show that the EKI is an efficient method to solve the resulting optimization problem. We provide a convergence analysis in the linear setting. To enhance the performance, we modify the algorithm motivated by the continuous version of the EKI and provide numerical evidence for its superiority. Numerical experiments demonstrate the robustness of the proposed algorithm, also in the nonlinear setting.

8.2.1 Introduction to inverse problems

In this section we briefly introduce the reader to inverse problems and the notation being used. In many problems in science and engeneering, the quantity of interest can not be observed directly, but only indirectly through observations of the underlying system. In such problems, typically referred to as inverse problems, one has to rely on measurements of the system to infer information about the quantity of interest. Inverse problems arise in many areas of application, e.g., biological problems, engineering and environmental systems. The information obtained from observations of the system can substantially reduce the uncertainty in predictions of the quantity of interest, and is hence indispensable in many applications.
Mathematically an inverse problem can be described as follows: recover the unknown parameter $z \in \mathcal{Z}$ in an abstract model or system

$$M(z, u) = 0 \tag{8.13}$$

from a finite number of observation of the state $u \in \mathcal{X}$ given by

$$O(u) = y \in \mathbb{R}^{n_y}\,, \tag{8.14}$$

which might be subject to measurement noise. The parameter space $\mathcal{Z}$ and the state space $\mathcal{X}$ are typically Banach spaces and the model equation, often a PDE defined on some domain D, holds in some Banach space $\mathcal{W}$. By $O : \mathcal{X} \to \mathbb{R}^{n_y}$ we denote the observation operator, mapping the state variables u to the finite-dimensional observations $y \in \mathbb{R}^{n_y}$.
A classical approach to solve an inverse problem is to minimize the data misfit in a suitable norm

$$\min_{z,u} \ \|O(u) - y\|^2_{\Gamma_{\text{obs}}} \tag{8.15}$$

$$\text{s.t. } M(z,u) = 0\,, \tag{8.16}$$

with $\Gamma_{obs} \in \mathbb{R}^{n_y \times n_y}$ symmetric and positive definit, given that the forward model is satisfied. Oftentimes the problem (8.15) – (8.16) is ill-conditioned, hence a regularization term on the unknown parameters is introduced in order to stabilize the optimization. Introducing a regularization the optimization problem becomes

$$\min_{z,u} \|O(u) - y\|^2_{\Gamma_{\text{obs}}} + \alpha_1 R_1(z) \tag{8.17}$$

$$\text{s.t. } M(z,u) = 0\,, \tag{8.18}$$

where the regularization is denoted by $R_1 : \mathcal{Z} \to \mathbb{R}$ and the positive scalar $\alpha_1 > 0$ is usually chosen according to prior knowledge on the unknown parameter z. Here and in the following of this chapter we denote by $\|\cdot\|$ the Euclidean norm and by $\langle\cdot,\cdot\rangle$ the corresponding inner product. For a given symmetric, positive definite matrix A, the weighted norm $\|\cdot\|_A$ is defined by $\|\cdot\|_A = \|A^{-1/2}\cdot\|$ and the weighted inner product by $\langle\cdot,\cdot\rangle_A = \langle\cdot, A^{-1}\cdot\rangle$.
Assuming that the forward model $M(u,p) = 0$ is well-posed, in the sense that for each parameter $z \in \mathcal{Z}$, there exists a unique state $u \in \mathcal{X}$ such that $M(z,u) = 0$ in $\mathcal{W}$, we can introducing the solution operator $S : \mathcal{Z} \to \mathcal{X}$ defined by $M(z, S(z)) = 0$. Using the solution operator, we can reformulate the optimization problems (8.15) – (8.16) and (8.17) – (8.18) as unconstrained optimization problems

$$\min_{z\in\mathcal{Z}} \ \|O(S(z)) - y\|^2_{\Gamma_{\text{obs}}}\,, \tag{8.19}$$

and

$$\min_{z\in\mathcal{Z}} \ \|O(S(z)) - y\|^2_{\Gamma_{\text{obs}}} + \alpha_1 R_1(z)\,, \tag{8.20}$$

respectively.

8.2.2 Bayesian approach to inverse problems

Adopting the Bayesian approach to inverse problems, we view the unknown parameters z as an $\mathcal{Z}$-valued random variable with prior distribution μ_0. The noise in the observations is assumed to enter the observations additive and described by a random variable $\eta \sim \mathcal{N}(0, \Gamma_{\text{obs}})$ with $\Gamma_{\text{obs}} \in \mathbb{R}^{n_y \times n_y}$ symmetric and positiv definit, i.e.,

$$y = O(S(z)) + \eta\,, \tag{8.21}$$

Further, we assume that the noise η is stochastically independent of z. By Bayes' theorem (see, e.g., [104]), we obtain the posterior distribution

$$\mu^*(\mathrm{d}z) \propto \exp\left(-\frac{1}{2}\|O(S(z)) - y\|^2_{\Gamma_{\text{obs}}}\right)\mu_0(\mathrm{d}z)\,, \tag{8.22}$$

the conditional distribution of the unknown given the observation y.
While the solutions of (8.19) and (8.20) are point estimates of the unknown parameter z, the solution of the Bayesian inverse problem is the conditional distribution of the unknown parameters given the data, the so-called posterior distribution (8.22). Since approximations of the posterior distribution are prohibitively expensive in many applications, one often uses point estimates instead. A popular choice is the maximum a posteriori (MAP) estimate, the most likely point of the unknown parameters under the posterior distribution. Denoting by ρ_0 the Lebesgue density of the prior distribution, the MAP estimate is defined as

$$\arg\max_{z\in\mathcal{Z}} \exp\left(-\frac{1}{2}\|O(S(z)) - y\|^2_{\Gamma_{\text{obs}}}\right)\rho_0(z)\,. \tag{8.23}$$

Assuming a Gaussian prior distribution, i.e., $\mu_0 = \mathcal{N}(z_0, C)$, the MAP estimate is given by the solution of the following minimization problem

$$\min_{z\in\mathcal{Z}} \frac{1}{2}\|O(S(z)) - y\|^2_{\Gamma_{\text{obs}}} + \frac{1}{2}\|z - z_0\|^2_C\,. \tag{8.24}$$

The Gaussian prior assumption leads to a Tikhonov-type regularization in the objective function, whereas the specific form of the first term in the objective function is due to the Gaussian assumption on the noise. Further details on MAP estimates can be found, e.g., in [34, 96, 159].

8.2.3 One-shot formulation for inverse problems

While (8.19), (8.20), and (8.23) are based on the reduced formulation of the problem, in this subsection we will formulate the inverse problem in the one-shot setting. The introduced one-shot approach solves an abstract inverse problem of the form (8.13) – (8.14).
Throughout Section 8.2, we derive the methods and theoretical results under the assumption that $\mathcal{Z}, \mathcal{X}$ and $\mathcal{W}$ are finite-dimensional, i.e., we assume that the forward problem $M(z, u) = 0$ has been discretized by a suitable numerical scheme and the parameter space is finite-dimensional as well, possibly after dimension truncation. Though most of the ideas and results can be generalized to the infinite-dimensional setting, we avoid the technicalities arising from the infinite-dimensional setting and focus on the discretized problem, i.e., we denote

$$\mathcal{Z} = \mathbb{R}^{n_z}, \quad \mathcal{X} = \mathbb{R}^{n_u}, \quad \mathcal{W} = \mathbb{R}^{n_w}.$$

While we usually denote vectors by boldface symbols, we do not follow this convention in this section.
Following the one-shot ideas, the abstract problem (8.13) – (8.14) can be written as

$$F(z, u) = \begin{pmatrix} M(z, u) \\ O(u) \end{pmatrix} = \begin{pmatrix} 0 \\ y \end{pmatrix} =: \tilde{y}\,, \tag{8.25}$$

Due to the noise in the observations, we rather consider

$$y = O(u) + \eta_{\text{obs}} \tag{8.26}$$

with normally distributed noise $\eta_{\text{obs}} \sim \mathcal{N}(0, \Gamma_{\text{obs}})$, and symmetric and positive definit matrix $\Gamma_{\text{obs}} \in \mathbb{R}^{n_y \times n_y}$. Similarly, we assume that

$$0 = M(z, u) + \eta_{\text{model}}\,, \tag{8.27}$$

i.e., we assume that the model error can be described by $\eta_{\text{model}} \sim \mathcal{N}(0, \Gamma_{\text{model}})$, and symmetric and positive definit matrix $\Gamma_{\text{model}} \in \mathbb{R}^{n_w \times n_w}$. Combining (8.26) and (8.27), we obtain the problem

$$\tilde{y} = F(z,u) + \begin{pmatrix} \eta_{\text{model}} \\ \eta_{\text{obs}} \end{pmatrix}. \tag{8.28}$$

The MAP estimate can then be computed by the solution of the following minimization problem

$$\min_{z,u} \frac{1}{2}\|F(z,u) - \tilde{y}\|^2_{\Gamma} + \alpha_1 R_1(z) + \alpha_2 R_2(u), \tag{8.29}$$

where $R_1 : \mathcal{Z} \to \mathbb{R}$ and $R_2 : \mathcal{X} \to \mathbb{R}$ are regularizations of the parameter $z \in \mathcal{Z}$ and the state $u \in \mathcal{X}$, $\alpha_1, \alpha_2 > 0$ and $\Gamma = \begin{pmatrix} \Gamma_{\text{model}} & 0 \\ 0 & \Gamma_{\text{obs}} \end{pmatrix} \in \mathbb{R}^{(n_w+n_y)\times(n_w+n_y)}$.
The proposed approach does not rely on a Gaussian noise model for the forward problem, i.e., non-Gaussian models can be straightforwardly incorporated. In this case, the Bayesian viewpoint may guide the choice of the regularization parameter or function. The model error can typically estimated from experimental data or more complex models, cf.,[102, 87]. We focus here on the Gaussian setting since the one-shot approach for inverse problems is typically formulated in a least-squares fashion (particularly when neural networks are used as surrogates for the forward problem [129, 130]). The focus of this work will be on the development of a methodology, which allows to satisfy the forward problem exactly. This is achieved by the connection to the Bayesian setting and working in the vanishing noise setting.

8.2.4 Vanishing noise and penalty methods

The case of an exact forward model, i.e., when forward equation is supposed to be satisfied exactly with $M(z,u) = 0$, can be modeled in the Bayesian setting by vanishing noise. In order to illustrate this idea we consider a parameterized noise covariance model $\Gamma_{\text{model}} = \gamma\hat{\Gamma}_{\text{model}}$ for $\gamma \in \mathbb{R}_+$ and a given symmetric and positiv definit matrix $\hat{\Gamma}_{\text{model}}$. The limit for $\gamma \to 0$ corresponds to the vanishing noise setting and can be interpret as reducing the uncertainty in our model. The MAP estimate in the one-shot framework is then given by

$$\min_{u,p} \frac{1}{2}\|O(u) - y\|^2_{\Gamma_{\text{obs}}} + \frac{\lambda}{2}\|M(z,u)\|^2_{\hat{\Gamma}_{\text{model}}} + \alpha_1 R_1(z) + \alpha_2 R_2(u) \tag{8.30}$$

with $\lambda = 1/\gamma$. This form of the optimization problem reveals the connection to penalty methods, which attempt to solve constrained optimization problems such as (8.15) – (8.16) by sequentially solving unconstrained optimization problems of the form (8.30) for a sequence of monotonically increasing penalty parameters λ. We present a well-known result about the convergence of such methods, see, e.g., [10].

Proposition 8.2.1. *Let the observation operator O, the forward model M and the regularization functions R_1, R_2 be continuous and the feasible set $\{(z,u)|M(z,u) = 0\}$ be nonempty. For $k = 0, 1, \ldots$ let (z_k, u_k) denote a global minimizer of*

$$\min_{z,u} \frac{1}{2}\|O(u) - y\|^2_{\Gamma_{obs}} + \frac{\lambda_k}{2}\|M(z,u)\|^2_{\hat{\Gamma}_{\text{model}}} + \alpha_1 R_1(u) + \alpha_2 R_2(p) \tag{8.31}$$

with $(\lambda_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$ strictly monotonically increasing and $\lambda_k \to \infty$ for $k \to \infty$. Then every accumulation point of the sequence (z_k, u_k) is a global minimizer of

$$\min_{z,u} \quad \frac{1}{2}\|O(u) - y\|^2_{\Gamma_{\text{obs}}} + \alpha_1 R_1(z) + \alpha_2 R_2(u) \tag{8.32}$$

$$s.t. \quad M(z,u) = 0\,. \tag{8.33}$$

This convergence result ensures the feasibility of the estimates, i.e., physical constraints can be incorporated and exactly satisfied in the limit in the proposed one-shot approach. We note that interesting questions for future research arise when considering exact penalty terms in the objective, which correspond to different noise models in the Bayesian setting. This setting will be the starting point of the incorporation of neural networks into the problem. Instead of minimizing with respect to the state u, we will approximate the solution of the forward problem u by a neural network $u_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta}$ denote the parameters of the neural network to be learned within this framework. Thus, we obtain the corresponding minimization problem

$$\min_{z,\boldsymbol{\theta}} \ \frac{1}{2}\|F(z,u_{\boldsymbol{\theta}}) - \tilde{y}\|_\Gamma^2 + \alpha_1\mathcal{R}_1(z) + \alpha_2\mathcal{R}_2(u_{\boldsymbol{\theta}},\boldsymbol{\theta})\,, \tag{8.34}$$

where $u_{\boldsymbol{\theta}}$ denotes the state approximated by the neural network.

Neural networks in inverse problems

Neural networks experienced a tremendous success in applications related to inverse problems, leading to a rapid increase in the number of publications in this area of research. Thus, we can only provide a excerpt of this fast growing research field, and focus on the most related work.

In [149] the authors show holomorphy of the data-to-QoI map $y \mapsto \mathbb{E}^{\mu^*}[\mathrm{QoI}]$, which relates observation data to the posterior expectation of an unknown quantity of interest (QoI), for additive, centered Gaussian observation noise in Bayesian inverse problems. Using the fact that holomorphy implies fast convergence of Taylor expansions, the authors derived an exponential expression rate bound in terms of the overall network size.

Our approach differs from the ideas above as we do not approximate the data-to-QoI map, but instead emulate the state u itself by a DNN. Hence, in our method the input of the neural network is a point in the spatial domain of the state, $x \in D$. The output of the neural network is an approximation of the state at this point, $u_{\boldsymbol{\theta}}(x) \in \mathbb{R}$, i.e., $N_L = 1$. Recall that a DNN is defined in (8.6). By a slight abuse of notation we denote by $u_{\boldsymbol{\theta}} \in \mathcal{X} = \mathbb{R}^{n_u}$ also a vector containing evaluations of the neural network at the n_u-many grid points of the state. In combination with a one-shot approach for the training of the neural network parameters, our method is closer related to the physics-informed neural networks (PINNs) in [129, 130]. In [129, 130] the authors consider PDEs of the form

$$f(t,x) := u_t + N(u,\lambda) = 0, \quad t \in [0,T],\, x \in D\,,$$

where N is a nonlinear differential operator parameterized by λ. The authors replace u by a neural network $u_{\boldsymbol{\theta}}$ and use automatic differentiation to construct the function $f_{\boldsymbol{\theta}}(t,x)$. The neural network parameters are then obtained by minimizing the mean squarred error $\mathrm{MSE} = \mathrm{MSE}_u + \mathrm{MSE}_f$, where

$$\mathrm{MSE}_u := \frac{1}{N_u}\sum_{i=1}^{N_u} |u_{\boldsymbol{\theta}}(t_u^i,x_u^i) - u^i|^2\,, \quad \mathrm{MSE}_f := \frac{1}{N_f}\sum_{i=1}^{N_f} |f_{\boldsymbol{\theta}}(t_f^i,x_f^i)|^2\,,$$

and $\{t_u^i, x_u^i, u^i\}_{i=1}^{N_u}$ denote the training data and $\{t_f^i, x_f^i\}_{i=1}^{N_f}$ are collocation points of $f_{\boldsymbol{\theta}}(t,x)$. For the minimization a L-BFGS method is used. The parameters λ of the differential

operator turn into parameters of the neural network $f_{\boldsymbol{\theta}}$ and can be learned by minimizing the MSE. Based on [129, 130] there has been a tremendous increase in research and applications of PINNs. For instance in [162], the authors consider so called Bayesian neural networks (BNNs), where the neural network parameters are updated according to Bayes' theorem. Hereby the initial distribution on the network parameters serves as prior distribution. The likelihood requires the PDE solution, which is obtained by concatenating the Bayesian neural network with a physics-informed neural network, which they call Bayesian physics-informed neural networks (B-PINNs). For the estimation of the posterior distributions they use the Hamiltonian Monte Carlo method and variational inference. In contrast to the PINNs, the Bayesian framework allows to quantify the aleatoric uncertainty associated with noisy data. In addition the numerical experiments in [162] indicate that B-PINNs beat PINNs in case of large noise levels on the observations.
In contrast to that, our proposed method is based on the MAP estimate and remains exact in the small noise limit. We propose a derivative-free optimization method, the EKI, which shows promising results without requiring derivatives with respect to the weights and design parameters.

8.2.5 Ensemble Kalman inversion

The ensemble Kalman inversion (EKI) generalizes the well-known ensemble Kalman Filter (EnKF) introduced by Evensen and coworker in the data assimilation context [48] to the inverse setting, see [91] for more details. Since the Kalman filter involves a Gaussian approximation of the underlying posterior distribution, we focus on an iterative version based on tempering in order to reduce the linearization error. Recall the posterior distribution μ^* given by

$$\mu^*(\mathrm{d}v) \propto \exp\Big(-\frac{1}{2}\|G(v)-y\|_\Gamma^2\Big)\mu_0(\mathrm{d}v)\,.$$

for an abstract inverse problem

$$y = G(v) + \eta,$$

where G maps the unknowns $v \in \mathbb{R}^{n_v}$ to the observations $y \in \mathbb{R}^{n_y}$ with $\eta \sim \mathcal{N}(0,\Gamma)$, and symmetric and positiv definit matrix $\Gamma \in \mathbb{R}^{n_y \times n_y}$. We define the intermediate measures

$$\mu_n(\mathrm{d}v) \propto \exp\Big(-\frac{1}{2}nh\|G(v)-y\|_\Gamma^2\Big)\mu_0(\mathrm{d}v) \quad n = 0,\dots,N \tag{8.35}$$

by scaling the data misfit by the step size $h = N^{-1}$, $N \in \mathbb{N}$. The idea is to apply the EnKF to the resulting artificial time dynamical system in order to evolve the prior distribution μ_0 into the posterior distribution $\mu_N = \mu^*$ by this sequence of intermediate measures. We account for the repeated use of the observations by amplifying the noise variance by $N = 1/h$ in each step. The intermediate measures μ_n are then approximated using the EKI with an ensemble of J particles $\{v_0^{(j)}\}_{j=1}^J$ with $J \in \mathbb{N}$

$$\mu_n \simeq \frac{1}{J}\sum_{j=1}^{J} \delta_{v_n^{(j)}} \tag{8.36}$$

with δ_v denoting the Dirac measure centered on $v_n^{(j)}$. The particles are transformed in each iteration by the application of the Kalman update formulas to the empirical mean $\bar{v}_n = \frac{1}{J}\sum_{j=1}^J v_n^{(j)}$ and empirical covariance $C(v_n) = \frac{1}{J-1}\sum_{j=1}^J (v_n^{(j)} - \bar{v}_n)\otimes(v_n^{(j)} - \bar{v}_n)$, i.e.,

$$\bar{v}_{n+1} = \bar{v}_n + K_n(y - G(\bar{v}_n))\,, \qquad C(v_{n+1}) = C(v_n) - K_n C^{y,v}(v_n)\,, \tag{8.37}$$

where $K_n = C^{v,y}(v_n)(C^{y,y}(v_n) + \frac{1}{h}\Gamma)^{-1}$ denotes the Kalman gain, and for $v = \{v^{(j)}\}_{j=1}^J$, the operators C^{yy} and C^{vy} given by

$$C^{y,y}(v) = \frac{1}{J}\sum_{j=1}^{J}\big(G(v^{(j)}) - \bar{G}\big) \otimes \big(G(v^{(j)}) - \bar{G}\big), \tag{8.38}$$

$$C^{v,y}(v) = \frac{1}{J}\sum_{j=1}^{J}\big(v^{(j)} - \bar{v}\big) \otimes \big(G(v^{(j)}) - \bar{G}\big), \tag{8.39}$$

$$C^{y,v}(v) = \frac{1}{J}\sum_{j=1}^{J}\big(G(v^{(j)}) - \bar{G}\big) \otimes \big(v^{(j)} - \bar{v}\big), \tag{8.40}$$

$$\bar{G} = \frac{1}{J}\sum_{j=1}^{J} G(v^{(j)}) \tag{8.41}$$

are the empirical covariances and empirical mean in the observation space. Since this update does not uniquely define the transformation of each particle $v_n^{(j)}$ to the next iteration $v_{n+1}^{(j)}$, the specific choice of the transformation leads to different variants of the EKI. We will focus here on the generalization of the EnKF as introduced by [91] resulting in a mapping of the particles of the form

$$v_{n+1}^{(j)} = v_n^{(j)} + C^{v,y}(v_n)(C^{y,y}(v_n) + h^{-1}\Gamma)^{-1}\big(y_{n+1}^{(j)} - G(v_n^{(j)})\big), \quad j = 1, \cdots, J, \tag{8.42}$$

where

$$y_{n+1}^{(j)} = y + \xi_{n+1}^{(j)}$$

The $\xi_{n+1}^{(j)}$ are i.i.d. random variables distributed according to $\mathcal{N}(0, h^{-1}\Sigma)$ with $\Sigma = \Gamma$ corresponding to the case of perturbed observations and $\Sigma = 0$ to the unperturbed observations.
The motivation via the sequence of intermediate measures and the resulting artificial time allows to derive the continuous time limit of the iteration, which has been extensively studied in [12, 142, 143] to build analysis of the EKI in the linear setting. This limit arises by taking the parameter h in (8.42) to zero resulting in

$$\frac{\mathrm{d}v^{(j)}}{\mathrm{d}t} = C^{v,y}(v)\Gamma^{-1}(y - G(v^{(j)})) + C^{v,y}(v^{(j)})\Gamma^{-1}\sqrt{\Sigma}\,\frac{dW^{(j)}}{dt}. \tag{8.43}$$

As shown in [46], the EKI does not in general converge to the true posterior distribution. Therefore, the analysis presented in [12, 142, 143] views the EKI as a derivative-free optimizer of the data misfit, which is also the viewpoint we adopt here.

Ensemble Kalman inversion for neural network based one-shot optimization

By approximating the state of the underlying PDE by a neural network, we seek to optimize with respect to the unknown parameter z and the parameters of the neural network $\boldsymbol{\theta}$. The idea is based on defining the function $H(v) := H(u, \boldsymbol{\theta}) = F(z, u_{\boldsymbol{\theta}})$, where $u_{\boldsymbol{\theta}}$ denotes the state approximated by the neural network and $v = (z, \boldsymbol{\theta})^\top$. This leads to the empirical summary statistics

$$\overline{(z, \boldsymbol{\theta})}_n = \frac{1}{J}\sum_{j=1}^{J}(z_n^{(j)}, \boldsymbol{\theta}_n^{(j)}), \quad \bar{H}_n = \frac{1}{J}\sum_{j=1}^{J} H(z_n^{(j)}, \boldsymbol{\theta}_n^{(j)}),$$

$$C_n^{z\boldsymbol{\theta},y} = \frac{1}{J}\sum_{j=1}^{J}\big((z_n^{(j)},\boldsymbol{\theta}_n^{(j)})^\top - \overline{(z,\boldsymbol{\theta})}_n^\top\big)\otimes\big(H\big(z_n^{(j)},\boldsymbol{\theta}_n^{(j)}\big)-\bar{H}_n\big),$$

$$C_n^{y,y} = \frac{1}{J}\sum_{j=1}^{J}\big(H(z_n^{(j)},\boldsymbol{\theta}_n^{(j)})-\bar{H}_n\big)\otimes\big(H(z_n^{(j)},\boldsymbol{\theta}_n^{(j)})-\bar{H}_n\big),$$

and the EKI update

$$(z_{n+1}^{(j)},\boldsymbol{\theta}_{n+1}^{(j)})^\top = (z_n^{(j)},\boldsymbol{\theta}_n^{(j)})^\top + C_n^{z\boldsymbol{\theta},y}\big(C_n^{y,y}+h^{-1}\Gamma\big)^{-1}\big(\tilde{y}_{n+1}^{(j)} - H(z_n^{(j)},\boldsymbol{\theta}_n^{(j)})\big), \tag{8.44}$$

where the perturbed observation are computed as before

$$\tilde{y}_{n+1}^{(j)} = \tilde{y} + \xi_{n+1}^{(j)}, \quad \xi_{n+1}^{(j)} \sim \mathcal{N}(0,h^{-1}\Sigma), \tag{8.45}$$

with

$$\tilde{y} = \begin{pmatrix}0\\ y\end{pmatrix}, \quad \Gamma := \begin{pmatrix}\Gamma_{\text{model}} & 0\\ 0 & \Gamma_{\text{obs}}\end{pmatrix}.$$

Figure 8.11 illustrates the basic idea of the application of the EKI to solve the neural network based one-shot formulation.

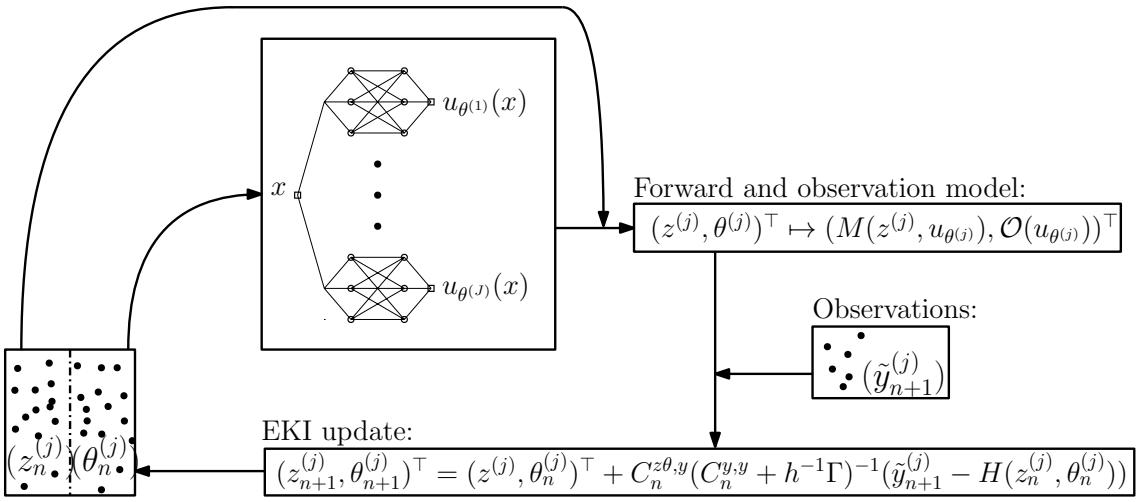


Figure 8.11: Description of the EKI applied to solve the neural network based one-shot formulation.

The EKI (8.44) will be used as a derivative-free optimizer of the data misfit $\|F(z,u_{\boldsymbol{\theta}})-\tilde{y}\|_\Gamma^2$. The analysis presented in [12, 142, 143] shows that the EKI in its continuous form is able to recover the data with a finite number of particles in the limit $t\to\infty$ under suitable assumptions on the forward problem and the set of particles. In particular, the analysis assumes a linear forward problem. Extensions to the nonlinear setting can be found, e.g., in [18, 160]. The limit $t\to\infty$ corresponds to the noise-free setting, as the inverse noise covariance scales with $n/N = nh$ in (8.35). To explore the scaling of the noise and to discuss regularization techniques, we illustrate the ideas in the following for a linear Gaussian setting, i.e., we assume that the forward response operator is linear $H(v)=Av$ with $A\in\mathcal{L}(\mathcal{Z}\times\Theta,\mathbb{R}^{n_w+n_y})$ and $\mu_0=\mathcal{N}(v_0,C_0)$. Considering the large ensemble size limit $J\to\infty$, the mean m and covariance C satisfy the equations

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} = -C(t)A^\top\Gamma^{-1}(Am(t)-y) \tag{8.46}$$

$$\frac{\mathrm{d}C}{\mathrm{d}t} = -C(t)A^\top \Gamma^{-1} A C(t) \tag{8.47}$$

for $\Sigma = \Gamma$ in (8.42). Considering the dynamics of the inverse covariance, it is straightforward to show that the solution is given by

$$C^{-1}(t) = C_0^{-1} + A^\top \Gamma^{-1} A t\,, \tag{8.48}$$

see, e.g., [55] and the references therein for details. Note that $C(1)$ corresponds to the posterior covariance and that $C(t) \to 0$ for $t \to \infty$. Furthermore, the mean is given by

$$m(t) = (C_0^{-1} + A^\top \Gamma^{-1} A t)^{-1}(A^\top \Gamma^{-1} y t + C_0^{-1} v_0)\,, \tag{8.49}$$

in particular the mean minimizes the data misfit in the limit $t \to \infty$. The application of the EKI in the inverse setting therefore often requires additional techniques such as adaptive stopping [143] or additional regularization [19] to overcome the ill-posedness of the minimization problem. To control the regularization of the data misfit and neural network individually, we consider the following system

$$F(z, u_{\boldsymbol{\theta}}) + \begin{pmatrix} \eta_{\text{model}} \\ \eta_{\text{obs}} \end{pmatrix} = \tilde{y} \tag{8.50}$$

$$\begin{pmatrix} z \\ \boldsymbol{\theta} \end{pmatrix} + \begin{pmatrix} \eta_{param} \\ \eta_{NN} \end{pmatrix} = 0 \tag{8.51}$$

with $\eta_{\text{model}} \sim \mathcal{N}(0, 1/\lambda\, \hat{\Gamma}_{\text{model}})$, $\eta_{\text{obs}} \sim \mathcal{N}(0, \Gamma_{\text{obs}})$, $z \sim \mathcal{N}(z_0, 1/\alpha_1\, C)$, $\boldsymbol{\theta} \sim \mathcal{N}(0, 1/\alpha_2\, I)$. The loss function corresponding to the augmented system (8.50) – (8.50) is given by

$$\frac{1}{2}\|O(u_{\boldsymbol{\theta}}) - y\|^2_{\Gamma_{\text{obs}}} + \frac{\lambda}{2}\|M(z, u_{\boldsymbol{\theta}})\|^2_{\hat{\Gamma}_{\text{model}}} + \frac{\alpha_1}{2}\|z - z_0\|^2_C + \frac{\alpha_2}{2}\|\boldsymbol{\theta}\|^2\,. \tag{8.52}$$

Assuming that the resulting forward operator

$$G(z, \boldsymbol{\theta}) = \begin{pmatrix} F(z, u_{\boldsymbol{\theta}}) \\ z \\ \boldsymbol{\theta} \end{pmatrix} \tag{8.53}$$

is linear, the EKI will converge to the minimum of the regularized loss function (8.52), cf., [142]. To ensure the feasibility of the EKI estimate (with respect to the underlying forward problem), we propose the following algorithm using the ideas discussed in Section 8.2.4.

Theorem 8.2.2. *Assume that the forward operator $G : \mathcal{Z} \times \Theta \to \mathbb{R}^{n_G}$, $n_G := n_w + n_y + n_z + n_{\boldsymbol{\theta}}$,*

$$G(u, \boldsymbol{\theta}) = \begin{pmatrix} F(u, p_{\boldsymbol{\theta}}) \\ u \\ \boldsymbol{\theta} \end{pmatrix}$$

is linear, i.e., $F(z, u_{\boldsymbol{\theta}}) = A(z, \boldsymbol{\theta})^\top$ with $A \in \mathcal{L}(\mathcal{Z} \times \Theta, \mathbb{R}^{n_w + n_y})$. Let $(\lambda_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$ be strictly monotonically increasing and $\lambda_k \to \infty$ for $k \to \infty$. Further, assume that the initial ensemble members are chosen so that $\operatorname{span}\{(z^{(j)}(0), \boldsymbol{\theta}^{(j)}(0))^\top, j = 1, \dots, J\} = \mathcal{Z} \times \Theta$. Then, Algorithm 8 generates a sequence of estimates $(\bar{z}_k, \bar{\boldsymbol{\theta}}_k)_{k\in\mathbb{N}}$, where $\bar{z}_k, \bar{\boldsymbol{\theta}}_k$ minimizes the loss function for the augmented system given by

$$\frac{1}{2}\|O(u_{\boldsymbol{\theta}}) - y\|^2_{\Gamma_{\text{obs}}} + \frac{\lambda_k}{2}\|M(z, u_{\boldsymbol{\theta}})\|^2_{\hat{\Gamma}_{\text{model}}} + \frac{\alpha_1}{2}\|z - z_0\|^2_C + \frac{\alpha_2}{2}\|\boldsymbol{\theta}\|^2$$

Algorithm 8 Penalty ensemble Kalman inversion for neural network based one-shot inversion

Require: initial ensemble $v_0^{(j)} = (z_0^{(j)}, \boldsymbol{\theta}_0^{(j)})^\top \in \mathcal{Z} \times \Theta, j = 1, \dots J,\ \lambda_0$.

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

2: Compute an approximation of the minimizer $(z_k, \boldsymbol{\theta}_k)^\top$ of

$$\min_{z,\boldsymbol{\theta}} \ \frac{1}{2}\|O(u_{\boldsymbol{\theta}}) - y\|^2_{\Gamma_{\text{obs}}} + \frac{\lambda_k}{2}\|M(z, u_{\boldsymbol{\theta}})\|^2_{\hat{\Gamma}_{\text{model}}} + \frac{\alpha_1}{2}\|z - z_0\|^2_C + \frac{\alpha_2}{2}\|\boldsymbol{\theta}\|^2 .$$

by solving

$$\frac{\mathrm{d}v^{(j)}}{\mathrm{d}t} \quad = \quad C^{vy}(v)\Gamma^{-1}(\hat{y} - G(v^{(j)})) + C^{vy}(v^{(j)})\Gamma^{-1}\sqrt{\Sigma}\,\frac{dW^{(j)}}{dt}$$

with $\hat{y} = (0, y, 0, 0)^\top$, $v^{(j)}(0) = v_0^{(j)}$ for the system (8.53) and $\Gamma = \operatorname{diag}\left(C, I, \frac{1}{\alpha_1}C, \frac{1}{\alpha_2}I\right)$.

3: Set $v_k = (z_k, \boldsymbol{\theta}_k)^\top = \lim_{T\to\infty} \bar{v}(T)$.

4: Increase λ_k.

5: Draw J ensemble members $v_0^{(j)}$ from $\mathcal{N}(v_k, \begin{pmatrix} C & 0 \\ 0 & I \end{pmatrix})$.

6: **end for**

with given $\alpha_1, \alpha_2 > 0$. Furthermore, every accumulation point of $(\bar{z}_k, \bar{\boldsymbol{\theta}}_k)_{k\in\mathbb{N}}$ is the (unique, global) minimizer of

$$\begin{aligned} \min_{z,\boldsymbol{\theta}} \quad & \frac{1}{2}\|O(u_{\boldsymbol{\theta}}) - y\|^2_{\Gamma_{\text{obs}}} + \frac{\alpha_1}{2}\|z - z_0\|^2_C + \frac{\alpha_2}{2}\|\boldsymbol{\theta}\|^2 \\ s.t. \quad & M(z, u_{\boldsymbol{\theta}}) = 0 \end{aligned}$$

Proof. Under the assumption of a linear forward model, the penalty function

$$\frac{1}{2}\|O(u_{\boldsymbol{\theta}}) - y\|^2_{\Gamma_{\text{obs}}} + \frac{\lambda_k}{2}\|M(z, u_{\boldsymbol{\theta}})\|^2_{\hat{\Gamma}_{\text{model}}} + \frac{\alpha_1}{2}\|z - z_0\|^2_C + \frac{\alpha_2}{2}\|\boldsymbol{\theta}\|^2$$

is strictly convex for all $k \in \mathbb{N}$, i.e., there exists a unique minimizer of the penalized problem. Choosing the initial ensemble such that $\operatorname{span}\{(z^{(j)}(0), \boldsymbol{\theta}^{(j)}(0))^\top, j = 1, \dots, J\} = \mathcal{X} \times \Theta$ ensures the convergence of the EKI estimate to the global minimizer, see [19, Theorem 3.13] and [142, Theorem 4]. The convergence of Algorithm 8 to the minimzer of the constrained problem then follows from Proposition 8.2.1. □

Remark 8.2.3. *The convergence result Theorem 8.2.2 is based on an assumption on the size of the ensemble, which is needed to ensure the convergence to the (global) minimizer of the loss function in each iteration. This is due to the well-known subspace property of the EKI, i.e., the EKI estimate will lie in the span of the initial ensemble when using the EKI in its variant. In case of a large or possibly infinite-dimensional parameter / state space, the assumption on the size of the ensemble can usually not be satisfied in practice. Techniques such as variance inflation, localization and adaptive ensemble choice are able to overcome the subspace property and thus might lead to much more efficient algorithms from a computational point of view.*
Furthermore, we stress the fact that the convergence result presented above is based on the linearity of the forward and observation operator. Thus the assumption is not fulfilled when considering a neural network with nonlinear activation function as approximation of

the solution of the forward problem. However, numerical experiments (see Section 8.2.6) show promising results even in the nonlinear setting. The generalization of the theoretical results, such as Theorem 8.2.2, is subject for future work.

Algorithm 8 requires the solutions of a sequence of optimization problems, i.e., for each λ the EKI is used to approximate the solution of the corresponding minimization problem. To avoid the repeated application of EKI, we propose a modified version of the algorithm in Algorithm 9. The idea of Algorithm 9 is to solve a single optimization problem with increasing regularization parameter λ. This can straightforwardly be incorporated in the continuous version of EKI by solving an additional differential equation for λ with nondecreasing right-hand side. The computational effort of Algorithm 9 is thus reduced compared to Algorithm 8 and numerical experiments suggest a comparable performance in terms of accuracy. The theoretical analysis of the convergence behavior will be subject to future work.

Algorithm 9 Simultaneous penalty ensemble Kalman inversion for neural network based one-shot inversion

Require: initial ensemble $v_0^{(j)} = (z_0^{(j)}, \boldsymbol{\theta}_0^{(j)})^\top \in \mathcal{Z} \times \Theta, j = 1, \dots, J,\ \lambda_0 \in \mathbb{R}_{\geqslant 0},\ f : \mathbb{R}_{\geqslant 0} \to \mathbb{R}_+$.

1: Compute an approximation of the minimizer of

$$\begin{aligned} \min_{z,\boldsymbol{\theta}} \quad & \frac{1}{2}\|O(u_{\boldsymbol{\theta}}) - y\|^2_{\Gamma_{\text{obs}}} + \frac{\alpha_1}{2}\|z - z_0\|^2_C + \frac{\alpha_2}{2}\|\boldsymbol{\theta}\|^2 \\ \text{s.t.} \quad & M(z, u_{\boldsymbol{\theta}}) = 0 \end{aligned}$$

by solving the following system

$$\begin{aligned} \frac{\mathrm{d}v^{(j)}}{\mathrm{d}t} &= C^{vy}(v)\Gamma^{-1}(\hat{y} - G(v^{(j)})) + C^{vy}(v^{(j)})\Gamma^{-1}\sqrt{\Sigma}\,\frac{dW^{(j)}}{dt} \\ \frac{\mathrm{d}\lambda}{\mathrm{d}t} &= f(\lambda) \end{aligned}$$

with $\hat{y} = (0, y, 0, 0)^\top$, $v^{(j)}(0) = v_0^{(j)}$ for the system (8.53), $\lambda(0) = \lambda_0$ and $\Gamma = \operatorname{diag}\left(C, I, \frac{1}{\alpha_1}C, \frac{1}{\alpha_2}I\right)$.

8.2.6 Numerical experiments

The following numerical experiments illustrate the one-shot inversion for different inverse problems. The first example is a one-dimensional problem, for which we compare the reduced optimization approach, quasi-Newton method (see, e.g., [49]) for the one-shot inversion, quasi-Newton method for the neural network based one-shot inversion (Algorithm 8), EKI for the one-shot inversion and EKI for the neural networks based one-shot inversion (Algorithm 9) in the linear setting. Moreover, we numerically explore the convergence behavior of the EKI for the neural networks based one-shot inversion Algorithm 9 also for a nonlinear forward model. The final experiment is concerned with the extension of the linear model to the two-dimensional problem to investigate the potential of the EKI for neural network based inversion in the higher-dimensional setting.

One-dimensional example

We consider the problem of recovering the unknown data $u^\dagger$ from noisy observations

$$y = \mathcal{O}(u^\dagger) + \eta^\dagger,$$

where $u^\dagger = \mathcal{A}^{-1}(z^\dagger)$ is the solution of the one-dimensional elliptic equation

$$\begin{aligned} -\frac{\mathrm{d}^2 u}{\mathrm{d}x^2} + u &= z^\dagger \quad \text{in } D := (0,\pi), \\ u &= 0 \quad \text{on } \partial D, \end{aligned} \tag{8.54}$$

with operator $\mathcal{O}$ observing the dynamical system at $n_y = 2^3 - 1$ equispaced observation points $x_i = \frac{i}{2^4} \cdot \pi$, $i = 1, \ldots, n_y$.
We approximate the forward-problem (8.54) numerically on a uniform mesh with mesh-width $h = 2^{-6}$ by a finite element method with continuous, piecewise linear ansatz functions. The approximated solution operator will be denoted by $S \in \mathbb{R}^{n_z \times n_z}$, with $n_z = 1/h$. The unknown parameter z is assumed to be Gaussian, i.e., $z \sim \mathcal{N}(0, C_0)$, with (discretized) covariance operator $C_0 = \beta(-\frac{\mathrm{d}^2}{\mathrm{d}x^2})^{-\nu}$ for $\beta = 5$, $\nu = 1.5$. For the inverse problem we assume a observational noise covariance $\Gamma_{\text{obs}} = 0.1 \cdot I_{n_y}$, a model error covariance $\hat{\Gamma}_{\text{model}} = 100 \cdot I_{n_z}$ and we choose the regularization parameter $\alpha_1 = 0.002$, while we turn off the regularization on u, i.e., we set $\alpha_2 = 0$. Further, we choose a feed-forward DNN with $L = 3$ layers, where we set $N_1 = N_2 = 10$ size of the hidden layers and $N_0 = N_L = 1$ size of the input and output layer. As activation function we choose the sigmoid function $\varrho(x) = \frac{1}{1+e^{-x}}$. The EKI method is based on the deterministic formulation represented through the coupled ODE system

$$\frac{\mathrm{d}v^{(j)}}{dt} = C^{vy}(v)\Gamma^{-1}(y - G(v^{(j)})), \tag{8.55}$$

which will be solved with the `MATLAB` function `ode45` up to time $T = 10^{10}$. The ensemble of particles $(z^{(j)})$, $(z^{(j)}, u^{(j)})$, and $(z^{(j)}, \boldsymbol{\theta}^{(j)})$ respectively will be initialized by $J = 150$ particles as i.i.d. samples, where the parameters $z_0^{(j)}$ are drawn from the prior distribution $\mathcal{N}(0, C_0)$, the states $u_0^{(j)}$ are drawn from $\mathcal{N}(0, 5\, I_{n_u})$, and the weights of the neural network are drawn from $\mathcal{N}(0, I_{n_{\boldsymbol{\theta}}})$, which are all independent from each other.
We compare the results to a classical gradient-based method, namely the quasi-Newton method with BFGS updates, as implemented by `MATLAB`.
We summarize the methods in the following and introduce abbreviations:

1. Reduced formulation: explicit solution (redTik).

2. One-shot formulation: we compare the performance of the EKI with Algorithm 8 (osEKI_1), the EKI with Algorithm 9 (osEKI_2) and the quasi-Newton method with Algorithm 8 (osQN_1).

3. Neural network based one-shot formulation: we compare the performance of the EKI with Algorithm 9 (nnosEKI_2) and the quasi-Newton method with Algorithm 8 (nnosQN_1).

Figure 8.12 shows the increasing sequence of the penalty parameter λ used for Algorithm 8 and the quasi-Newton method and Algorithm 9 (over time).

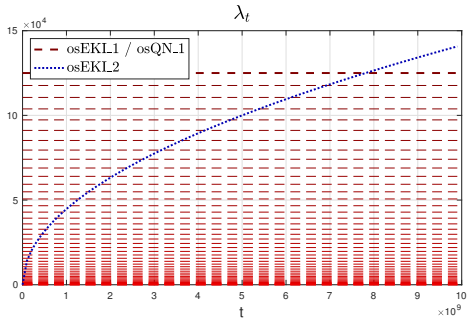


Figure 8.12: Scaling parameter λ depending on time for Algorithm 8, $\lambda_k = k^3$ for $k = 1, 2, \ldots, 50$, and Algorithm 9, $\mathrm{d}\lambda/\mathrm{d}t = 1/\lambda$.

One-shot inversion

In order to illustrate the convergence result of the EKI and to numerically investigate the performance of Algorithm 9, we start the discussion by a comparison of the one-shot inversion based on the FEM approximation of the forward problem in the 1-dimensional example.

Figure 8.13 shows the difference of the estimates given by EKI with Algorithm 8 (osEKI_1), the EKI with Algorithm 9 (osEKI_2) and the quasi-Newton method with Algorithm 8 (osQN_1) compared to the Tikhonov solution and the truth (on the left-hand side) and in the observation space (on the right-hand side). All three methods lead to an excellent approximation of the Tikhonov solution. Due to the linearity of the forward problem, the quasi-Newton method as well as the EKI with Algorithm 8 are expected to converge to the regularized solution. The EKI with Algorithm 9 demonstrates a similar performance while reducing the compuational effort significantly compared to Algorithm 8.

The comparison of the data misfit and the residual of the forward problem shown in Figure 8.14 reveals a good performance of the EKI (for both algorithms) with feasibility of the estimates (with respect to the forward problem) in the range of 10^{-10}.

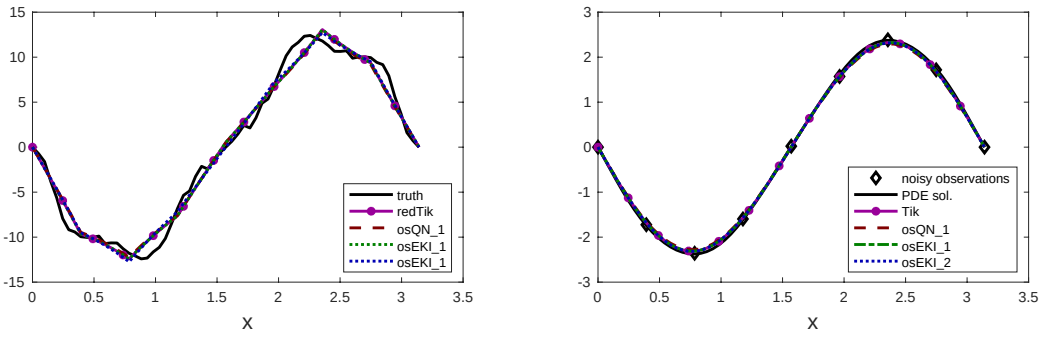


Figure 8.13: Comparison of parameter estimation given by EKI with Algorithm 8 (osEKI_1), the EKI with Algorithm 9 (osEKI_2) and the quasi-Newton method with Algorithm 8 (osQN_1) compared to the Tikhonov solution and the truth (on the left-hand side) and in the observation space (on the right-hand side).

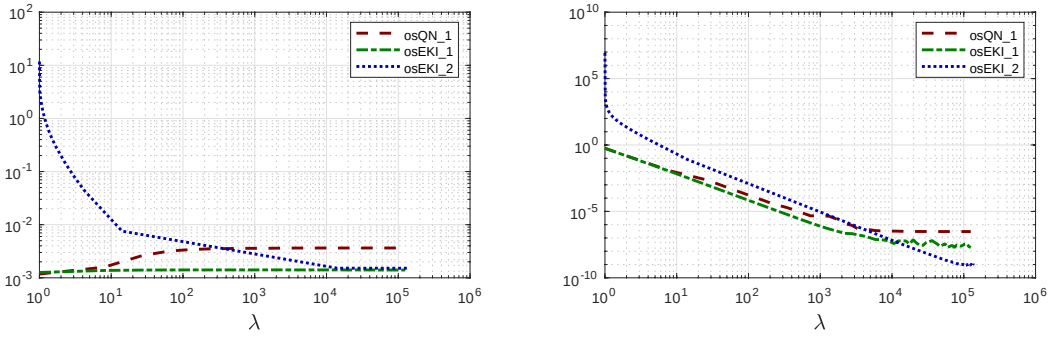


Figure 8.14: Comparison of the data misfit given by EKI with Algorithm 8 (osEKI_1), the EKI with Algorithm 9 (osEKI_2) and the quasi-Newton method with Algorithm 8 1 (osQN_1) (on the left-hand side) and residual of the forward problem (on the right-hand side), both with respect to λ.

One-shot method with neural network approximation

We next replace the solution of the forward problem by a neural network in the one-shot setting. Due to the excellent performance of Algorithm 9 in the previous experiment, we focus in the following on this approach for the neural network based one-shot inversion.

The EKI for the neural network based one-shot inversion leads to a good approximation of the regularized solution (cf., Figure 8.15), whereas the performance of the quasi-Newton approach is slightly worse, which might be due to the nonlinearity introduced into the problem by the neural network approximation.

The comparison of the data misfit and residual of the forward problem reveals an excellent convergence behaviour of the EKI for the neural network based one-shot optimization, whereas the quasi-Newton method does not converge to a feasible estimate, cf., Figure 8.16.

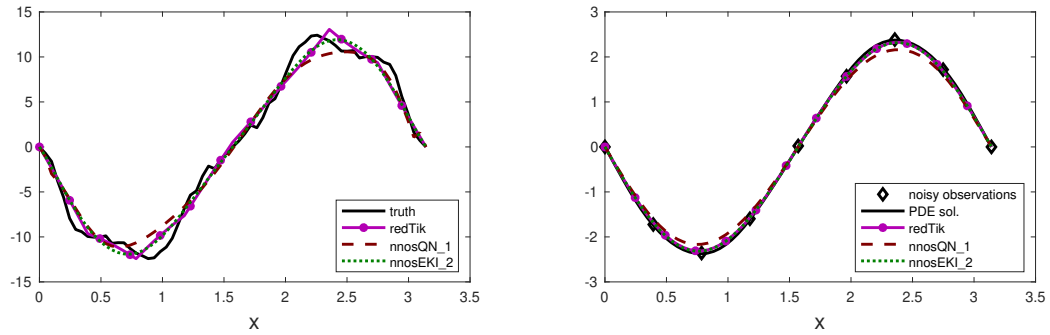


Figure 8.15: Comparison of parameter estimation given by the EKI with Algorithm 9 (nnosEKI_2) and the quasi-Newton method with Algorithm 8 (nnosQN_1) for the neural network based one-shot inversion compared to the Tikhonov solution and the truth (on the left-hand side) and in the observation space (on the right-hand side).

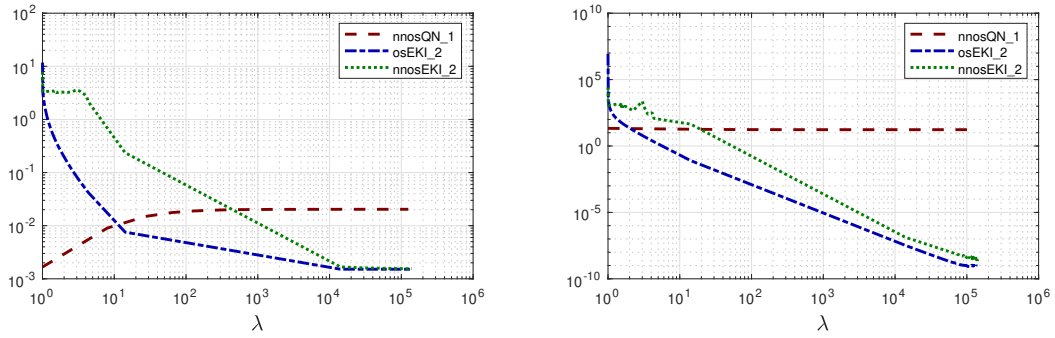


Figure 8.16: Comparison of the data misfit given by the EKI with Algorithm 9 (nnosEKI_2) and the quasi-Newton method with Algorithm 8 (nnosQN_1) for the neural network based one-shot inversion compared to EKI with Algorithm 9 (osEKI_2) from the previous experiment (on the left-hand side) and residual of the forward problem (on the right-hand side), both with respect to λ.

Nonlinear forward model

We consider in the following a nonlinear forward model of the form

$$\begin{aligned} -\nabla \cdot (\exp(z^\dagger) \cdot \nabla u) &= 10 \quad \text{in } D := (0, \pi), \\ u &= 0 \quad \text{on } \partial D\,. \end{aligned} \tag{8.56}$$

Note that the mapping from the unknown parameter function to the state is nonlinear. We use the same discretization as in the linear problem. The unknown parameter $z^\dagger$ is assumed to be Gaussian with zero mean and $C_0 = \beta(-\frac{\mathrm{d}^2}{\mathrm{d}x^2})^{-\nu}$ where we choose $\beta = 1$, $\nu = 2$. Furthermore, we set $\Gamma_{\text{obs}} = 0.0001 \cdot I_{n_y}$, $\hat{\Gamma}_{\text{model}} = 10 \cdot I_{n_u}$, $\alpha_1 = 2$ and $\alpha_2 = 0$. The structure of the DNN remains the same as in the linear case.
We compare the one-shot method with neural network approximation resulting from the EKI with Algorithm 9 with the Tikhonov solution of the reduced formulation, which has been approximated by a quasi-Newton method. The scaling parameter λ in Algorithm 9 is determined by the ODE $\mathrm{d}\lambda/\mathrm{d}t = 1$, i.e., the scaling parameter grows linearly. Similarly to the linear case, we find that the one-shot method with neural network approximation leads to a good approximation of the Tikhonov solution for the reduced model, cf., Figure 8.17. In Figure 8.18, we observe that the penalty parameter λ drives the estimate towards feasibilty, i.e. towards the solution of the constrained optimization problem.

Two-dimensional example

Our final numerical example is based on the two-dimensional Poisson equation

$$\begin{aligned} -\Delta u &= z^\dagger \quad \text{in } D := (0, 1)^2, \\ u &= 0 \quad \text{on } \partial D, \end{aligned} \tag{8.57}$$

for which we consider again the problem of recovering the unknown source term $z^\dagger$ from noisy observations

$$y = \mathcal{O}(u^\dagger) + \eta^\dagger, \tag{8.58}$$

with $u^\dagger$ denoting the solution of (8.57). We consider an observation operator $\mathcal{O}$ observing $n_y = 50$ randomly picked observation points x_i, $i = 1, \ldots, n_y$, as illustrated in Figure 8.19.

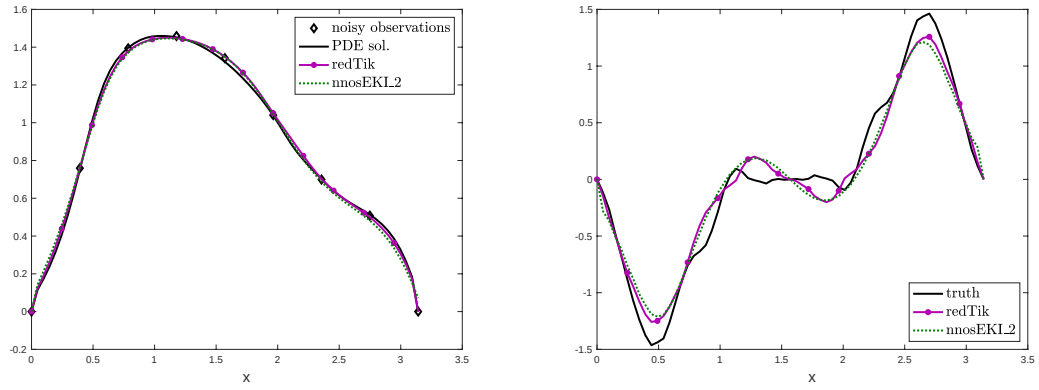


Figure 8.17: Comparison of parameter estimation given by the EKI with Algorithm 9 (osEKI_2) and the Tikhonov solution (on the left-hand side) and corresponding PDE solution (on the right-hand side).

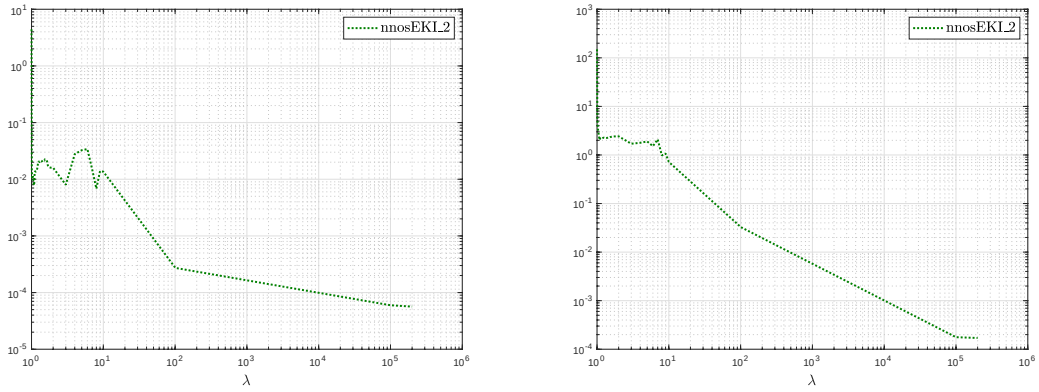


Figure 8.18: Data misfit given by the EKI with Algorithm 9 (osEKI_2) for the neural network based one-shot inversion compared (on the left-hand side) and residual of the forward problem (on the right-hand side), both with respect to λ.

We numerically approximate the forward model (8.57) with continuous, piecewise linear finite element basis functions on a mesh with 95 grid points in D and 40 grid points on ∂D using the `MATLAB Partial Differential Equation Toolbox`. We again denote the approximated solution operator by $S \in \mathbb{R}^{n_z \times n_z}$, with $n_u = 95$. Similar as before, we assume the unknown parameter z to be Gaussian, with (discretized) covariance operator $C_0 = \beta(\tau \cdot \mathrm{id} - \Delta)^{-\nu}$ for $\beta = 100$, $\nu = 2$ and $\tau = 1$. We assume the observational noise covariance to be $\Gamma_{\text{obs}} = 0.01 \cdot I_{n_y}$, whereas we assume a model covariance $\hat{\Gamma}_{\text{model}} = 0.1 \cdot I_{n_z}$. We set the regularization parameters $\alpha_1 = 0.002$ and $\alpha_2 = 0$. The DNN consists of $L = 3$ layers, with $N_1 = N_2 = 10$ hidden neurons, $N_0 = 2$ input neurons and $N_L = 1$ output neuron, and sigmoid activation function. The setting of the EKI is as described above with $J = 300$ particles drawn as i.i.d. sample from the prior. Figure 8.19 shows the truth and the corresponding PDE solution.

In the following, we compare the neural network based one-shot formulation, solved by the EKI with Algorithm 9, to the explicit Tikhonov solution of the reduced formulation. The scaling parameter λ in Algorithm 9 is determined by the ODE $\mathrm{d}\lambda/\mathrm{d}t = 1/\lambda^2$. Figure 8.20

demonstartes that the EKI leads to a comparable solution. The proposed approach leads to a feasible solution with respect to the forward problem, cf. Figure 8.21.

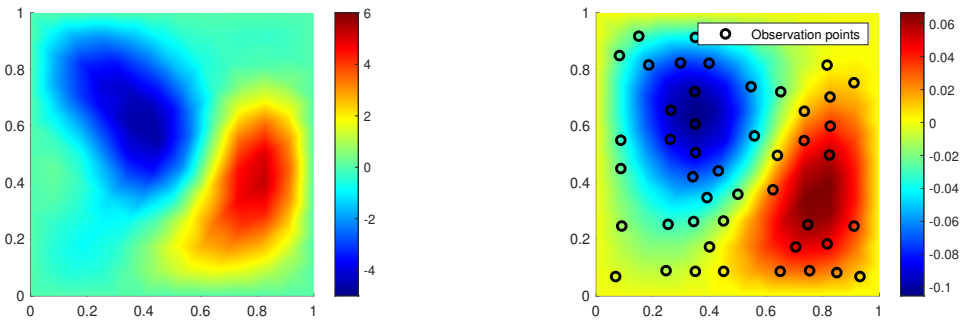


Figure 8.19: Ground truth (left-hand side) and the corresponding PDE solution (right-hand side).

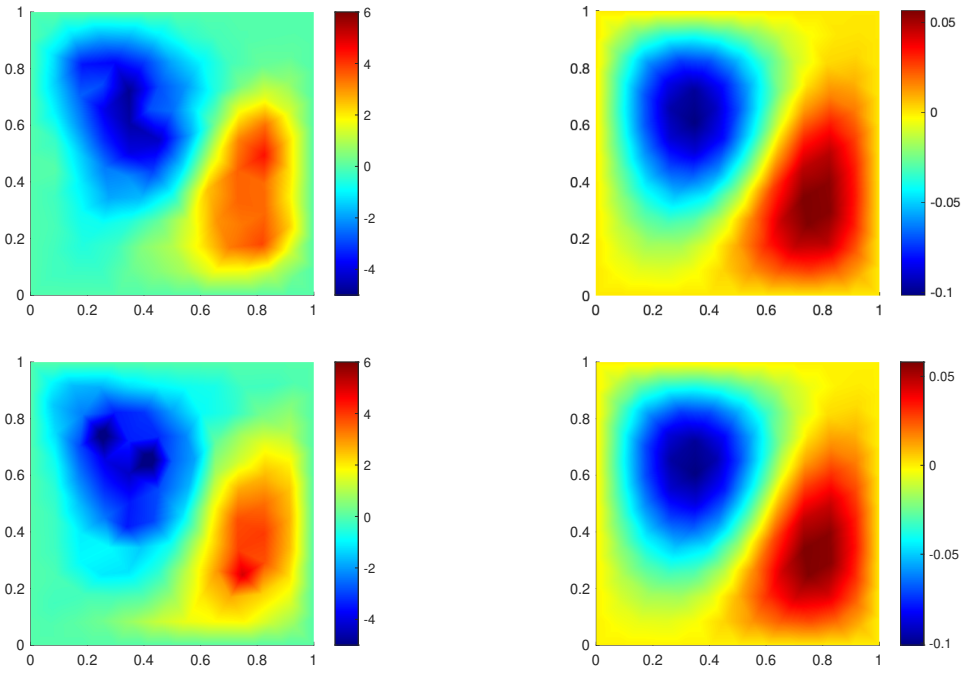


Figure 8.20: Comparison of parameter estimation given by the EKI with Algorithm 9 (osEKI_2) (below) and the Tikhonov solution (above) (on the left-hand side) and corresponding PDE solution (on the right-hand side).

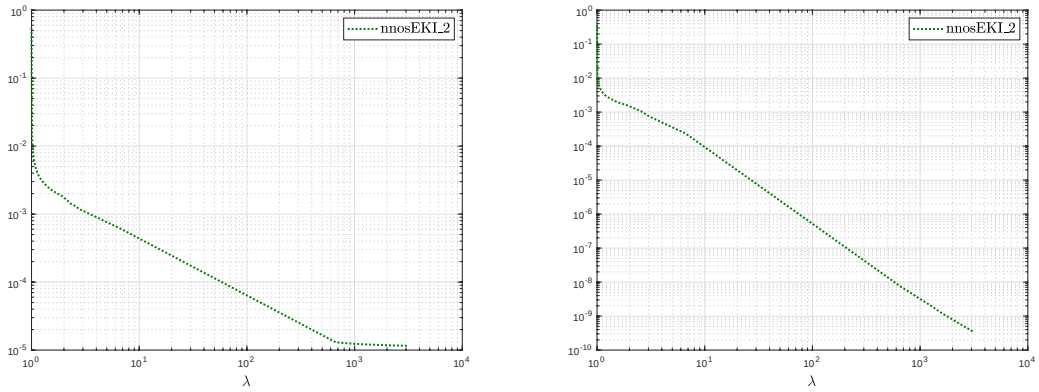


Figure 8.21: Data misfit given by the EKI with Algorithm 9 (osEKI_2) for the neural network based one-shot inversion compared (on the left-hand side) and residual of the forward problem (on the right-hand side), both with respect to λ.

9 Conclusions and outlook

We close this thesis with a short summary and a brief discussion about interesting future research directions.
We discussed the PDE-constrained optimization problem under uncertainty in a very general setting in Chapter 3. Particularly, we presented results about the existence and optimality of solutions along with optimality conditions under different sets of assumptions on the risk measure, the cost functional, and the constraint.
For the development of efficient methods to solve the optimal control problem under uncertainty and the error analysis, we focused on tracking-type objective functionals composed with sufficiently smooth risk measures and constraints that are sufficiently regular with respect to the uncertainty. We study three optimal control problems in detail in Chapter 4. In particular, we investigate their parametric regularity in order to apply to them the error bounds and concergence results which are developed in the following chapters.
In fact, many of the presented results, in particular in Chapter 5 and Chapter 6, are not limited to the application to optimal control problems, but are derived in such a general setting that they find applications in different areas in uncertainty quantification and related fields. This proofs that the problems considered in this thesis are not only interesting on their own, but also have the potential to reveal interesting insights into research questions arising in related fields.

Chapter 5

Chapter 5 is devoted to the dimension truncation error analysis for a class of high-dimensional integration problems. A popular approach to derive dimension truncation error rates in the context of PDEs with random coefficients is based on the Neumann series. This technique heavily relies on the parametric structure of the problem and is thus practically constrained to affine parametric operator equations. In contrast to the Neumann series approach, we utilize the parametric regularity of the integration problem to derive error bounds and convergence rates based on Taylor series. Our proposed technique appears to be quite robust as we were able to improve dimension truncation convergence rates even in a non-affine setting, for instance for elliptic PDEs with lognormal random coefficients. We analyze the dimension truncation error in the general setting of separable Banach spaces and with respect to the generalized β-Gaussian distribution. Thus our dimension truncation error rates immediately apply for spatially discretized PDE solutions obtained using a conforming finite element method. Furthermore, our proposed method enables the development of dimension truncation rates for sufficiently smooth nonlinear quantities of interest of the PDE response, provided that the composition of the nonlinear quantity of interest with the PDE solution satisfies the assumptions of our dimension truncation result.

Chapter 6

In Chapter 6 we recall error bounds and convergence rates for randomly shifted rank-1 lattice rules for integration of real-valued functions. By exploiting the concept of duality in Banach spaces, we succeeded in generalizing these well-known bounds and convergence rates for real-valued integrands to integrands that take values in separable Banach spaces. This generalization was motivated by the fact that the integrals involved in optimal control problems subject to PDE constraints with random coefficients are typically Bochner integrals, i.e., integrals over Banach space-valued objects. However the derived results are not at all restricted to optimal control problems and open up many interesting areas of application for QMC integration. A possible extension of our new results would be the generalization to different distributions, such as the lognormal or generalized β-Gaussian distribution. Together with our results from Chapter 5 this would provide a very general and uniform framework for QMC integration in Banach spaces. In particular, the application of QMC integration in the context of PDEs with random coefficients would simplify to checking the regularity assumptions in our results.

Chapter 7

We presented a MLQMC method for the estimation of gradients for PDE-constrained optimization problems. Specifically the objective function is a tracking-type functional composed with the expected value and the constraint is an elliptic PDE with lognormal random diffusion coefficient. Numerical results for this particular problem show that the MLQMC method outperforms the MLMC and the QMC method. Its superior performance is due to the faster decay of the variances of each term in the telescopic sum (7.29) defining the multilevel method. Based on the parametric regularity of the problem, we derived a rigorous analysis of our MLQMC method confirming the faster decay of the relevant variances. While the numerical experiments and the analysis are performed for the specific elliptic model problem, we expect that the results carry over to other problems as well, such as the parabolic PDE-constrained optimal control problem. The numerical and theoretical evidence for a more general problem class remains to be investigated.

Chapter 8

In Chapter 8 we focus on the incorporation of surrogates into the optimization problems. In particular, we aim to replace the computational intense solution of the underlying model, typically a PDE, by a surrogate which is cheap to evaluate.
In Section 8.1 we apply this strategy to the PDE-constrained optimal control problem under uncertainty and in Section 8.2 we transfer the ideas to Bayesian inverse problems.
Our proposed framework is based on a quadratic penalization on the PDE residual and very flexible in the sense that it allows for different surrogates, such as polynomial expansions, reduced basis approaches, or neural networks. In our framework the surrogate is trained only for the optimal control during the optimization of the underlying problem. This should be contrasted with the expensive offline training, where a surrogate is trained for all admissible controls and is substituted afterwards into the underlying optimization problem. The numerical experiments for the optimal control problem subject to the elliptic PDE constraint show promising results and applications to more complex optimization

problems under uncertainty will be subject to future work.
We analyzed the stochastic gradient method for the optimization of the penalized problem. In more complex situations, gradients might not be available due to the use of black-box solvers or computational limits. In this case, the application of derivative-free optimization techniques, in particular Kalman based methods, are expected to be efficient in this setting. We studied the application of the ensemble Kalman filter to a penalized problem in the context of Bayesian inverse problems in Section 8.2. In this section we demonstrated that the ensemble Kalman inversion for neural network based one-shot inversion is a promising method, regarding both estimation quality of the unknown parameter and computational feasibility. The connection between the penalized optimization problem and the Bayesian inverse problem setting with vanishing noise allowed to establish a convergence result in a simplified linear setting. Several directions for future work arise naturally from the presented ideas. For instance, the theoretical analysis of the neural network based one-shot inversion using recent results about the expressivity of neural networks in the context of parametric PDEs is a promising direction for future research. Furthermore, a comparison to state-of-the-art optimization algorithms in the machine learning community should be discussed. Moreover, it would be interesting to investigate if the emulation of the underlying dynamics of the PDEs in the considered problems can be improved by choosing more sophisticated architectures of neural networks, such as residual neural networks or convolutional neural networks, see [138].

Bibliography

[1] R. J. Adler. *The Geometry of Random Fields*. Society for Industrial and Applied Mathematics, 2010. doi: 10.1137/1.9780898718980.

[2] S. Agapiou, M. Burger, M. Dashti, and T. Helin. Sparsity-promoting and edge-preserving maximum a posteriori estimators in non-parametric Bayesian inverse problems. *Inverse Probl.*, 34(4):045002, 2018. doi: 10.1088/1361-6420/aaacac.

[3] A. Ahmadi-Javid. Entropic Value-at-Risk: a new coherent risk measure. *J. Optim. Theory Appl.*, 155(3):1105–1123, 2012. doi: 10.1007/s10957-011-9968-2.

[4] A. 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(1):466–492, 2017. doi: 10.1137/16M109870X.

[5] T. S. Angell and A. Kirsch. *Optimization Methods in Electromagnetic Radiation*. Springer Monographs in Mathematics. Springer New York, 2006. ISBN 9780387218274. doi: 10.1007/b97629.

[6] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb. Solving inverse problems using data-driven models. *Acta Numer.*, 28:1–174, 2019. doi: 10.1017/S0962492919000059.

[7] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Math. Financ.*, 9(3):203–228, 1999. doi: 10.1111/1467-9965.00068.

[8] M. Bachmayr, A. Cohen, and W. Dahmen. Parametric PDEs: sparse or low-rank approximations? *IMA J. Numer. Anal.*, 38(4):1661–1708, 09 2017. doi: 10.1093/imanum/drx052.

[9] M. Badiale and E. Serra. *Semilinear Elliptic Equations for Beginners: Existence Results via the Variational Approach*. Universitext. Springer London, 2010. ISBN 9780857292278. doi: 10.1007/978-0-85729-227-8.

[10] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999. ISBN 978-1-886529-05-2.

[11] K. Bhattacharya, B. Hosseini, N. B. Kovachki, and A. M. Stuart. Model reduction and neural networks for parametric PDEs. *J. Comput. Math.*, 7:121–157, 2021. doi: 10.5802/smai-jcm.74.

[12] D. Blömker, C. Schillings, P. Wacker, and S. Weissmann. Well posedness and convergence analysis of the ensemble Kalman inversion. *Inverse Probl.*, 35(8):085007, 2019. doi: 10.1088/1361-6420/ab149c.

[13] V. Bogachev. *Measure Theory*, volume 1 of *Measure Theory*. Springer Berlin Heidelberg, 2007. ISBN 9783540345145. doi: 10.1007/978-3-540-34514-5.

[14] A. Borzì and V. Schulz. *Computational Optimization of Systems Governed by Partial Differential Equations*. Society for Industrial and Applied Mathematics, USA, 2012. ISBN 1611972043. doi: 10.1137/1.9781611972054.

[15] A. Borzì and G. von Winckel. Multigrid methods and sparse-grid collocation techniques for parabolic optimal control problems with random coefficients. *SIAM J. Sci. Comput.*, 31(3):2172–2192, 2009. doi: doi.org/10.1137/070711311.

[16] A. Borzì and G. von Winckel. A POD framework to determine robust controls in PDE optimization. *Comput. Vis. Sci.*, 14(3):91–103, 2011. doi: 10.1007/s00791-011-0165-5.

[17] G. Castiglione, A. Frosini, E. Munarini, A. Restivo, and S. Rinaldi. Combinatorial aspects of L-convex polyominoes. *European J. Combin.*, 28(6):1724–1741, 2007. doi: 10.1016/j.ejc.2006.06.020.

[18] N. K. Chada and X. Tong. Convergence acceleration of ensemble Kalman inversion in nonlinear settings. *Math. Comput.*, 91(335):1247–1280, 2022. doi: 10.1090/mcom/3709.

[19] N. K. Chada, A. M. Stuart, and X. T. Tong. Tikhonov regularization within ensemble Kalman inversion. *SIAM J. Numer. Anal.*, 58(2):1263–1294, 2019. doi: 10.1137/19M1242331.

[20] N. K. Chada, C. Schillings, X. T. Tong, and S. Weissmann. Consistency analysis of bilevel data-driven learning in inverse problems. *Commun. Math. Sci.*, 20(1):123 – 164, 2021. doi: 10.4310/CMS.2022.v20.n1.a4.

[21] G. Chan and A. T. Wood. Algorithm AS 312: An algorithm for simulating stationary Gaussian random fields. *J. Appl. Stat.*, pages 171–181, 1997. doi: 10.1111/1467-9876.00057.

[22] J. Charrier. Strong and weak error estimates for elliptic partial differential equations with random coefficients. *SIAM J. Numer. Anal.*, 50(1):216–246, 2012. doi: 10.1137/100800531.

[23] J. Charrier, R. Scheichl, and A. L. Teckentrup. Finite element error analysis of elliptic PDEs with random coefficients and its application to multilevel Monte Carlo methods. *SIAM J. Numer. Anal.*, 51(1):322–352, 2013. doi: 10.1137/110853054.

[24] P. Chen and O. Ghattas. Sparse polynomial approximations for affine parametric saddle point problems. *Vietnam J. Math.*, 2022. doi: 10.1007/s10013-022-00584-1.

[25] P. Chen and A. Quarteroni. Weighted reduced basis method for stochastic optimal control problems with elliptic PDE constraint. *SIAM/ASA J. Uncertain. Quantif.*, 2(1):364–396, 2014. doi: 10.1137/130940517.

[26] Y. Chen, B. Hosseini, H. Owhadi, and A. M. Stuart. Solving and learning nonlinear PDEs with Gaussian processes, 2021. arXiv:2103.12959 [math.NA].

[27] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North-Holland, 1978. ISBN 978-0-89871-514-9. doi: 10.1137/1.9780898719208.

[28] C. Clason, T. Helin, R. Kretschmann, and P. Piiroinen. Generalized modes in Bayesian inverse problems. *SIAM/ASA J. Uncertain. Quantif.*, 7(2):652–684, 2019. doi: 10.1137/18M1191804.

[29] K. A. Cliffe, M. B. Giles, R. Scheichl, and A. L. Teckentrup. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Comput. Vis. Sci.*, 14(3), 2011. doi: 10.1007/s00791-011-0160-x.

[30] A. Cohen and R. DeVore. Approximation of high-dimensional parametric PDEs. *Acta Numer.*, 24:1–159, 2015. doi: 10.1017/S0962492915000033.

[31] A. Cohen, R. DeVore, and C. Schwab. Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs. *Found. Comput. Math.*, 10(6):615–646, 2010. doi: 10.1007/s10208-010-9072-2.

[32] D. Cohn. *Measure Theory: Second Edition*. Birkhäuser Advanced Texts Basler Lehrbücher. Springer New York, 2013. ISBN 9781461469568. doi: 10.1007/978-1-4899-0399-0.

[33] R. Cools, F. Y. Kuo, and D. Nuyens. Constructing embedded lattice rules for multivariate integration. *SIAM J. Sci. Comput.*, 28(6):2162–2188, 2006. doi: 10.1137/06065074X.

[34] M. Dashti and A. M. Stuart. *The Bayesian approach to inverse problems*, pages 311–428. Springer International Publishing, 2017. ISBN 978-3-319-12385-1. doi: 10.1007/978-3-319-12385-1_7.

[35] M. Dashti, K. J. H. Law, A. M. Stuart, and J. Voss. MAP estimators and their consistency in Bayesian nonparametric inverse problems. *Inverse Probl.*, 29(9):095017, 2013. doi: 10.1088/0266-5611/29/9/095017.

[36] R. Dautray and J.-L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology: Volume 5 Evolution Problems I*. Springer, Heidelberg, 2012. ISBN 10: 354050205X.

[37] R. DeVore. Nonlinear approximation. *Acta Numer.*, 7:51–150, 1998. doi: 10.1017/S0962492900002816.

[38] J. Dick, F. Y. Kuo, and I. H. Sloan. High-dimensional integration: the quasi-Monte Carlo way. *Acta Numer.*, 22:133–288, 2013. doi: 10.1017/S0962492913000044.

[39] J. Dick, F. Y. Kuo, Q. T. L. Gia, D. Nuyens, and C. Schwab. Higher order QMC Petrov–Galerkin discretization for affine parametric operator equations with random field inputs. *SIAM J. Numer. Anal.*, 52(6):2676–2702, 2014. doi: 10.1137/130943984.

[40] J. Dick, Q. T. L. Gia, and C. Schwab. Higher order quasi-Monte Carlo integration for holomorphic, parametric operator equations. *SIAM/ASA J. Uncertain. Quantif.*, 4(1):48–79, 2016. doi: 10.1137/140985913.

[41] J. Dick, R. N. Gantner, Q. T. L. Gia, and C. Schwab. Higher order quasi-Monte Carlo integration for Bayesian PDE inversion. *Comput. Math. Appl.*, 77(1):144–172, 2019. doi: 10.1016/j.camwa.2018.09.019.

[42] J. Diestel. *Sequences and Series in Banach Spaces*. Graduate Texts in Mathematics. Springer New York, 1984. ISBN 9780387908595. doi: 10.1007/978-1-4612-5200-9.

[43] C. R. Dietrich and G. N. Newsam. Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. *SIAM J. Sci. Comput.*, 18(4):1088–1107, 1997. doi: 10.1137/S1064827592240555.

[44] Dong, Guozhi, Hintermüller, Michael, and Papafitsoros, Kostas. Optimization with learning-informed differential equation constraints and its applications. *ESAIM: COCV*, 28:3, 2022. doi: 10.1051/cocv/2021100.

[45] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*. Mathematics and Its Applications. Springer Netherlands, 1996. ISBN 9780792341574.

[46] O. G. Ernst, B. Sprungk, and H.-J. Starkloff. Analysis of the ensemble and polynomial chaos Kalman filters in Bayesian inverse problems. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):823–851, 2015. doi: 10.1137/140981319.

[47] L. C. Evans. *Partial Differential Equations*. Graduate studies in mathematics. American Mathematical Society, 2010. ISBN 9780821849743. doi: 10.1090/gsm/019.

[48] G. Evensen. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53(4):343–367, 2003. doi: 10.1007/s10236-003-0036-9.

[49] C. A. Floudas and P. M. Pardalos. *Encyclopedia of Optimization*. Encyclopedia of Optimization. Springer US, 2008. ISBN 9780387747583. doi: 10.1007/978-0-387-74759-0.

[50] H. Föllmer and T. Knispel. *Convex risk measures: basic facts, law-invariance and beyond, asymptotics for large portfolios*, pages 507–554. World Scientific, 2013. doi: 10.1142/8557.

[51] P. Frauenfelder, C. Schwab, and R. Todor. Finite elements for elliptic problems with stochastic coefficients. *Comput. Methods Appl. Mech. Engrg.*, 194:205–228, 2005. doi: 10.1016/j.cma.2004.04.008.

[52] R. N. Gantner. *Computational Higher-Order Quasi-Monte Carlo for Random Partial Differential Equations*. PhD thesis, ETH Zurich, 2017.

[53] R. N. Gantner. Dimension truncation in QMC for affine-parametric operator equations. In A. B. Owen and P. W. Glynn, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2016*, pages 249–264, Stanford, CA, 2018. Springer. doi: 10.1007/978-3-319-91436-7_13.

[54] R. N. Gantner, L. Herrmann, and C. Schwab. Multilevel QMC with product weights for affine-parametric, elliptic PDEs. In J. Dick, F. Y. Kuo, and H. Woźniakowski, editors, *Contemporary Computational Mathematics - A Celebration of the 80th Birthday of Ian Sloan*, pages 373–405. Springer International Publishing, 2018. doi: 10.1007/978-3-319-72456-0_18.

[55] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart. Interacting Langevin diffusions: gradient structure and ensemble Kalman sampler. *SIAM J. Appl. Dyn.*, 19(1):412–441, 2020. doi: 10.1137/19M1251655.

[56] C. Geiersbach and T. Scarinci. Stochastic proximal gradient methods for nonconvex problems in Hilbert spaces. *Comput. Optim. Appl.*, 78(3):705–740, 2021. doi: 10.1007/s10589-020-00259-y.

[57] M. Geist, P. Petersen, M. Raslan, R. Schneider, and G. Kutyniok. Numerical solution of the parametric diffusion equation by deep neural networks. *J. Sci. Comput.*, 88 (1):22, 2021. doi: 10.1007/s10915-021-01532-w.

[58] R. G. Ghanem and P. D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Courier Corporation, Mineola, NY, 2003. ISBN 978-1-4612-7795-8. doi: 10.1007/978-1-4612-3094-6.

[59] D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer-Verlag, 2001. ISBN 978-3-540-41160-4. doi: 10.1007/978-3-642-61798-0. Reprint of the 1998 edition.

[60] A. D. Gilbert, I. G. Graham, F. Y. Kuo, R. Scheichl, and I. H. Sloan. Analysis of quasi-Monte Carlo methods for elliptic eigenvalue problems with stochastic coefficients. *Numer. Math.*, 142(4):863–915, 2019. doi: 10.1007/s00211-019-01046-6.

[61] A. D. Gilbert, F. Y. Kuo, and I. H. Sloan. Analysis of preintegration followed by quasi-Monte Carlo integration for distribution functions and densities, 2021. arXiv:2112.10308 [math.NA].

[62] M. B. Giles. Multilevel Monte Carlo methods. *Acta Numer.*, 24:259–328, 2015. doi: 10.1017/S096249291500001X.

[63] C. J. Gittelson. Stochastic Galerkin discretization of the log-normal isotropic diffusion problem. *Math. Models Methods Appl. Sci.*, 20(2):237–263, 2010. doi: 10.1142/S0218202510004210.

[64] C. J. Gittelson. *Adaptive Galerkin methods for parametric and stochastic operator equations*. PhD thesis, ETH Zurich, 2011. doi: 10.3929/ethz-a-006380316.

[65] C. J. Gittelson and C. Schwab. Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs. *Acta Numer.*, 20:291–467, 2011. doi: 10.1017/S0962492911000055.

[66] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org` (27 September 2022).

[67] H. W. Gould. *Combinatorial Identities: A Standardized Set of Tables Listing 500 Binomial Coefficient Summations*. Morgantown, 1972.

[68] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, Amsterdam, seventh edition, 2007. doi: 10.1016/C2010-0-64839-5.

[69] I. G. Graham, F. Y. Kuo, D. Nuyens, R. Scheichl, and I. H. Sloan. Quasi-Monte Carlo methods for elliptic PDEs with random coefficients and applications. *J. Comput. Phys.*, 230(10):3668–3694, 2011. doi: 10.1016/j.jcp.2011.01.023.

[70] I. G. Graham, F. Y. Kuo, J. A. Nichols, R. Scheichl, C. Schwab, and I. H. Sloan. Quasi-Monte Carlo finite element methods for elliptic PDEs with lognormal random coefficients. *Numer. Math.*, 131(2):329–368, 2015. doi: 10.1007/s00211-014-0689-y.

[71] I. G. Graham, F. Y. Kuo, D. Nuyens, R. Scheichl, and I. H. Sloan. Analysis of circulant embedding methods for sampling stationary random fields. *SIAM J. Numer. Anal.*, 56(3):1871–1895, 2018. doi: 10.1137/17M1149730.

[72] I. G. Graham, F. Y. Kuo, D. Nuyens, R. Scheichl, and I. H. Sloan. Circulant embedding with QMC: analysis for elliptic PDE with lognormal coefficients. *Numer. Math.*, 140(2):479–511, 2018. doi: 10.1007/s00211-018-0968-0.

[73] S. Günther, L. Ruthotto, J. B. Schroder, E. C. Cyr, and N. R. Gauger. Layer-parallel training of deep residual neural networks. *SIAM J. Math. Data Sci.*, 2(1): 1–23, 2020. doi: 10.1137/19M1247620.

[74] P. A. Guth and V. Kaarnioja. Generalized dimension truncation error analysis for high-dimensional numerical integration: lognormal setting and beyond, 2022. arXiv:2209.06176 [math.NA].

[75] P. A. Guth and A. van Barel. Multilevel quasi-Monte Carlo for optimization under uncertainty, 2021. arXiv:2109.14367 [math.NA].

[76] P. A. Guth, V. Kaarnioja, F. Y. Kuo, C. Schillings, and I. H. Sloan. A quasi-Monte Carlo method for optimal control under uncertainty. *SIAM/ASA J. Uncertain. Quantif.*, 9(2):354–383, 2021. doi: 10.1137/19M1294952.

[77] P. A. Guth, V. Kaarnioja, F. Y. Kuo, C. Schillings, and I. H. Sloan. Parabolic PDE-constrained optimal control under uncertainty with entropic risk measure using quasi-Monte Carlo integration, 2022. arxiv:2208.02767 [math.NA].

[78] P. A. Guth, C. Schillings, and S. Weissmann. 14 Ensemble Kalman filter for neural network-based one-shot inversion. In R. Herzog, M. Heinkenschloss, D. Kalise, G. Stadler, and E. Trélat, editors, *Optimization and Control for Partial Differential Equations: Uncertainty quantification, open and closed-loop control, and shape optimization*, pages 393–418, Berlin, Boston, 2022. De Gruyter. doi: 10.1515/9783110695984-014.

[79] P. A. Guth, C. Schillings, and S. Weissmann. A general framework for machine learning based optimization under uncertainty, 2022. arXiv:2112.11126 [math.OC].

[80] E. Haber, F. Lucka, and L. Ruthotto. Never look back - a modified EnKF method and its application to the training of neural networks without back propagation, 2018. arXiv:1805.08034 [math.NA].

[81] P. Halmos. *Measure Theory*. Graduate Texts in Mathematics. Springer New York, 1976. ISBN 9780387900889. doi: 10.1007/978-1-4684-9440-2.

[82] J. Han, A. Jentzen, and W. E. Solving high-dimensional partial differential equations using deep learning. *Proc. Natl. Acad. Sci. U. S. A.*, 115(34):8505–8510, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1718942115.

[83] T. Helin and M. Burger. Maximum a posteriori probability estimates in infinite-dimensional Bayesian inverse problems. *Inverse Probl.*, 31(8):085009, jul 2015. doi: 10.1088/0266-5611/31/8/085009.

[84] L. Herrmann and C. Schwab. QMC integration for lognormal-parametric, elliptic PDEs: local supports and product weights. *Numer. Math.*, 141(1):63–102, 2019. doi: 10.1007/978-1-4684-9440-2.

[85] L. Herrmann, C. Schwab, and J. Zech. Deep ReLU neural network expression rates for data-to-QoI maps in Bayesian PDE inversion. Technical Report 2020-02, Seminar for Applied Mathematics, ETH Zürich, Switzerland, 2020.

[86] L. Herrmann, M. Keller, and C. Schwab. Quasi-Monte Carlo Bayesian estimation under Besov priors in elliptic inverse problems. *Math. Comp.*, 90:1831–1860, 2021. doi: 10.1090/mcom/3615.

[87] D. Higdon, M. Kennedy, J. C. Cavendish, J. A. Cafeo, and R. D. Ryne. Combining field data and computer simulations for calibration and prediction. *SIAM J. Sci. Comput.*, 26(2):448–466, Feb. 2005. ISSN 1064-8275. doi: 10.1137/S1064827503426693.

[88] M. Hinze. A variational discretization concept in control constrained optimization: the linear-quadratic case. *Comput. Optim. Appl.*, 30:45–61, 01 2005. doi: 10.1007/s10589-005-4559-5.

[89] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints*, volume 23. Springer Netherlands, 2008. ISBN 9781402088391. doi: 10.1007/978-1-4020-8839-1.

[90] T. Hytönen, J. van Neerven, M. Veraar, and L. Weis. *Analysis in Banach Spaces: Volume I: Martingales and Littlewood-Paley Theory.* Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics. Springer International Publishing, 2016. ISBN 9783319485201. doi: 10.1007/978-3-319-48520-1.

[91] M. A. Iglesias, K. J. H. Law, and A. M. Stuart. Ensemble Kalman methods for inverse problems. *Inverse Probl.*, 29(4):045001, 2013. doi: 10.1088/0266-5611/29/4/045001.

[92] K. Ito and K. Kunisch. An augmented Lagrangian technique for variational inequalities. *Appl. Math. Optim.*, 21(1):223–241, 1990. doi: 10.1007/BF01445164.

[93] S. Joe and F. Y. Kuo. Constructing Sobol sequences with better two-dimensional projections. *SIAM J. Sci. Comput.*, 30(5):2635–2654, 2008. doi: 10.1137/070709359.

[94] V. Kaarnioja, F. Y. Kuo, and I. H. Sloan. Uncertainty quantification using periodic random variables. *SIAM J. Numer. Anal.*, 58(2):1068–1091, 2020. doi: 10.1137/19M1262796.

[95] V. Kaarnioja, Y. Kazashi, F. Y. Kuo, F. Nobile, and I. H. Sloan. Fast approximation by periodic kernel-based lattice-point interpolation with application in uncertainty quantification. *Numer. Math.*, 150(1):33–77, 2022. doi: 10.1007/s00211-021-01242-3.

[96] J. Kaipio and E. Somersalo. *Statistical and computational inverse problems*. Applied mathematical sciences; Volume 160. Springer Science & Business Media, New York, NY, 2010. ISBN 9781441919649. doi: 10.1007/b138659.

[97] B. Kaltenbacher. Regularization based on all-at-once formulations for inverse problems. *SIAM J. Numer. Anal.*, 54:2594–2618, 2016. doi: 10.1137/16M1060984.

[98] B. Kaltenbacher. All-at-once versus reduced iterative methods for time dependent inverse problems. *Inverse Probl.*, 33(6):064002, 2017. doi: 10.1088/1361-6420/aa6f34.

[99] B. Kaltenbacher, A. Neubauer, and O. Scherzer. *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*. De Gruyter, Berlin, New York, 2008. ISBN 9783110208276. doi: 10.1515/9783110208276.

[100] K. Karhunen. Über lineare methoden in der wahrscheinlichkeitsrechnung. *Annales Academiae Scientiarum Fennicae. Series A. 1, Mathematica-physica*, 37:1–79, 1947.

[101] Katana, 2010. doi: 10.26190/669X-A286.

[102] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *J. R. Stat. Soc., B: Stat. Methodol.*, 63(3):425–464, 2001. doi: 10.1111/1467-9868.00294.

[103] D. P. Kingma and J. Ba. Adam: a method for stochastic optimization, 2014. arXiv:1412.6980 [cs.LG].

[104] A. Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer London, 2007. ISBN 9781848000483. doi: 10.1007/978-1-4471-5361-0.

[105] D. P. Kouri and T. M. Surowiec. Risk-averse PDE-constrained optimization using the conditional value-at-risk. *SIAM J. Optim.*, 26(1):365–396, 2016. doi: 10.1137/140954556.

[106] D. P. Kouri and T. M. Surowiec. Existence and optimality conditions for risk-averse PDE-constrained optimization. *SIAM/ASA J. Uncertain. Quantif.*, 6(2):787–815, 2018. doi: 10.1137/16M1086613.

[107] D. P. Kouri and T. M. Surowiec. Epi-regularization of risk measures. *Math. Oper. Res.*, 45(2):774–795, 2020. doi: 10.1287/moor.2019.1013.

[108] 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(4):A1847–A1879, 2013. doi: 10.1137/120892362.

[109] N. B. Kovachki and A. M. Stuart. Ensemble Kalman inversion: a derivative-free technique for machine learning tasks. *Inverse Probl.*, 35(9):095005, 2019. doi: 10.1088/1361-6420/ab1c3a.

[110] D. Kressner and C. Tobler. Low-rank tensor Krylov subspace methods for parametrized linear systems. *SIAM J. Matrix Anal. Appl.*, 32(4):1288–1316, 2011. doi: 10.1137/100799010.

[111] 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(3):2442–2471, 2013. doi: 10.1137/110847597.

[112] F. Y. Kuo. Lattice rule generating vectors, 2022. `https://web.maths.unsw.edu.au/~fkuo/lattice/index.html` (15 July 2022).

[113] F. Y. Kuo and D. Nuyens. Application of quasi-Monte Carlo methods to elliptic PDEs with random diffusion coefficients: a survey of analysis and implementation. *Found. Comput. Math.*, 16(6):1631–1696, 2016. doi: 10.1007/s10208-016-9329-5.

[114] F. Y. Kuo and D. Nuyens. QMC4PDE, 2022. `https://people.cs.kuleuven.be/~dirk.nuyens/qmc4pde/` (15 July 2022).

[115] F. Y. Kuo, I. H. Sloan, G. W. Wasilkowski, and B. J. Waterhouse. Randomly shifted lattice rules with the optimal rate of convergence for unbounded integrands. *J. Complex.*, 26(2):135–160, 2010. doi: 10.1016/j.jco.2009.07.005.

[116] 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(6):3351–3374, 2012. doi: 10.1137/110845537.

[117] F. Y. Kuo, C. Schwab, and I. H. Sloan. Multi-level quasi-Monte Carlo finite element methods for a class of elliptic PDEs with random coefficients. *Found. Comput. Math.*, 15(2):411–449, 2015. doi: 10.1007/s10208-014-9237-5.

[118] F. Y. Kuo, D. Nuyens, L. Plaskota, I. H. Sloan, and G. W. Wasilkowski. Infinite-dimensional integration and the multivariate decomposition method. *J. Comput. Appl. Math.*, 326:217–234, 2017. doi: 10.1016/j.cam.2017.05.031.

[119] F. Y. Kuo, R. Scheichl, C. Schwab, I. H. Sloan, and E. Ullmann. Multilevel quasi-Monte Carlo methods for lognormal diffusion problems. *Math. Comp.*, 86(308): 2827–2860, 2017. doi: 10.1090/mcom/3207.

[120] G. Kutyniok, P. Petersen, M. Raslan, and R. Schneider. A theoretical analysis of deep neural networks and parametric PDEs. *Constr. Approx.*, 2021. doi: 10.1007/s00365-021-09551-4.

[121] M. Loève. Fonctions aléatoires de second ordre. *Revue Scientifique*, pages 195–206, 1946.

[122] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nat. Mach. Intell.*, 3(3):218–229, 2021. doi: 10.1038/s42256-021-00302-5.

[123] J. A. Nichols and F. Y. Kuo. Fast CBC construction of randomly shifted lattice rules achieving $O(n^{-1+\delta})$ convergence for unbounded integrands over $\mathbb{R}^s$ in weighted spaces with POD weights. *J. Complex.*, 30(4):444–468, 2014. doi: 10.1016/j.jco.2014.02.004.

[124] D. Nuyens and R. Cools. Fast component-by-component construction of rank-1 lattice rules with a non-prime number of points. *J. Complex.*, 22(1):4–28, 2006. doi: 10.1016/j.jco.2005.07.002.

[125] D. Nuyens and R. Cools. Fast algorithms for component-by-component construction of rank-1 lattice rules in shift-invariant reproducing kernel Hilbert spaces. *Math. Comp.*, 75(254):903–920, 2006. ISSN 00255718, 10886842.

[126] J. A. A. Opschoor, P. C. Petersen, and C. Schwab. Deep ReLU networks and high-order finite element methods. *Anal. Appl.*, 12 2019. doi: 10.1142/S0219530519410136.

[127] B. T. Polyak. The convergence rate of the penalty function method. *USSR Computational Mathematics and Mathematical Physics*, 11(1):1–12, 1971. ISSN 0041-5553. doi: 10.1016/0041-5553(71)90094-2.

[128] J. Quaintance and H. W. Gould. *Combinatorial Identities for Stirling Numbers: The Unpublished Notes of H. W. Gould.* World Scientific Publishing Company, River Edge, NJ, 2015. doi: 10.1142/9821.

[129] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics informed deep learning (part I): data-driven solutions of nonlinear partial differential equations, 2017. arXiv:1711.10561 [cs.AI].

[130] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics informed deep learning (part II): data-driven discovery of nonlinear partial differential equations, 2017. arXiv:1711.10566 [cs.AI].

[131] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

[132] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Stat.*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586.

[133] R. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *J. Bank. Financ.*, 26(7):1443–1471, 2002. doi: 10.1016/S0378-4266(02)00271-6.

[134] G. Rozza, D. B. P. Huynh, and A. T. Patera. Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. *Arch. Comput. Methods Eng.*, 15(3):229 – 275, Sept. 2008. doi: 10.1007/s11831-008-9019-9.

[135] W. Rudin. *Functional Analysis.* International series in pure and applied mathematics. McGraw-Hill, 1991. ISBN 9780071009447.

[136] A. Ruszczynski and A. Shapiro. Optimization of convex risk functions. *Math. Oper. Res.*, 31(3):433–452, 2006. doi: 10/1287/moor.10500186.

[137] A. Ruszczynski and A. Shapiro. *Chapter 6: Risk Averse Optimization*, pages 271–385. MOS-SIAM, 2014. doi: 10.1137/1.9781611973433.ch6.

[138] L. Ruthotto and E. Haber. Deep neural networks motivated by partial differential equations. *J. Math. Imaging Vis.*, 62(3):352–364, 2020. doi: 10.1007/s10851-019-00903-1.

[139] R. A. Ryan. *Introduction to Tensor Products of Banach Spaces*. Springer Monographs in Mathematics. Springer London, 2002. ISBN 9781852334376. doi: 10.1007/978-1-4471-3903-4.

[140] S. A. Sauter and C. Schwab. *Boundary Element Methods*. Springer Series in Computational Mathematics. Springer Berlin Heidelberg, 2010. ISBN 9783540680925. doi: 10.1007/978-3-540-68093-2.

[141] T. H. Savits. Some statistical applications of Faa di Bruno. *J. Multivariate Anal.*, 97(10):2131–2140, 2006. doi: 10.1016/j.jmva.2006.03.001.

[142] C. Schillings and A. M. Stuart. Analysis of the ensemble Kalman filter for inverse problems. *SIAM J. Numer. Anal.*, 55(3):1264–1290, 2017. doi: 10.1137/16M105959X.

[143] C. Schillings and A. M. Stuart. Convergence analysis of ensemble Kalman inversion: the linear, noisy case. *Appl. Anal.*, 97(1):107–123, 2018. doi: 10.1080/00036811.2017.1386784.

[144] C. Schillings, S. Schmidt, and V. Schulz. Efficient shape optimization for certain and uncertain aerodynamic design. *Comput. Fluids*, 46(1):78–87, 2011. doi: 10.1016/j.compfluid.2010.12.007.

[145] K. D. Schmidt. *Maß und Wahrscheinlichkeit*. Springer-Lehrbuch. Springer, 2009. ISBN 9783540897293. doi: 10.1007/978-3-642-21026-6.

[146] C. Schwab. QMC Galerkin discretization of parametric operator equations. In J. Dick, F. Y. Kuo, G. W. Peters, and I. H. Sloan, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2012*, pages 613–629, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-41095-6. doi: 10.1007/978-3-642-41095-6_32.

[147] C. Schwab and R. Stevenson. Space-time adaptive wavelet methods for parabolic evolution problems. *Math. Comp.*, 78:1293–1318, 2009. doi: 10.1090/S0025-5718-08-02205-9.

[148] C. Schwab and R. A. Todor. Karhunen–Loève approximation of random fields by generalized fast multipole methods. *J. Comput. Phys.*, 217(1):100–122, 2006. doi: 10.1016/j.jcp.2006.01.048.

[149] C. Schwab and J. Zech. Deep learning in high dimension: neural network expression rates for generalized polynomial chaos expansions in UQ. *Anal. Appl.*, 17(01):19–55, 2019. doi: 10.1142/S0219530518500203.

[150] A. Shapiro. On concepts of directional differentiability. *J. Optim. Theory Appl.*, 66 (3):477–487, 1990. doi: 10.1007/BF00940933.

[151] Y. Shin, J. Darbon, and G. E. Karniadakis. On the convergence and generalization of physics informed neural networks, 2020. arXiv:2004.01806v2 [math.NA].

[152] M. A. Shubin. *Pseudodifferential Operators and Spectral Theory*. Springer Ser. Sov. Math., Springer Verlag, 1987. doi: 10.1007/978-3-642-56579-3.

[153] A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numer.*, 19:451–559, 2010. doi: 10.1017/S0962492910000061.

[154] V. Thomée. *Galerkin Finite Element Methods for Parabolic Problems*. Springer Series in Computational Mathematics. Springer Berlin Heidelberg, 2007. ISBN 9783540331223. doi: 10.1007/3-540-33122-0.

[155] F. Tröltzsch. *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, volume 112. American Mathematical Soc., 2010. ISBN ISBN: 978-1-4704-1174-9. doi: 10.1090/gsm/11.

[156] N. N. Vakhania, V. I. Tarieladze, and S. A. Chobanyan. *Probability Distributions on Banach Spaces*. Mathematics and its Applications. Springer Dordrecht, 1987. ISBN 9789027724960. doi: 10.1007/978-94-009-3873-1.

[157] A. Van Barel and S. Vandewalle. Robust optimization of PDEs with random coefficients using a multilevel Monte Carlo method. *SIAM/ASA J. Uncertain. Quantif.*, 7(1):174–202, 2019. doi: 10.1137/17M1155892.

[158] A. Van Barel and S. Vandewalle. MG/OPT and multilevel Monte Carlo for robust optimization of PDEs. *SIAM J. Optim.*, 31(3):1850–1876, 2021. doi: 10.1137/20M1347164.

[159] P. Wacker. MAP estimators for nonparametric Bayesian inverse problems in banach spaces, 2020. arXiv:2007.12760 [math.PR].

[160] S. Weissmann. Gradient flow structure and convergence analysis of the ensemble Kalman inversion for nonlinear forward models. *Inverse Probl.*, 38, 2022. doi: 10.1088/1361-6420/ac8bed.

[161] A. T. Wood and G. Chan. Simulation of stationary Gaussian processes in $[0,1]^d$. *J. Comput. Graph. Stat.*, 3(4):409–432, 1994. doi: 10.2307/1390903.

[162] L. Yang, X. Meng, and G. E. Karniadakis. B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data. *J. Comput. Phys.*, page 109913, 2020. doi: 10.1016/j.jcp.2020.109913.

[163] D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Netw.*, 94:103–114, 2017. doi: 10.1016/j.neunet.2017.07.002.

[164] K. Yosida. *Functional Analysis*. Springer, Heidelberg, 1980. doi: 10.1007/978-3-642-61859-8.

[165] J. Zech. *Sparse-Grid Approximation of High-Dimensional Parametric PDEs*. PhD thesis, ETH Zurich, 2018. doi: 10.3929/ethz-b-000340651.